Answers to referees for the submitted manuscript (revised title, as requested):

Evolution of oxygen, stratification and their relationship in the North Pacific Ocean in CMIP6 Earth System Models

L. Novi, A. Bracco, T. Ito and Y. Takano

Dear Editor,

Thank you very much for your consideration. We report below our point-by-point answers to the referees, which also detail the implemented changes.

Referee#3 – Second review.

"Summary
Novi et al have improved the manuscript in this revision."

We thank the referee for this positive evaluation.

"Clarifications to the methods for non-specialists:
I am pleased with the edits and additions that Novi & colleagues have made to explain their approaches. I also agree with them that these approaches are useful for oceanography and biogeochemistry, and I hope myself to learn and implement them in my own work. On this note, can the authors provide a link to their analysis code on their Github? If the authors are serious about uptake of these techniques by the community, I think that providing their "publication ready" code would go a long way to achieving this goal."

We thank the referee for this positive feedback and appreciation of the improvements in our work. The Data Availability section contains the codes that we used in this work. We changed the title of that section to "Data and codes availability" to point the reader to that section. In particular, that section contains the GitHub links to the python code for δ-MAPS and for the Information Entropy calculations. For the hotspots calculation, we used a collection of command line arguments via CDO-climate data operator, cited in the References as Schulzweida, U.: CDO User Guide (2.1.0). Zenodo. https://doi.org/10.5281/zenodo.7112925, 2022. We pointed to CDO also in the Data and codes availability section. Additionally, we made available a sample code for the hotspots calculation, now available at https://github.com/superlju/IPVO2hotspots/tree/main

"Signposting
The addition of hypotheses has improved the signposting."

We thank the referee for this positive feedback.

"Line 340: I think a more informative title would be "Climate indices and their relationship with the IPV and O2". Something like this or related."

We thank the referee for this recommendation. We changed the title to “Estimation of climate indices and their relationship with IPV* and O2**”.**

Line 399: Can you end this section with something that tells the reader what we have learned?

We thank the referee for this suggestion. We have restructure and updated this paragraph in order to satisfy this request.

“Further analyses
Thanks for including the analysis on AOU and O2 solubility. While I was disappointed that this analysis wasn’t carried through the paper into the predictability (hypothesis 2) and hotspot (hypothesis 3) sections, I understand that the paper is large enough as it is.”

We thank the referee for appreciating our work. The discussion we included on the AOU and O2sol analysis (Lines 309-314 of the first revision and Suppl Fig S5) relates indeed to the predictability potential (paragraph 4.1). We acknowledge that we didn’t include the AOU and O2sol in the hotspots analysis because – as also stated by the referee – we preferred to keep the manuscript length concise. We believe however that the analysis of hotspots of AOU and O2sol could be an interesting topic to address in a separate work.

Specific comments:
• Line 73: Replace “oceans” with “North Pacific”.
Thank you, we have updated the sentence as requested.

• Line 543: What do you mean by “challenge the overarching hypothesis?”
The overarching hypothesis in this work was that in the North Pacific the spatial-temporal variability of O_2 reflects that of ocean ventilation, which can be measured by the magnitude of the isopycnic potential vorticity (IPV). By that sentence we meant to test the hypothesis stated above. We reframed the sentence in the manuscript to better clarify this point.

Thank you for considering my input to your research,
Pearse J. Buchanan.

Referee#2 – Second review.

Thanks to the authors for providing an updated manuscript. Many of the author’s responses have clarified points for me, and after gaining a better understanding of the research I find I still have many questions about the work. Therefore I recommend that there be a second major revision for the manuscript. I hope that my points help make the manuscript more accessible to an interested audience, as I still believe there can be a lot of value in using IE for ocean biogeochemistry. I am a bit disappointed that some of my concerns have not been addressed in this new manuscript. I have repeated those concerns within my review.

Major points:

Overall manuscript

"There still needs to be much serious consideration on how best to frame this work for its intended audience. I appreciate the consideration about the mathematics behind information entropy and the mapping procedure; however, *Biogeosciences* reaches a wide variety of scientists interested in the Earth system with varying backgrounds in mathematics. Some of the references, like those to topology, can be easily misinterpreted or misunderstood by readers unfamiliar with the mathematics. Even if the terms are correct they can present a barrier to the audience."

We introduce methods not commonly used in the ocean biogeochemistry community, providing a novel framework to assess predictability or changes, focusing on oxygen and IPV. In our first submission, we limited the description of the methods to their generalities, but we were asked to provide more details. Referee#3 explicitly asked for "Clarifications to the methods for non-specialists" (see first report of Referee#3) with a particular focus on the IE and the SED calculation. In response to our changes Referee#3 commented "I am pleased with the edits and additions that Novi & colleagues have made to explain their approaches. I also agree with them that these approaches are useful for oceanography and biogeochemistry, and I hope myself to learn and implement them in my own work.".
We think, therefore, that the mathematical description is useful. We now explain the meaning of topology.

"I wonder if the title can more clearly reflect how the work uses stratification as a proxy to understand ocean oxygen and its evolution. Currently the manuscript sounds like a model intercomparison which it is not."
To address this new point we changed the title to:
"Evolution of oxygen, stratification and their relationship in the North Pacific Ocean in CMIP6 Earth System Models".

There are numerous errors in language use, generally in the use of gerunds and matching up singular/plural nouns and verbs. It makes parts of the manuscript quite difficult to understand. I cannot provide every example but I would strongly recommend the authors conduct a thorough proofreading for the manuscript.

We will proofread the manuscript again. It would have been helpful, however, to mention a few, as all authors have proofread the manuscript. We found few instances of docx to pdf conversion problems (which are not under our control) but not many typos.

Introduction

"There is no reference to current reconstruction schemes of ocean oxygen, or indeed the fact that observations of oxygen are sparse in space and time. As it is part of both the abstract and the discussion it should be included in the introduction."

In the previous review, the referee asked us to include a discussion on oxygen reconstructions in the discussion and conclusion part, which we acknowledged and included. Using the referee's words, they asked to restructure the discussion and conclusion section as follows: "I would recommend changing Section 5 to "Discussion and Conclusions" (or adding a separate Discussions section) and including a critical analysis on the relationship of this study to other work on the PDO and North Pacific oxygen (e.g., Ito et al 2019) or to reconstruction efforts (e.g. Sharp et al., 2022)." We addressed this concern (see our first answer report). To address this new point, we added a reference in the Introduction (Bindoff et al. 2019) together with a mentioning of the sparseness of oxygen observations and uncertainties associated with them.

"I appreciate the highlighting of the hypotheses and the questions in this new version, but the paper should be framed such that it references, in order:

1. The aim/objectives of the paper
2. The questions the paper aims to answer
3. The hypotheses for these questions

Right now the order is objective -> hypotheses -> questions, which makes it difficult to read. Additionally I do not think that the hypotheses need to be referenced as HYP1/2/3, as this just makes the manuscript more difficult to read and reference."

We re-arranged the sequence of our presentation.

We were asked in the previous review stage from this referee to make a clearer connection between the results and hypotheses and to add signposting to help the reader following this. To satisfy those comments, which we believe improved the manuscript presentation, we needed to add an explicit signpost to point the reader to the three hypotheses (HYP1/2/3), along with the major re-organization of the sections. We wish to keep them, as they were found useful by a different reviewer.

"There is no literature review on the use of IE in oceanography. Any citations about its application exist much later and they should absolutely be included in the introduction to provide justification for this work."

We now add references to IE in the Introduction.

"In addition, there needs to be a section in which the use of IE for physical properties like temperature is compared to its potential use for ocean biogeochemical properties. If space is an issue, this is something that could replace parts of the beginning discussion on ocean oxygen cycling which is a bit high-level, has very few references, and also mentions quite a few biological processes which are ignored for the rest of the manuscript."

We respectfully disagree with this comment (which we do not fully understand). Is the reviewer asking for a new section where we compare IE use for Temperature to its use for oxygen? Why? What is the motivation? This request was not pointed out as concern in the first review stage. What is the added value given the hypothesis to be tested in the manuscript?

"Lines 113-115 seem like an additional objective in the manuscript. If this is the case it needs to be reformatted."

We thank the referee for this comment. We moved the comment.

"Generally an introduction ends with an outline of the following sections. In the current introduction there is too much emphasis on the methods (which should be in the immediately-following section). A brief outline of what is included in the methods can be in the section introducing the following sections."

We respectfully notice that this is a personal preference and many papers in Biogeosciences do not include the outline at the end of the Introduction. We added a short outline, while not finding it useful.

Materials and methods

"Which historical experiments are being used? ESMs have emissions-based historical forcings and concentration-based historical forcings. In either case the radiative forcing is not necessarily the same between models with different atmospheric components and atmospheric chemistry."

This point is irrelevant to the main outcome of the manuscript, because the relationship between O2 and stratification could display commonalities across models (different patterns potentially, but common behaviors) independently of the radiative forcing. However, we are more explicit in saying that we are using historical runs with CO2 concentration-based historical forcing.

It would help to include which variables you take from the CMIP6 models. Also, please define how you calculate density from temperature and salinity and how it is implemented in Equation 4. Is it using EOS80 or TEOS10? Are all model outputs gridded so that 0-200m are full cells, or is there some interpolation involved?

TEOS10 (but in the upper 200 m differences will not really matter much). The models have different grids and levels. A linear interpolation is applied in the vertical and all datasets are bi-linearly interpolated to a 1x1 degree in the horizontal.

"I still find the names of these indices (Ind1, Ind2) to be very difficult to internalize, both in the methods section and afterwards. I would strongly suggest eliminating the shorthand entirely."

In the revised version, we already added a subscript to the indices to recall what changes they refer to: *means* for the changes in means, *variability* for the changes in varaibility, and *extremes* for the changes in exterems. We think that eliminating the shorthand entirely would burden the readability. We ackowledge that it might not be the best format for all the readers, but the notation adopted was consistent with previous literature. We anyway replaced the names with Δ_{means}, $\Delta_{variability}$ etc).

"For the definitions on lines 191-193 I would strongly suggest using an overbar to indicate time averages rather than angle brackets."

This is a matter of personal preference. We prefer angle brackets.

“I am still unsure about the arguments about the robustness of the extreme metric. If a 32 year time series of, say, winter stratification is used, the extreme is effectively the 97th percentile. This definition is necessarily dependent on both the length of the time series used and the number of ensemble members used. Additionally, is there a reason why you use seasonal averages to calculate the extremes, but then divide by the total number of months within that season? I am not sure if I am missing something major here but this continues to confuse me.”

To compute the extremes we proceed as follows: (1) in each season of the first period, say 1950-1981, we compute the multi-year seasonal min (or max) using monthly data, not seasonal averages. This is the min (max) value ever reached by that variable in any of the three months considered for a given season. We set this value as threshold, then we count how many times (still at monthly frequency) over 1983-2014 the threshold is exceeded (above or below) in months belonging to that season. Therefore, as we work with monthly data, the following division is done over the total number of months considered. So that, for example, in winter we compare the number of times that a threshold is exceeded in the winter months over 1983-2014 (which we counted at monthly frequency), with the number of months contained in all the winters in that period, i.e. 3months*32 years.

We have verified that the definition is robust and different than using a fixed percentile. The reviewer may verify it as the codes are public. Indeed, the hotspot identification does not vary much with the percentile chosen (given the hotspot definition), or with the period selection, as shown below for the GFDL-ESM4 model.

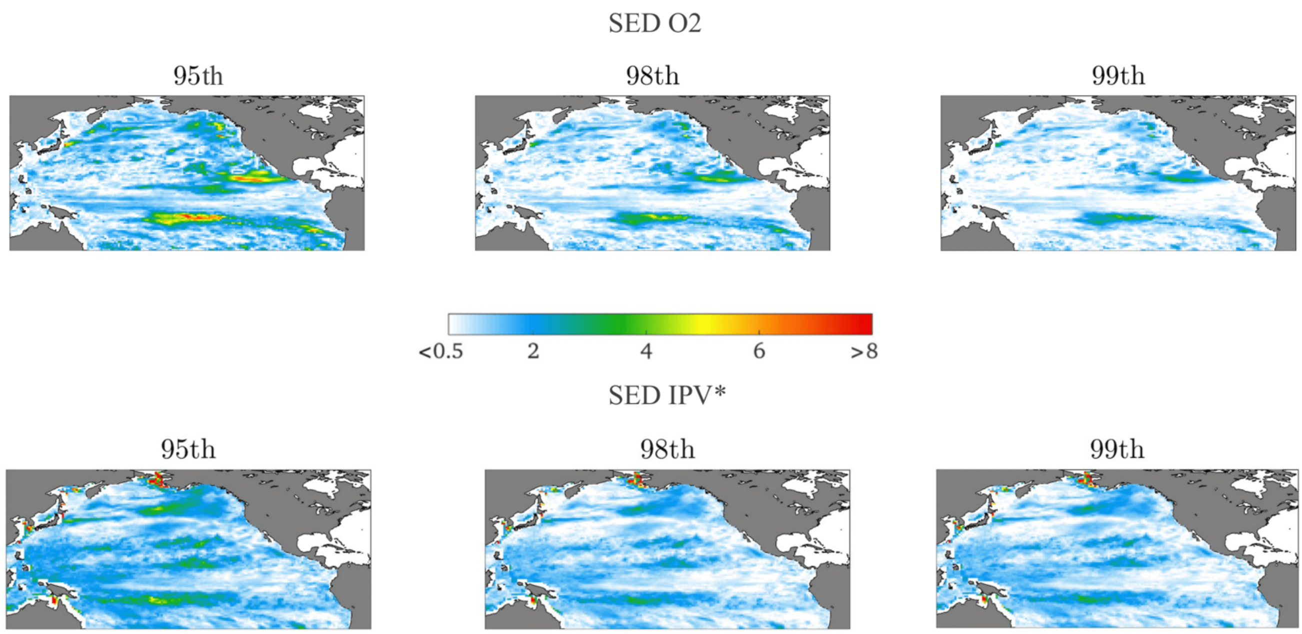

Fig. 1 : Historical SED computed with different thresholds.

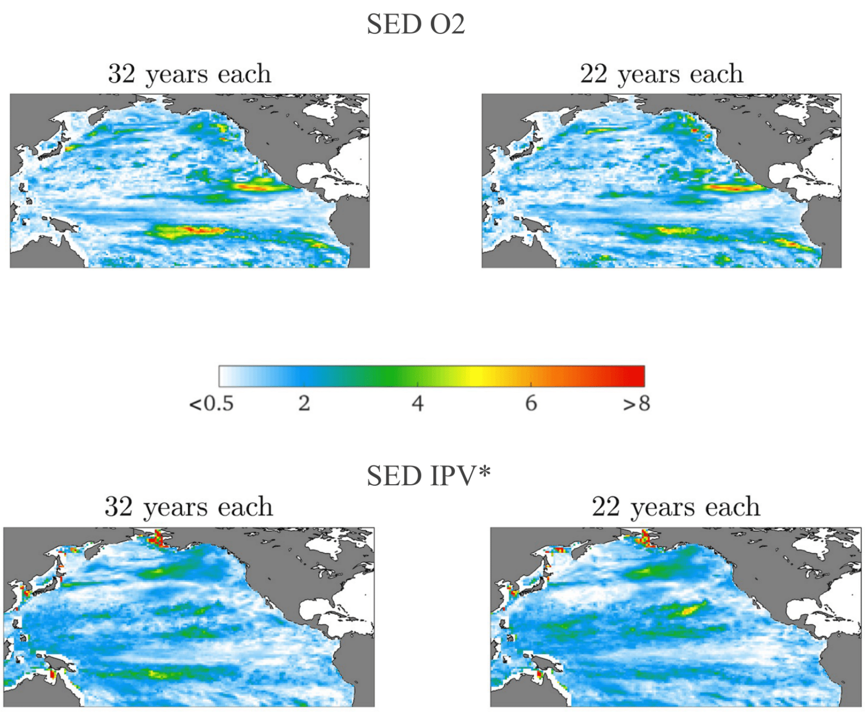

Fig. 2: Historical SED computed over different periods.

Results

The manuscript mentions that the hindcast "not surprisingly" displays the best fit for IPV and O2. This makes me wonder about the usefulness of a fully-coupled ESM compared to other ocean BGC models forced by reanalyses. What are the benefits of using a fully-coupled ESM? If the representation of the PDO is the largest issue in reproducing IPV and O2 (although I'm not convinced this is true) then I would assume this work would be more successful using ocean-only runs. Coupled models exhibit a "signal to noise paradox" (e.g., Scaife and Smith 2018, Zhang et al 2021) which results in an apparent decrease in predictability that may be restricted to model world. How would this impact the interpretability of your results?

Noting that none of the above issues were raised in the first review, we would like to point out that with the goals stated for this (our) manuscript, ocean only models would have not been useful. In addition:
1) We never say in the manuscript that the representation of the PDO is the largest issue. We say that even after removing the PDO signal (whichever it is) the O2 – stratification relationship changes across models and is not robust by and large outside the Equatorial region. We do say, however, that the O2 evolution has been shown to be influenced by the PDO, therefore the PDO focus.

2) E3SM-2G is introduced and used as reference model dataset (the best we can do with models run at climate-scale resolutions).

3) The benefit of using fully coupled ESMs is that they can project into the future, ocean only runs cannot, and we explicitly want to explore how the IPV-O2 relationship may evolve (and if such evolution displays common behaviors across models)

4) Coupled models (some at least) may exhibit a signal to noise paradox, but as it turns out their predictability potential varies greatly and indeed some models are more predictable (have overall lower entropy) than the reanalysis or hindcast counterpart (Figure 2).

Considering the detailed analysis presented in Falasca and Bracco (2022) on two models, coupled climate models continue to be more predictable than observations whenever predictability is defined on their topological properties. The papers cited by the reviewer apply different definitions of predictability to specific processes that differ from what we are exploring here.

I am a bit confused about the argument about how low-frequency PDO modulation alongside higher-frequency variability can produce low predictability. By that argument there is information associated with the PDO, but it is just more difficult to isolate. Is this not a case where the timescales of interest need to be controlled for? You use interannual changes in seasonal values for IPV and oxygen, but if we know the PDO to operate on decadal timescales, is there any way that the IE method can be altered to take these longer timescales into account?

The method is looking for recurrences. At those scales, a signal that emerges from the superposition of different time scales (noise at short time scales and a quasi-sinusoidal modulation at long time scales) will show very little predictability at intermediate time scales – those of interest here. It is possible in principle to apply a running mean - let's say decadal- on each grid point of the domain and each variable, and re-run the IE analysis to isolate the PDO signal, but this strategy will require much longer time-series (a couple of centuries at least) to be meaningful and will not answer the questions posed in this manuscript. We filtered/isolated the PDO signal instead.

"Please ensure that all figures follow inclusivity guidelines. In particular Figures 4, 5, 9, and 13 need to be revised so that readers with deuteranomaly (red-green colourblindness) can interpret the figures. "

Thanks for the comment. All figures were tested already with the free software *Coblis – Color Blindness Simulator* to verify that they are visible by deuteranomaly, protanomaly, tritanomaly and blue cone monocromacy. We report below how the software shows the figures. We only show variations for Fig. 13 and for Fig. 4 because Fig. 5 has the same color palette as Fig. 4, and Fig. 9 has the same colorpalette of Fig. 13.

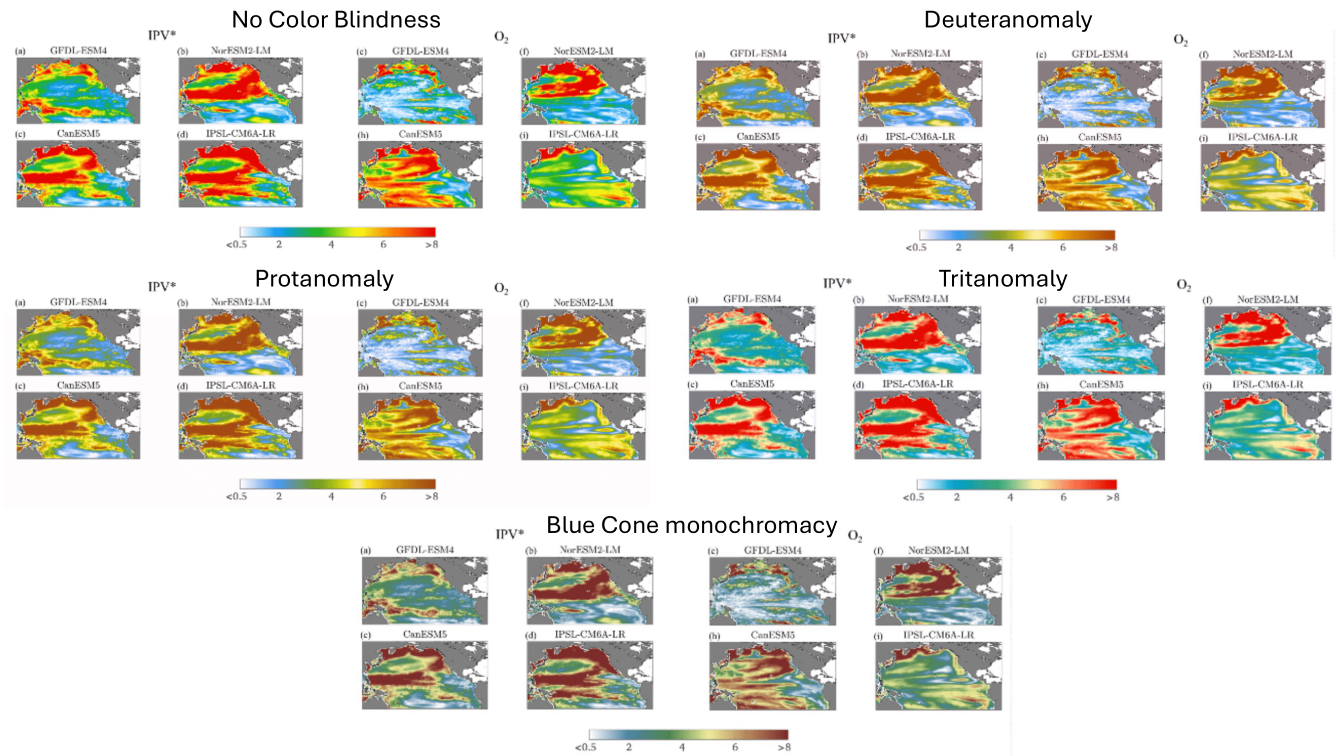

Fig.3: variations of Fig 13 (same colorpalette as Fig. 9 of the submitted manuscript)

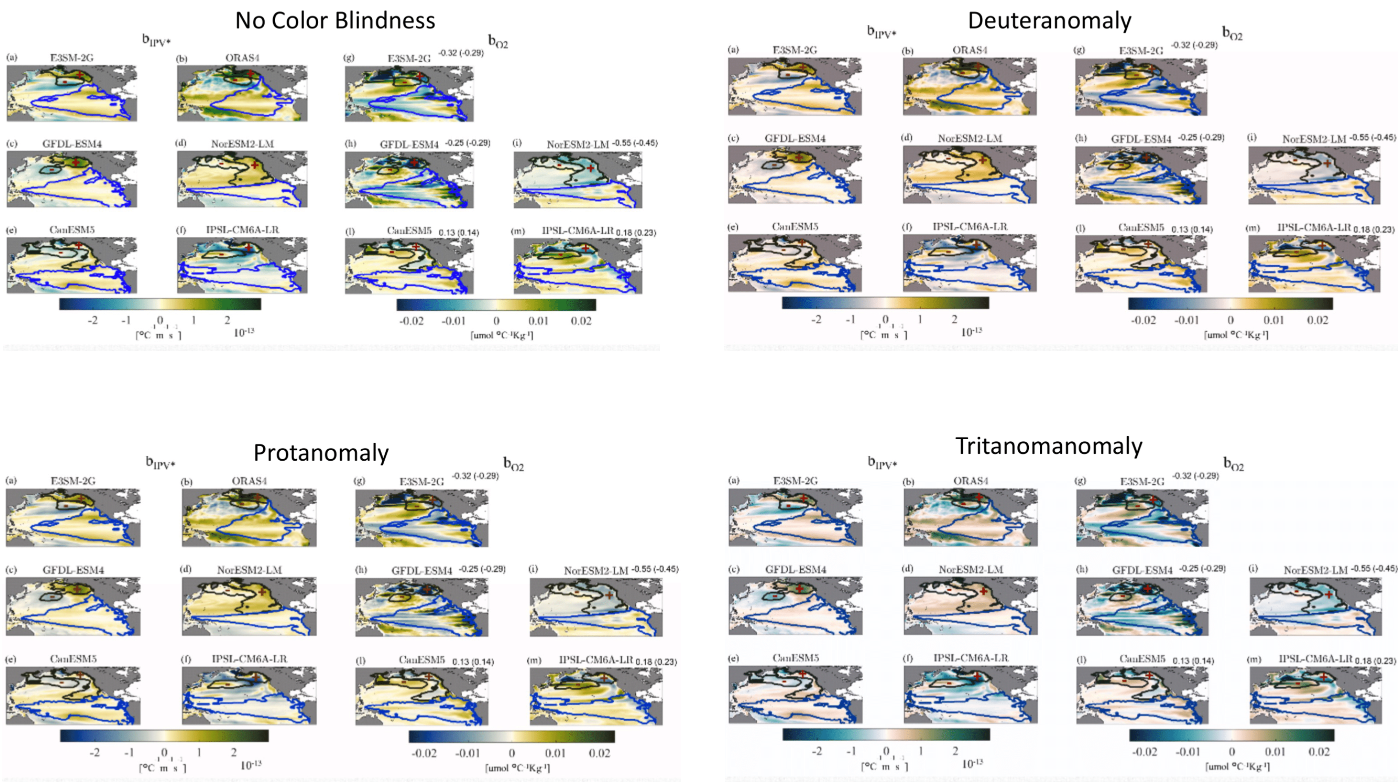

Fig. 4: Variations of Fig.4 (same color palette of Fig. 5 of the Submitted manuscript)

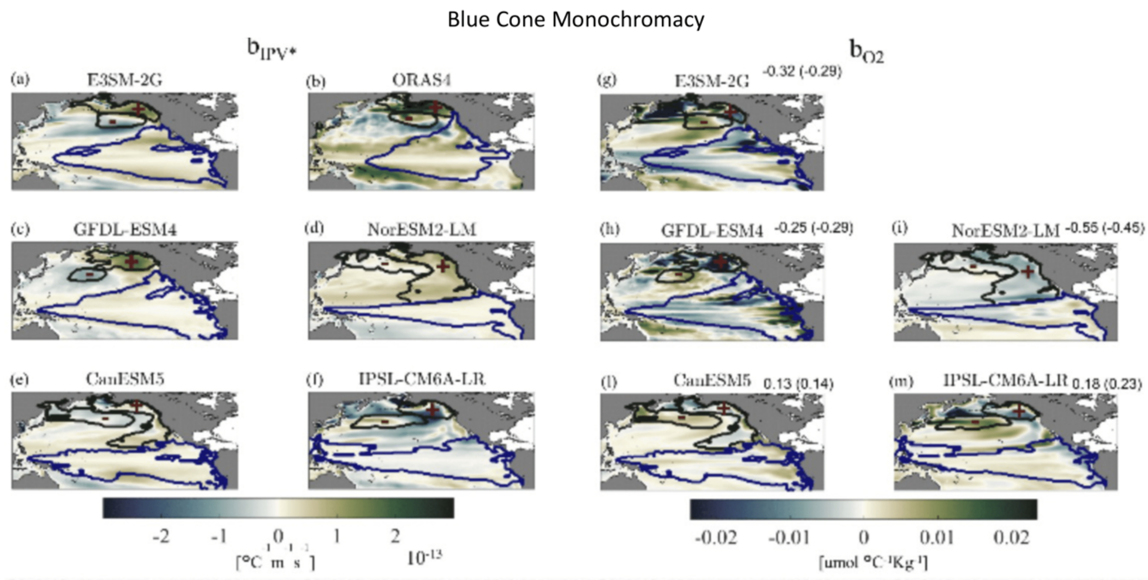

Fig.5: variation of Fig.4 (Same color palette as Fig.5 of the submitted manuscript)

Additional points on figure legibility:

● Figure 1: Please indicate which panels illustrate biases within the figure itself and ensure that "RMSE = " is included in each subplot.

We modified the figure accordingly (the title lines got too busy though, which required introducing shorter names for the models)

● Figure 2: The years for the historical and future runs should be included in the figure, as well as some reference to IE so that the reader can differentiate the figure from later figures. Also, the colorbar direction seems counterintuitive to me, as I would normally associate green with "good" (i.e., lots of information) and this is actually for high entropy.

The reason why the years are not explicitly included in the figure(s) is to avoid redundancy. The "historical" period– as stated multiple times in the manuscript and in the caption – is 1950-2014 for the modes but 1960-2014 for the hindcast and the reanalysis. Adding the information in the figure would force us to either (i) repeat the period atop of each panel or (ii) adding the period below the title but specifying that it is 1960-2014 for hindcast and reanalysis. In both cases the figure would be contain a lot of text without adding any information as it is repeatedly stated in caption and text.
The colorbar direction is a personal preference. Also, the colorbar has not changed since the first submitted version and was not problematic in the previous version.
The reference to IE is given in the caption but we added a title to explicitly indicate it.

● Figure 3: Change the x-ticks for the historical run to green and the future scenario run to black. Subpanel b needs a similar label to differentiate the green and black dashed lines.
Thank you for this comment, we modified as requested.

● Figures 4 & 5: the contours for ENSO and PDO are difficult to separate for some panels. Please use dashed lines for one of the indices so that the reader can more easily discern which

regions are related to ENSO and PDO.
Thank you for this comment, we modified as requested.

● Figure 6: The years in the caption are wrong and they should reflect the two periods used in the averaging rather than combine them into one period.
We modified the caption but we retain the period as it is correct (while divided in 2 for the calculation as explained in the text).

Discussion/Conclusions

Thanks for including a discussion. I think there still needs to be many more citations. For instance, the first page cites only Ito et al 2019. On line 534-535 you reference a wide range of mechanisms but there is no relevant literature included.

We repeated some of the references contained in the Introduction and added some new ones.

In the discussion, the manuscript references that reconstructions need to interpolate onto a regular grid. This is fleshed out more in the authors' response. This is not true – there is no inherent requirement to interpolate observations *or* innovations onto a regular grid when creating a reconstruction. There is utility in doing this, particularly as reconstructions are often used to compare or benchmark models, but any further assertion should be removed from the manuscript.

We thank the referee for noticing this, we will reframe that sentence to avoid any superflous implication. However, the sentence is -we believe – for the IE calculation (not reconstructions), which can be performed on ARGO data interpolated to a regular grid and cannot be performed as defined on quasi-Lagrangian observations.

Please include a concluding paragraph that restates the objectives and main findings of the paper.
Thank you for the recommendation, we added what requested.