Answers to referees for the submitted manuscript

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

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

Referee #2

The manuscript would benefit from much more explicit signposting throughout.
The analyses are impressive but are presented in such a way that they feel separate and the reader must make their own logical steps. In addition, the manuscript lacks a critical analysis of the results and their connection to existing literature on BGC reconstructions. Adding text to connect all the sections together and references to other literature will allow readers to better understand the method and its applicability to the problem of sparse observations of BGC variables, even outside of ocean oxygen.

We thank the reviewer for this general comment. We modified the presentation of the results following the above suggestions.

General Comments:

- "The manuscript would benefit from explicit signposting throughout its sections. The authors set out their hypotheses in the introduction but there are jumps between the results sections. The hypotheses focus on the regulation of O2 variability by IPV* yet most of the analyses approach IPV* and O2 separately (albeit connected by the PDO index)."

We added signposting explicitly linking each subsection of the results to one of hypotheses and we significantly restructured the main text to help clarify our message.

- "Additionally, hypothesis (3) seems to relate to regions where predictability is high, but in the results the identification of hotspots uses the residuals from the PDO regression, which is where predictability is low. The lack of signposting makes it difficult for the reader to understand the full implications of the previous analyses when a new analysis is introduced, and there is ambiguity as to how well the hypotheses have been answered in the conclusions section."

In the revision, we clarify that we isolate the PDO signal to verify if the low predictability in the N. Pacific (north of the ENSO region) is an issue of time scales (i.e. there is a low frequency PDO modulation with a high frequency 'noise' due to both atmospheric and oceanic variability. The PDO is indeed a lower frequency mode compared to ENSO and has most loading at higher latitudes, where weather 'noise' is greater). However, even when the PDO is isolated we do not find a strong anticorrelation between PDO-induced changes in IPV and PDO-induced changes in O2 in the E3SM forced simulation and there is large intermodel difference in the CMPI6 runs.
In the original manuscript, the rational for running the hotspot analysis on the residual fields was motivated by the fact that the PDO-forced component is low frequency variability and does not vary much over time (b, i.e. the regression coefficient between the PDO-forced physical fields and the PDO index, does not depend on time, which is only contained in the PDO index, which is shown). In addition, we had verified that the changes in the mean essentially coincide with the trends (therefore we were not removing any significant information). Taken together, these two considerations imply that the extremes will not depend on the PDO forced contribution. This is indeed the case, but we should have used the whole signal. We do so now, and both the maps of the indicators and SED are essentially unchanged (see for example the figures below for historical and future SED). We thank the referee for this comment.

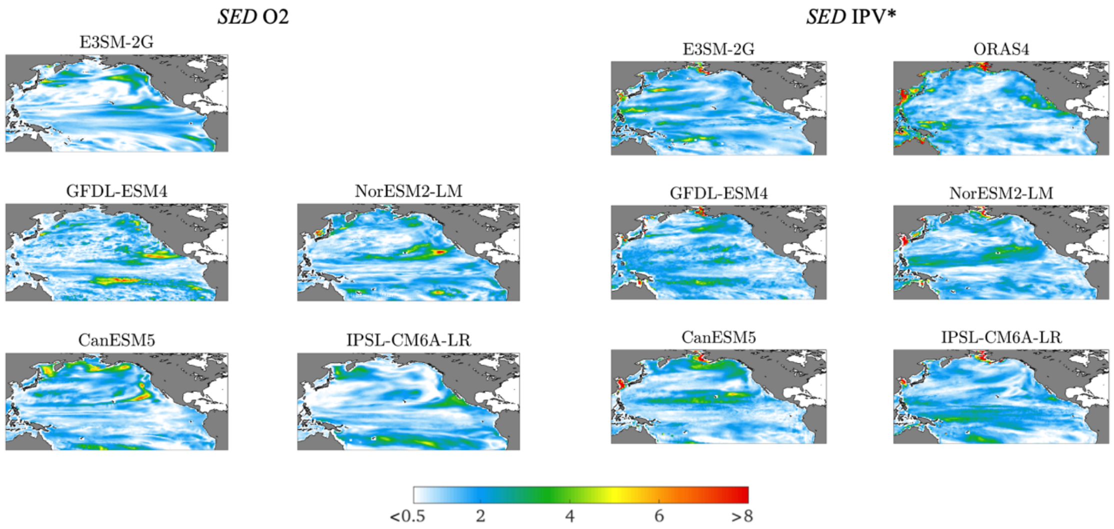

Fig.1 Historical (and 1960-2014 period for the hindcast and ORAS4) SED index for the whole fields.

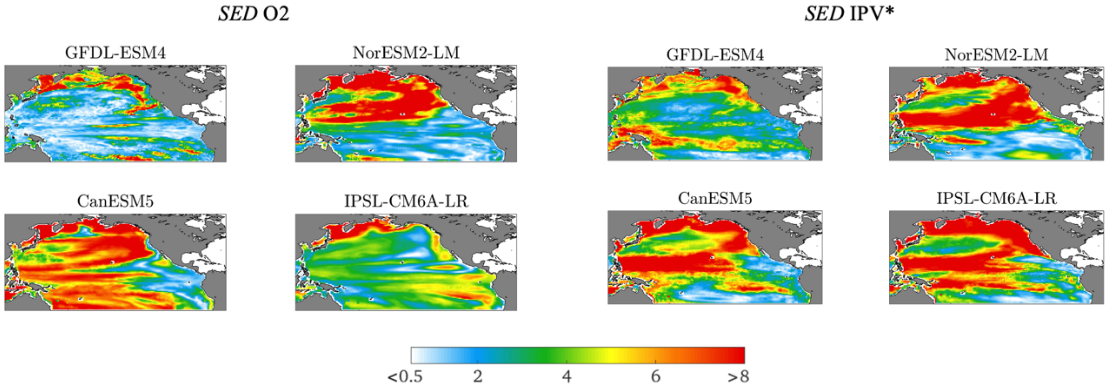

Fig.2 Future SED index for the whole fields.

- "I am not sure why the PDO was chosen as the potential proxy for understanding upper-ocean O2 concentrations. A priori I would assume that ENSO would play a significant role, and the results show a connection between ENSO and predictability of IPV* and O2, whereas when looking at the PDO alone there appears to be a limited connection. In the introduction Ito et al (2019) is mentioned but some further explanation would be helpful."

  Our focus is on the North Pacific, as stated in the Introduction, and outside the tropical band the PDO modulates the variability at climate scales more so than ENSO. As also explored in previous literature (for example, Ito et al. (2019)), the dominant mode of oxygen variability in the northern Pacific Ocean is correlated with the PDO index which explains about 25% of the variance. To further verify it in the present work, we computed the first EOF for the E3SM-2G hindcast 0-200m O2 and IPV*anomalies over 1960-2014 over the northern Pacific (20.5°N-69.5°N;115.5°E-60.5°W) and the corresponding time series for the first principal component (*pc1)*. The first EOF explains 25% of the oxygen variance and about 12% of the IPV*variance. The *–pc1* shows a significant and strong correlation (Pearson's R coefficient) with PDO timeseries computed using SST anomalies as detailed in the manuscript (R = 0.83, pval < 0.01 for O2, and R = -0.44, pval <0.01 for ipv*, after applying a 5-year moving means). The correlation with the PDO is higher than the one with ENSO, which is at most R = 0.34, pval < 0.01 for O2, after applying a 3-month moving mean. This is consistent with previous knowledge that a coherent basin-wide pattern of oxygen variability is mostly associated with PDO in the northern Pacific Ocean. We included in the revised *Introduction* a clarification of this point and we discuss the outcome of the EOF analysis when indicating how we separate the PDO forced-component.

- “Additionally, I would be interested in seeing the regression analysis using both ENSO and the PDO as predictors and seeing how the residuals depend on the climate indices used for the regression.”
  We computed the linear regression analysis using the ENSO index, with the same procedure used for PDO, obtaining similar b shapes of the coefficients -as to be expected - but much lower absolute values, further confirming that the effect of PDO is overall dominant in the regression. Color limits are +/-3 standard deviations of the ensembles as in the main text. We included this finding in the Supplementary Material.

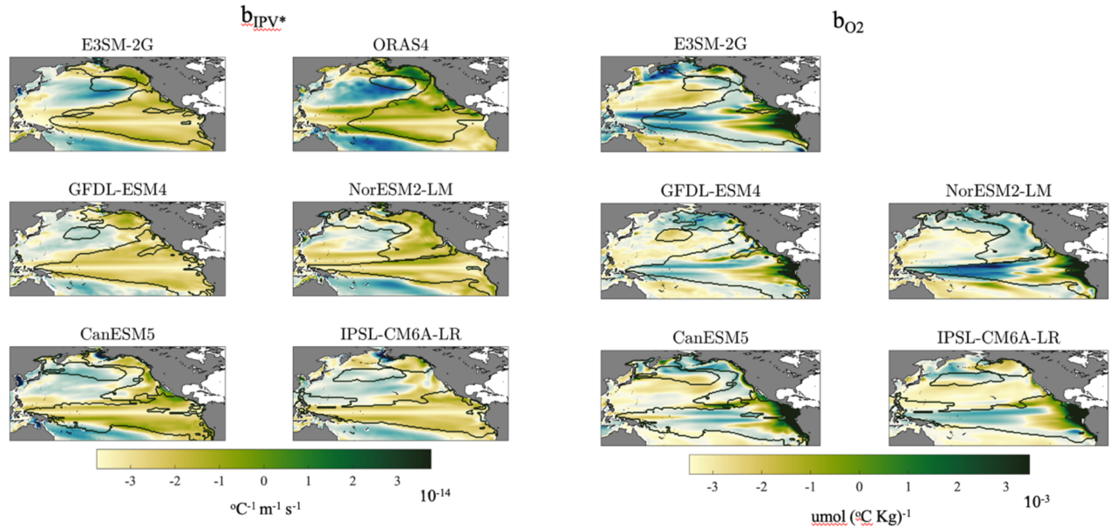

Fig. 3:

Fig. 3 bIPV* (left) and bO2 (right) ENSO-regression coefficient maps with superposed contours of the ENSO and of the PDO+ and PDO- domains. The correlation coefficients among the corresponding maps for the same model or hindcast are also indicated. Color limits are fixed as +/- 3 standard deviations of the ensemble for each variable over the whole area.

- “There is no discussion section, and the conclusions section seems to repeat the findings of the study without any connection to the wider literature. 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 thank the referee for this comment. In the revision we provide a “Discussion and Conclusion” section that addresses the requested points. We extensively restructured the text to improve and extend it as suggested. Few examples of what we included:
  We added a discussion on how our findings align with Ito et al (2019) in identifying a cohesive basin-wide prevailing pattern of oxygen variability in the northern Pacific primarily associated to the Pacific Decadal Oscillation.
  We also discussed that reconstruction efforts (such as the one proposed in Sharp et al. (2022) for example) need to interpolate on a regular distribution, leading to additional uncertainty.
  We stress that our framework could help determine the large-scale predictability potential of any field of interest.
  We also included a discussion of the limitations of our work, as also stated elsewhere in this revision file. These include, for example, recognizing that oxygen

concentrations in coastal areas are influenced by complex biophysical interactions and physical processes that state-of-the-art climate models currently cannot fully resolve. As a result, projected oxygen trends may exhibit variability even within subregions under the same scenario, as demonstrated in studies like Bograd et al. (2023) for EBUS. Analyzing coastal dynamics at the required scales would necessitate higher resolution models, which, if projected into the future, would use, however, CMIP6 runs as boundary conditions.
A more extensive discussion and critical analysis with regard of existing literature is also provided in the revised manuscript.

- "I would like to see more discussion on the use of the CMIP6 ESMs. What are the implications of using relatively coarse-resolution models in this analysis? Particularly in the higher latitudes where eddy mixing is parameterized. What are the implications of having such a broad inter-model range in the results? It is not clear to me whether the emergent relationships are consistent enough across the models to justify the emphasis between IPV*, O2, and the PDO."

CMIP6 models are state-of-the-art in terms of climate prediction capabilities, but they are imperfect and limited in resolution. However, they are the only tool we have that allow for an evaluation of the decadal modes of climate variability, their impacts and their potential changes in a warming planet. Recognizing the biases these models have, we also analyzed a hindcast (E3SM-2G). Our goal is precisely to see if there is a relationship that is robust across models between large-scale climate modes of variability in the N Pacific and their impact on IPV and O_2. If this was the case, independently of the PDO or ENSO representation - which may differ in each model - we could reasonably conclude that IPV, which can be monitored, for example through ARGO floats, could be used to track the large-scale variability of O_2. This could also be done through models, where the resolution can be more easily increased if a simple (or no) biogeochemical module is included, and for which a validation can be conducted with less uncertainty on physical (instead of biogeochemical) variables due to the greater abundance of observations.
Indeed, the variability across the models is, for some of our hypotheses, too large to reach any conclusion and the relationship not as strong as hypothesized on the base of the available sparse observations. At the same time, at least for the historical period, our analysis allows for identifying which models may be more realistic in its representation of the large-scale variability of the North Pacific.
We expanded the Discussion and Conclusion section to include the points above.

- "I would like to see some discussion on the use of predictability studies for real-world reconstructions. As I understand it, the predictability mentioned in this study assumes perfect knowledge at a time *t* of a specific field, either IPV* or O2 (for Section 4.1) or sea surface temperatures (for the PDO index used in the regression analysis in Section 4.2). However, T and S profiles from Argo are still irregularly distributed, which is a nontrivial problem for ocean reconstructions of both heat and salinity themselves (e.g., Smith and Murphy 2007, Cheng and Zhu 2016) and ocean BGC (e.g., Turner et al. 2023, Keppler et al. 2023)".

Models and analyses or hindcasts allow for testing if a system is predictable - notwithstanding their biases -, or to calculate the potential predictability of a system. When using observations such as ARGO profiles, it is necessary to interpolate onto a regular distribution, which will increase the uncertainty. If the predictability potential is high, such an exercise will be worth it, if the potential predictability is low, futile. This assumes that the model(s) are capturing the main features driving the -in our case large scale – predictability. We added a discussion point on the final section of the manuscript to reflect this comment.

Specific Comments:

Line 174: I am not familiar with this definition of extremes. Is there a reason you have not used a general quantile threshold or a distribution fit to characterize the extremes? As you use only one realization for each model, there is a nontrivial chance of "significant" changes in extremes due to internal variability.

We thank the referee for this comment. Building on previous works (Falasca et al. (2019); Falasca et al (2022)), we expect the topology of a given model to remain relatively stable, i.e. we do not expect the member choice to significantly influence the calculation of extremes and hotspots with the chosen definition. We verified the robustness of our results, computing the extremes indicators of four randomly-chosen ensemble members of the CanESM5 model for the whole signals of IPV* and O2 for the historical periods. We found no significant changes in extremes and SED, as shown in the figures below.

A major advantage of the hotspot definition chosen is that it accounts for changes in mean, variability and extremes at the same time in the identification of the hotspots. In other words, it accounts for the topology of the simulated climate fields, which can be characterized by considering all the three aspects together (as done also, implicitly, in δ-Maps).

The definition of extremes adopted aims at including information on seasons exceeding corresponding baseline extremes, without choosing a priori a threshold on the current distribution, which is especially relevant for comparing changes with respect to a reference baseline.

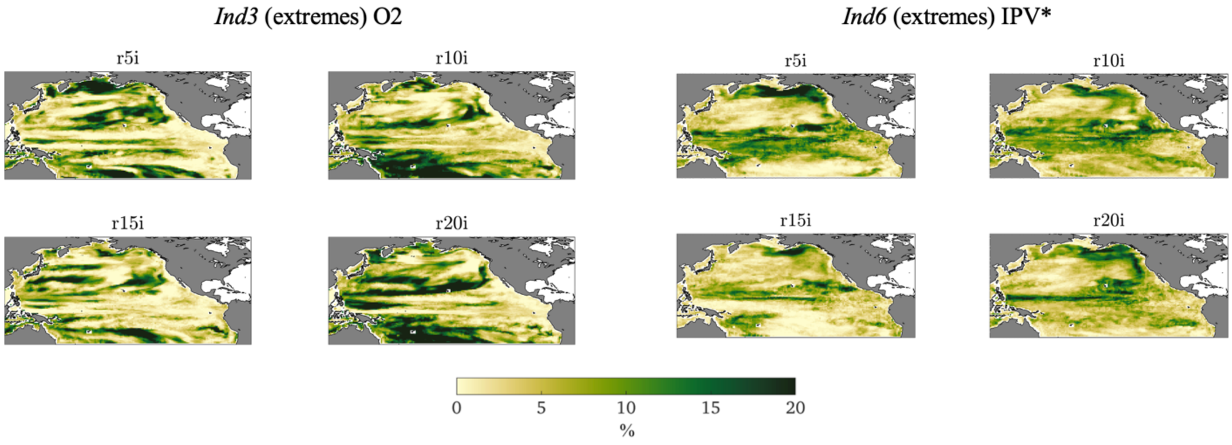

Fig4. O2 (left) and IPV*(right) indicators for changes in extremes (historical, whole signal) for four different ensemble members of the CanESM5 model.

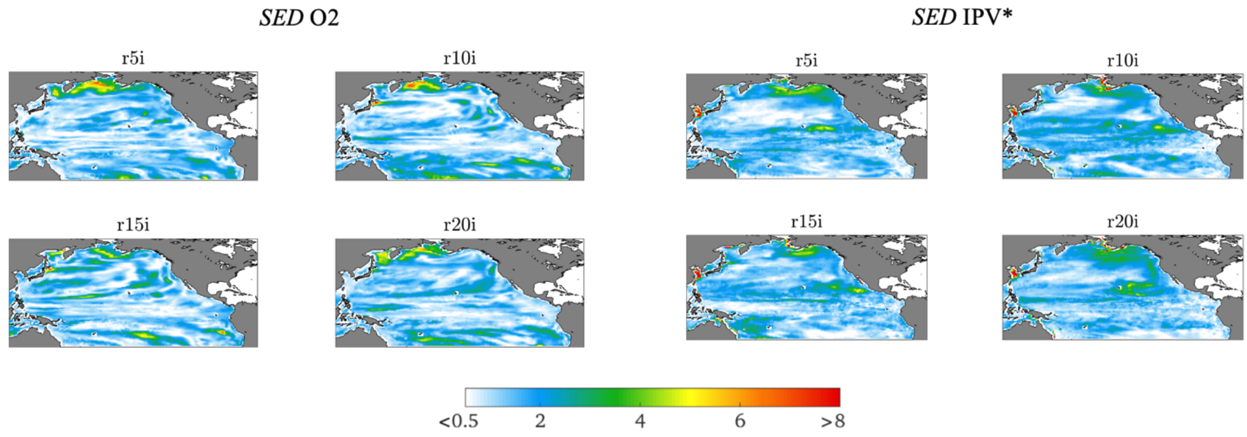

Fig5. O2 (left) and IPV*(right) SED indices (historical, whole signal) for four different ensemble members of the CanESM5 model.

Line 189: Why is the depth horizon set to the top 200 m?
We set the depth horizon at 200m as it represents a reasonable trade-off between being in the upper thermocline, being deep enough to smooth the gas-exchange effects dominant at the surface, but not too deep as models tend to lose a good representation of variability (interannual and longer) when compared to observations.

Line 193: What is the reasoning behind the choice of ESM models for the ensemble? Without choosing multiple realizations for each model and using 4 models, the ensemble seems quite small relative to the available CMIP6 output. Also it would be good to know which biogeochemical models each ESM employs in Table 1, even if biological oxygen cycling is outside the scope of this manuscript.

The objective of this work is to present a set of metrics (a framework) that may help in identifying relationships and quantifying predictability potential across physical and biogeochemical variables.
With the goal above in mind, we chose a subsample of the CMIP6 catalog that did not resemble each other in terms of components and/or resolution. Adding more models will not challenge the main conclusions: 1) There remain significant intermodal differences in the representation of climate variability in the North Pacific. This is not just reflected in the patterns, but also in the representation of the relationships between physical (IPV) and biogeochemical (O2) variables, which is the focus of our investigation. 2) Such a relationship appears weaker than we hoped in all datasets analyzed, but is statistically significant under several metrics, in the hindcast and in some models (GFDL being the best example).
We clarify this point in the revised manuscript.

Line 206: JRA-55do v1.4 has an anticyclonic tropical cyclone in the NE Pacific in 1959 (as well as multiple anticyclonic tropical cyclones in the Atlantic, see https://climate.mri-jma.go.jp/pub/ocean/JRA55-do/). The issue is fixed in v1.5. It would be ideal to re-run the hindcast with the corrected atmospheric forcing. If that is not possible 1959 should be excluded from your analysis, perhaps using 1960-2014 as your historical period.

We thank the reviewer for the comment, as we were not aware of this problem. We re-ran all the computations involving E3SM-2G and ORAS4 (for consistency of comparison) over 1960-2014. Results are nearly identical, as one year alone does not modify the PDO or the overall Pacific variability (see figure below). In particular, for both E3SM-2G and ORAS4, we re-run the entropy regression, and hotspots analyses over the new period. For the latter, we divided 1960-2014 in two intervals of equal length, 1960-1986 and 1988-2014. We replaced these new analyses in the revised manuscript.

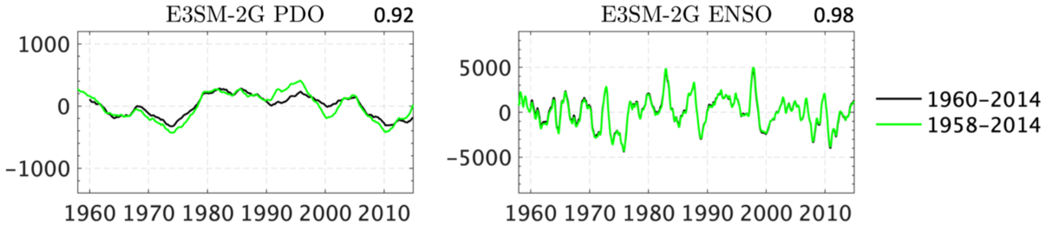

Fig. 6: PDO (left) and ENSO (right) indices (SST cumulative anomalies) calculated using δ-MAPS in the historical period, using E3SM-2G over 1958-2014 (green) and 1960-2014 (black). The numbers on top-right of the panels are the correlation coefficients between the two curves where they overlap in time, i.e. 1960-2014.

Line 238: I am not sure exactly what predictability means. Based on (1), this assumes some perfect knowledge of t=0 everywhere in the North Pacific for each of these models? What is the length of time used to calculate the IE? Perhaps I misunderstand something in the methods with these questions, but clearer definitions for predictability both here and in the methods would be helpful.

We thank the referee for pointing out the need for additional clarifications. We added a more detailed explanation of predictability as linked to the entropy measure in the method section. The quantification of IE relies on recurrence. In each point, the entropy of the field under investigation is associated with recurring microstates in its time series (that define the system and thereby impacts its predictability). The higher the predictability of a time series the more recurrent are its temporal dynamics, i.e. the easiest will be to predict its future evolution. We computed IE using monthly data over the whole historical and future periods.

Line 244: What do you mean by “area most impacted by ENSO”? Has there been a regression analysis done for ENSO in each of the models? Figure 2 a-f seem to have quite high IE (low predictability) in the equatorial upwelling region across all the models.

We rephrased it to indicate more clearly that we mean. Here we meant the domain identified as ENSO-related by δ-Maps, which well follows what would be identified by an EOF analysis over the SST field as region where the variance explained by PC1 is greatest. As stated in the manuscript “The predictability potential is higher along two stripes enclosing the ENSO domain but excluding the upwelling cold tongue regions.” This result is not new. The predictability of the cold tongue has been found to be low over much longer periods in the IPSL model (Falasca et al. (2020)), and in SODA reanalysis and a large suite of CMIP5 models in Ikuyajolu et al (2021).

Line 319-320: How do the regression coefficients stay relatively stable if the domain for the PDO evolves? Also, what is the implication about the residuals dominating the evolution of both IPV* and O2 in terms of predictability (and, in particular, predictability related to the PDO)?

The regression coefficients are computed point by point using the local value of O2 (or IPV*) timeseries, and the PDO time-series (the same for all grid points). The PDO signal is computed using the timeseries associated to the PDO domains, which capture the overall decadal climate variability regardless of their exact shape, and this decadal variability is not changing significantly in most cases (the PDO dynamics are not changing significantly, which is not surprising).

Technical Comments:

Line 119: Should xi be multidimensional?

Thank you for catching this typo. Yes, xi is indeed multi-dimensional. Also, in equation (1) of the revised version, we replaced xi$\in\mathbb{R}$ with **xi**$\in\mathbb{R}^d$, being d the **xi** space dimension**.**

Line 135: What is the reasoning behind the use of 4 microstates?

We ran a sensitivity analysis of the entropy field to m, within a meaningful range according to previous literature (Ikuyajolu et al. (2021)). We tested m = 2,3,4,5 for GFDL-ESM4 over 1950-2014. In the figure below we show the results when the same color scale is used (panels a-d) and when each case has a different color scale to highlight the spatial features (panels e-h). The pattern, i.e. areas more (less) predictable relative to the surroundings are substantially unchanged, i.e. the geographical patterns are robust, as also found in Ikuyajolu et al. m = 4 and m = 5 show reasonable entropy values, but we chose m=4 is because it spans the widest range of possible values, as also shown by the histogram below.

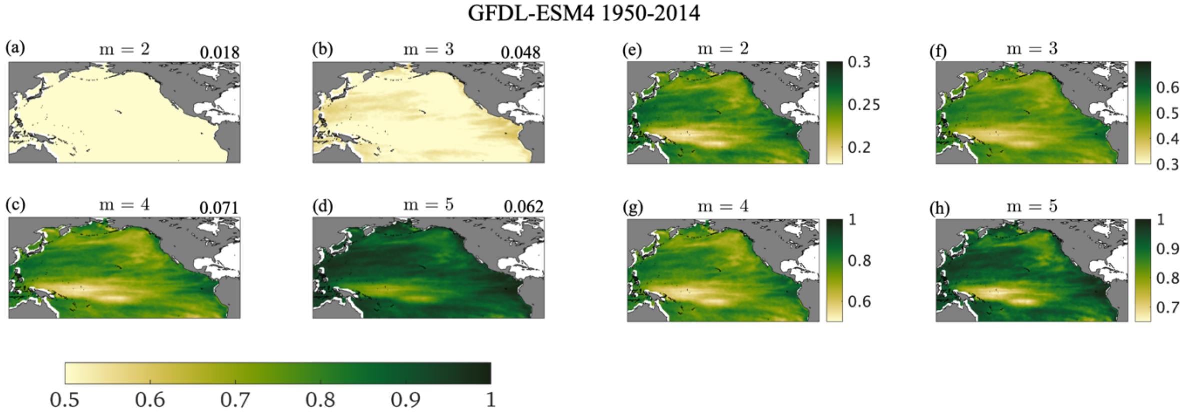

Fig.7: Historical (1950-2014) GFDL-ESM4 Entropy maps computed using m=1,2,3,4. Panels (a)-(d) have the same color scale than the one used in the main text. Each panel from (e) to (h) (same fields as panels (a)-(d)) has a diffent color scale, to show the spatial pattern.

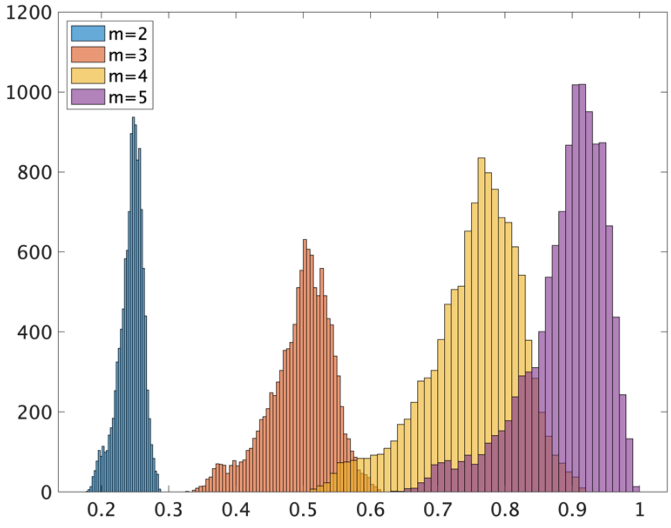

Fig.8: Histograms for historical (1950-2014) GFDL-ESM4 Entropy maps computed using m=1,2,3,4.

Line 160: Why have you not used the same years across the models and reanalysis and hindcast product for each period?
We are looking at a decadal variability mode, the PDO, as main focus of our study, and wanted to cover the longest possible period for the models and as further as possible in the future projections (historical, 1950-2014 and future 2036-2100), but the reanalysis and hindcast are only available over a shorter time period. Therefore, we decided to keep the model on a longer time range to capture the PDO temporal scales as best as possible.

Line 162: I find the use of shortcuts like *Period 1/2, Ind 1/2/3, yseasm* to hinder my understanding, particularly when examining the figures. More descriptive shortcuts (e.g., \overbar{DJF1983-2014} - \overbar{DJF1950-1981} instead of *Ind1*) would greatly help readability
We changed notations in the revised manuscript as recommended.

Answers to all minor points is after the comments list:

Line 186: perhaps define N2 here?
Line 188: Do you mean Equation 4?
Line 195: Which variables are used from the CMIP6 models? If T and S, it would be helpful to know which models use EOS-80 and which use TEOS-10 for their density fields.
Line 196: Do you mean SSP?
Line 207: ORAS4 could use a description in this section. Also to explain about the lack of O2 results (I presume the reanalysis has no biogeochemistry?)
Line 221: RMSE values embedded within Figure 1 rather than presented as a list would increase readability of this section.
Line 223 and elsewhere: Please use consistent units formatting with superscripts.
Line 279: Is it possible to include a scaled version of the NOAA PDO time series in Figure 3 for comparison?
Line 294: Is there one bO2 and bIPV* for all scenarios or are the coefficients calculated for each scenario separately?
Line 335: ORAS4
Line 435: Repeat here the vertical domain (0-200m)
We thank the referee for catching some typos and giving recommendations for improvements. In the revised manuscript we implemented all the corrections and requested changes. We answer to the questions as follows:
L195: That is correct, the variables from the CMIP6 models are potential temperature and salinity. The requested information will be included in the revised table describing the models.
L207: Yes, that is correct. We will add a clarification in the revised text.
L294: The coefficients are computed for each dataset separately.

References used in the answer to Referee#2

Bograd, S. J. et al.: Climate change impacts on eastern boundary upwelling systems. *Annu. Rev. Mar. Sci*. **15**, 303–328, 2023.

Ikuyajolu, O. J., Falasca, F. and Bracco, A.: Information Entropy as Quantifier of Potential Predictability in the Tropical Indo-Pacific Basin. Front. Clim. 3:675840. doi: 10.3389/fclim.2021.675840, 2021.

Ito, T., Long, M. C., Deutsch, C., Minobe, S., and Sun, D.: Mechanisms of low-frequency oxygen variability in the North Pacific. Global Biogeochemical Cycles, 33(2), 110–124. https://doi.org/10.1029/2018GB005987, 2019

Falasca, F., Bracco, A., Nenes, A., and Fountalis, I.: Dimensionality Reduction and Network Inference for Climate Data Using δ-MAPS: Application to the CESM Large Ensemble Sea Surface Temperature, J. Adv. Model. Earth Syst., 11, 1479–1515, 2019.

Falasca, F., Crétat, J., Braconnot, P., and Bracco, A.: Spatiotemporal complexity and time-dependent networks in sea surface temperature from mid- to late Holocene, Europ. Phys. J. Plus, 135, 1–21, https://doi.org/10.1140/epjp/s13360-020-00403-x, 2020