List of most significant changes to the manuscript:

1. The aim of the study has been stated more clearly in the introduction (lines 156:160).

2. The novelty and relevance of the study have been stated more clearly in the abstract (lines 47:49) and introduction (lines 154:156), and discussed more thoroughly the discussion section and in the conclusions (lines 554:557).

3. The ensemble description (section 2.1) has been thoroughly revised, details about boundary schemes have been added and sources of ambiguity removed.

4. The validation metrics (in section 2.2) have been explicitly described.

5. Our analysis method for oxygen change has been discussed more thoroughly. This includes relations to more commonly used metrics like AOU (lines 247:251, 305:308, 542:544), its assumptions and potential limitations (lines 247:255, 542:547), and why we believe it is still applicable to the case study of the NWES (lines 253:255, 547:551).

6. The methods' equations have been revised according to Reviewer 3 suggestions and the whole analyses re-run. This didn't change the results significantly.

7. While calculating correlation the linear trend has been retained as explaining long term trends is one of the objects of the study. This did not change the bulk of results but removed the need for [former] figure 8 as when the long-term trend is retained there is indeed a significant negative correlation between SS and PEA in the Norwegian Trench in both HADGEM and IPSL.

8. Additional details about the performed analyses and the considered variables have been introduced explicitly in the methods (lines 285:296).

9. Throughout the result section we added detail on all previously undiscussed correlations and on the possible causal mechanisms and/or covariances that can explain them.

10. We completely re-wrote the discussion section focussing on our results and on how they fit in the context of available literature and on the limits of our methodology, also linking at what stated in the introduction.

11. We completely re-wrote the conclusions section summarising what the main finding of this study is, how it contributes to the study of oxygen dynamics in the NWES and in coastal and shelf ecosystems, and finally we make recommendations for future research directions.

NB: All line number cited in the answers to comments refer to the edited manuscript without track-changes.

REVIEWER 1

GENERAL COMMENTS:

REV1:

The manuscript investigates the processes driving near-bed oxygen changes on the Northwest European Continental Shelf under a high-emissions climate change scenario, with a focus on the intermodel uncertainties in these processes and their effects on oxygen. This work extends and qualifies the results of a previous study (Wakelin et al., 2020) by adding two additional sets of regionally downscaled model projections within the high-emissions forcing scenario (RCP 8.5).

Ocean deoxygenation and coastal hypoxia under climate change pose a serious threat to marine ecosystems. Robust understanding and projection of these processes is important for effective adaptation of ecosystem services. Given the lack of skill of coarse resolution global ESMs in coastal regions, regional downscaling of ESM projections will likely play a critical role in exploring this topic.

Although these additional model simulations provide valuable new insights into the fate of the oxygen in the region, some of the main conclusions reached by the authors are not well supported by the evidence presented. The scope of the study is not well defined and the manuscript overall lacks focus and rigor. While the scientific premise of the study is valuable, major revisions are required for this work to be fit for publication in Biogeosciences.

ANSWER:

We thank Reviewer 1 for the useful comments. We thoroughly revised the manuscript according to all reviewers' comments, we took special care in providing additional support to our conclusions, we better defined the scope of the study and its limitations.

SPECIFIC COMMENTS:

REV1:

(1)

A major result of the paper is the attribution of the deoxygenation hotspot in the Norwegian Trench to a relaxation reversal of the Norwegian Trench Current; but this interpretation is not well supported or well argued. The authors argue that (1) a relaxation of the advective current causes a freshening of the shelf region causing increased stratification, and (2) correlation suggests that the increased stratification is responsible for deoxygenation. Holt et al (2018) argue that changes in stratification are responsible for the relaxation of the current, opposite to the authors’ explanation. In most cases, an increase in stratification would come from surface warming and precipitation changes; this null hypothesis should be disproved before seeking alternative explanations.

(2)

It is also not clear in the results whether vertical mixing or horizontal advective transport is dominating oxygen supply to the Norwegian Trench region, which should guide the conclusions made. Note that Wakelin et al (2020) do link reduced current to a recoupling of export with near-bed respiration; perhaps this is connected to the change in sign of correlation between SS and stratification (320).

(3)

Lastly, ‘tight coupling’ in Figure 11 is not necessarily convincing by eye. A stronger link has to be made.

ANSWER:

(1)

We appreciate how this may have not been entirely clear in the text but our conclusions about the causes of the relaxation of the WNT current are not at odds with Holt et al. 2018. In both works it is increased stratification at the northern entrance of the trench that reduces oceanic inflow into the North Sea, this in turn increases retention of fresh water from continental Europe and the Baltic within the North Sea, driving freshening (Holt et al. fig1e,f, this study, fig 3.) and a further increase in stratification in the North Sea. Then there certainly is a component of the increase in stratification due to the atmospheric temperature forcing, but this cannot explain the hotspot of increased stratification as surface warming is homogeneous across the domain (fig. 3.).

We revised the manuscript (lines 503:508) to make all of this clearer.

(2)

We acknowledge the reviewer’s comment, however in our model configuration, the lateral transport of oxygen from the Baltic open boundary does not change in time. We acknowledge that this was not clarified in the manuscript, but in GFDL and IPSL, the Baltic open boundary has fixed climatological values for all tracers, including oxygen. This choice was made because ESMs are scarcely reliable for an enclosed sea such as the Baltic. This ensures that the deoxygenation signal we detected in the Norwegian Trench does not originate from the Baltic boundary through lateral transport.

HADGEM also uses a fixed climatology at the boundary for both biogeochemical variables (including oxygen) and freshwater input, with the difference that the Baltic boundary is treated like a river, rather than an open boundary.

Both boundary treatment choices do have some limitations, however they also rule out lateral transport from the Baltic as the source of the deoxygenation signal in the Norwegian Trench.

We clarified this in the methods (lines 209:213), and added some discussion about the limitation from having a climatological boundary at the Baltic while being able to rule out lateral transport as a contributing factor for deoxygenation hotspots (lines 516:518).

(3)

We complemented section 3.9 (now 3.7) with correlation coefficients for the analysed timeseries and revised the text according to the results (lines 449:459).

REV1:

The title of the manuscript suggests that the focus of the paper is on ‘intra-scenario variability’; however, it is unclear what the scope of this is and how effectively it can actually be investigated with available tools. Uncertainty in ESM climate projections (and by extension, downscaled projections) fall broadly into three categories, regarding (1) internal model variability, (2) intermodel uncertainty, (3) and scenario uncertainty. The term ‘intra-scenario’ would suggest that you look at both internal variability and intermodel uncertainty, which is not really the case. Due to the small sample size (three models) and inconsistencies in the model and methods used for downscaling in the older HADGEM run versus the IPSL and GFDL simulations, neither internal variability nor intermodel uncertainty is well sampled nor well isolated. Perhaps the term ‘multi-model comparison’ used in the abstract is more appropriate here. This is already addressed somewhat in the introduction (125-135), but should be clarified and given more thought. Claims like “we added an intra-scenario variability dimension (375)” are unclear and misleading, and should be changed.

ANSWER:

We appreciate the focus on "Intra-scenario variability" may be misleading, and we concur with the reviewer that the scope here is to compare the projections of oxygen from the small “multi-model” ensemble. Therefore, we changed "Intra-scenario variability" in the title and throughout the text with "multi-model comparison", or deleted it, revised the introduction by mentioning the categories of uncertainty in projections (lines 143:145) and by more clearly stating the aim of the study (lines 156:160), shifting the focus away from variability estimation. We revised the discussion by stating which sources of variability were not addressed in this study (lines 534:537). We revised the conclusions highlighting the importance of sampling different sources of variability while building regional climate model ensembles (lines 566:575).

REV1:

Throughout the study, the authors claim that oxygen changes in the study region across the three simulations scale with the climate sensitivity of the parent ESMs. If quantifications of these sensitivities are available, they should be presented here. Additionally, an issue with this claim is that the differences in downscaling methods for HADGEM vs the IPSL and GFDL simulations provide uncontrolled degrees of freedom. The authors should provide an argument whether the differences in downscaling techniques should significantly impact the magnitude of oxygen changes. If possible, the authors could run some short sensitivity experiments using the new (used for IPSL, GFDL) setup to test sensitivity to e.g. vertical resolution.

ANSWER:

We added the estimates for the global equilibrium climate sensitivity of the three parent ESMs (these are 4.59, 4.12 and 2.39K for HADGEM2-ES, IPSL-CM5A-MR and GFDL-ESM2G respectively, lines 181:184).

We also added in section 4 a more detailed discussion on how the different downscaling methods from HADGEM may influence the results (lines 519:533). Unfortunately producing conclusive evidence requires ad-hoc experiments, and as the reviewer suggest, this can be quite an expensive task that is not always feasible for multiple reasons, including availability of resources. While we agree that the differences in the model set-up may play a role in the dynamic, these will not be the driving cause of the patterns projected by the model.

Nonetheless, despite some noticeable different responses, the bulk of the behaviour of our ensemble members is still coherent with the tested climate change intensities. This we think shows that our results are still robust with respect to the [limited] model variability represented in our ensemble.

REV1:

In the model used by Wakelin et al (2020), oxygen is not included in open boundary conditions of the regional model so that changes in open ocean oxygen is not included. Is this the case here? This is very important for how the results may be interpreted and should be documented carefully.

ANSWER:

We appreciate this was not explicitly mentioned in the text. In our IPSL and GFDL members oxygen is indeed included in the open ocean and Baltic boundaries, whilst Wakelin et al. (2020), that is our HADGEM, uses a zero gradient-scheme (i.e. boundary concentration equals concentration inside the domain) for most biogeochemical tracers, including oxygen, at the open ocean boundary. The only tracers that are forced with external data at the open ocean boundary are nutrients and inorganic carbon. At the Baltic boundary HADGEM uses climatological values for all tracers, including oxygen.

We improved the Ensemble description in the manuscript, so that all boundary schemes are clearly described (lines 185:220).

The impact of the different treatment of the boundary will largely impact the open ocean part of the domain (that is excluded from the analysis), while the shallow depth and intense winter mixing of the NWES makes so that ocean-atmosphere exchange will reset the oxygen to saturation every winter, or more frequently, throughout the water column, so that oxygen on shelf is scarcely coupled with oceanic oxygen.

REV1:

The authors need to be careful when interpreting correlation as causation. Correlations are only meaningful when there is a process that can explain the relationship. Please be thorough about when a physical/ biogeochemical mechanism can explain a correlation and when a correlation cannot be explained. For example, why would you have a positive correlation between SS and stratification in some regions (Fig. 7)? If strong but erroneous correlations are prevalent, why can we still trust the results? The authors should also provide a discussion of any covariances that may influence the results (e.g. between temperature, stratification, respiration, NPP)

ANSWER:

We agree with the reviewer that correlation does not automatically imply causation, and we can support our interpretation by improving the presentation of the results and the discussion of the attribution of correlations. In particular we added detail about:

[1] corr(SSO2, Tatm)>0 in southern coastal regions, all members, covariance explained by increasing NPP (lines 384:389),

[2] corr(SSO2, PEA)>0 in coastal regions, covariance mediated by the seasonality in NPP (lines 403:408).

[3] corr(SSO2, Tatm)<0 in the Trench and Eastern North Sea, all members, (new results without detrending see later, covariation with increasing PEA, lines 390:392),

[4] corr(BResp, SSO2)>0 in the Norwegian Trench, IPSL, covariance explained by decreasing BResp, due to decreasing NPP, together with decreasing SSO2 due to increased stratification (no strong direct causal link, lines 440:443).

REV1:

In calculating correlations, the long-term trend is removed. I see how this avoids false positives, but how can you assess the drivers of forced changes after removing the long term trends? In this case, it seems that correlations just classify the drivers of short-term variability, which is not what you purport to be investigating. Please explain/ clarify.

ANSWER:

We appreciate this may be of concern regarding our methodology. While analysing the data we did conduct exploratory analyses where trends were not removed, which only resulted in slight improvements of some detected correlations, with no relevant changes in sign. We concluded that trend removal was in this case the most conservative practice.

This perhaps could be justified in systems with short turnover rates where drivers of short- and long-term trend overlap. For example, warming reduces oxygen solubility both on the long-term, through increasing mean temperatures, and on short-term, e.g. during summer months. Or increasing NPP produces oxygen both on the short-term, during a bloom, and the long-term (if coupled with enough mixing) if the productivity of a region increases over time.

Nonetheless, we see a solid point can indeed be made in favour of retaining the trend when calculating correlations, if the aim is explaining the trend, and taking care that, when interpreting results, some patterns will be explained by covariances rather than causal links (false positives). This is what we did (lines 229:304). As for the revised results the only relevant changes are:

1) Negative correlation between O2sat and atmospheric temperature in the Norwegian Trench (all members, instead of non-significant).

2) Negative correlation between SSO2 and atmospheric temperature in the Norwegian Trench and eastern part of the North Sea (HADGEM and IPSL, instead of non-significant).

3) Negative correlation between SSO2 and PEA along the Norwegian Trench (HADGEM and IPSL, instead of non-significant)

for 1) and 3) a case can be made for a causal link, for 2) the pattern is more easily explained by covariance with PEA.

None of these change our conclusions substantially but 3) removes the need for Fig 8. “running correlations between SSO2 and PEA mediated over the Norwegian Trench”.

The correlations involving biogeochemical variables (NPP, BResp) didn't show any significant change.

We attach a revised version of the part of results that changed for a more complete exposition.

REV1:

What is gained by decomposition in section 2.3 as opposed to a traditional O2sat, AOU decomposition? Why is O2phys-ch (O2sat scaled by the initial saturation state) a more meaningful metric than O2sat? The authors end up using O2sat and SS (a.k.a 1-AOU) anyway, so this section can be removed entirely.

ANSWER:

Please note that we re-worked the methods section according to Reviewer3's comments, as a result the definition of ΔO2_other changed slightly and O2,phy-ch,t is no longer present, the comment still applies though.

The ΔO2_phys-ch and ΔO2_other metrics we presented in the methods are indeed related to the traditional O2sat, AOU decomposition, with the difference that they describe the partitioning of oxygen change relative to a reference period, rather than the distance from equilibrium at any specific moment.

This renders them interesting as metrics because they are directly comparable, being both Δ concentrations (unlike Osat (concentration), AOU (Δ)) and they sum up to the total ΔO2. This allows to quantify how much of the observed change can be attributed to each component.

AOU estimates oxygen consumption (and production) since a water parcel was last in contact with the atmosphere, assuming Osat doesn't change. Our metrics, by explicitly considering changes in Osat, allow to partition oxygen change into the two separate components.

We included a section in the methods explaining this, the relation between our metrics and AOU, and the hypotheses and limitation of both methods (lines 246:255).

Results about ΔO2,phys-ch and ΔO2,other are presented in section 3.4 Contributions to near-bed oxygen change.

REV1:

How are there negative values in the root-mean-square distance calculation (Fig. 2)? Need to provide formulae here for nbias and nurmsd.

ANSWER:

we appreciate the metrics from Jollif et al. may not be as widely known as others, we addressed this by adding their definition in the methods (although we merely described the equations, rather than writing them down, as they are indeed trivial, lines 233:239). As for the negative values of nurmsd, they arise by multiplying rmsd by the sign of the difference of model and data stds, so that a negative value indicates that the model's std is lower than that of the observations, and vice-versa for positive values. We explained also this in the text.

REV1:

Bias-correction for hypoxia measurements should be included in methods

ANSWER:

we included bias correction procedure in the methods (lines 309:315).

TECHNICAL COMMENTS:

REV1:

Grammatical errors and inconsistent capitalization throughout. Please proofread carefully.

ANSWER:

We carefully proof-read the manuscript and corrected errors and capitalization.

REV1:

In all figures, panel labels need to be included.

ANSWER:

Panel labels have been added to all figures.

REV1:

Use consistent terminology for region names. Is the Danish strait the same as Skagerrak? Eastern North Sea is referenced throughout but not delineated on the map in Fig 1.

ANSWER:

We replaced Danish strait with Skagerrak, we also replaced 'Eastern North Sea' with 'eastern part of the North Sea', or similar, throughout the text.

REV1:

Nearly all instances of ‘in fact’ can be removed

ANSWER:

all instances of 'in fact have been removed or replaced'

NB: All line number cited in the answers to comments refer to the edited manuscript without track-changes.

REVIEWER 2

GENERAL COMMENTS

REV2:

Giovanni Galli et al. have evaluated the trends and controls of O2 changes due to biogeochemical and physical changes in the NWES using model data for the 21st century. Unfortunately, the methods are unclear to me (ensemble description, analysis of O2 change) as well as the research questions/novelty of the study. I agree it is important to assess biogeochemical changes and drivers (and their uncertainty) in such a heavily exploited and strongly changing region. As I could not follow the methods everywhere, I can only give an incomplete review of the manuscript at this stage. Here are some comments that may help to improve the manuscript:

ANSWER:

We thank Reviewer 2 for the useful comments, we thoroughly revised the manuscript according to all reviewers' comments. Among other change we improved the ensemble description removing possible sources of ambiguity as well as the methods description and stated the study objectives more clearly.

SPECIFIC COMMENTS

REV2:

The title does not really seem to cover the results (How about ‘ 21st century trends and controls of near-bed oxygen change on the Northwest European Continental Shelf’ or so?

ANSWER:

We acknowledge "intra-scenario variability" was misplaced here. We would refrain to use '21st century trends' because our ensemble is not large enough to produce a robust assessment of expected trends, or to quantify uncertainties. However, under Reviewer1's suggestion, we changed 'intra-scenario variability' to 'multi-model comparison' to reflect this concern.

REV2:

Abstract: Could you quantify some of your statements? What is new here?

ANSWER:

We added some quantification of results in the abstract, and some more in the results section.

We also improved the last paragraphs of the abstract (lines 47:52) and introduction (lines 154:160) by stating the novelty and relevance of this study. In short, we expand on Wakelin et al (2020) showing that the projections of near-bed oxygen presented there are robust in a multi model context and that trends and drivers of change remain coherent at different warming intensities. Finally we want to assess the impact of the change in circulation presented by Holt et al. (2018) on oxygen change similarly across models. These are new findings that allow a better understanding of the projection of near-bed oxygen in the NWES and its drivers. We believe (and reviewer 1 seem to concur) that these are important questions to assess the impact of climate change on shelf seas and plan the needed adaptation, and it may suggest directions for future research.

REV2:

l85-92: You make an elaborate comparison here, but then also underlines the limited usability of Kwiatkowski et al. (2020). I would just highlight the limitation of ESMs to quantify this region if you like, but not compare.

ANSWER:

Our aim here was not really comparing the two results, which we believe may be equally valid, but showing that when looking at oxygen change results may vary also according to the spatial scales and domains being analysed (coastal vs shelf). Then global ESMs have indeed some known limitations in resolving coastal and shelf processes and we rephrased the part of text when we explain this to make it clearer and more specific (lines 101:104).

REV2:

L116-122: which reference(s) is this all based on?

ANSWER:

It's all Wakelin et al. 2020, we are summarising the results of that paper and that took some text, we appreciate this may not have been entirely clear, so we repeated the citation in the text (lines 128:142).

REV2:

L128: for regional models boundary conditions are also highly relevant. Spinup-times could also be mentioned here. Do you wish to provide an exhaustive list here?

ANSWER:

We rephrased the sentence so that we refer to the components of variability (Frölicher et al. 2016): internal, model and scenario variability, and provided some examples of variability sources for internal and model variability (147:150). Since this particular sentence is general, it applies to both regional and global models, to atmosphere, ocean or land, it may not be advisable to mention all possible sources of variability.

REV2:

L131: Why not CMIP6?

ANSWER:

These simulations were implemented before outputs from CMIP6 were available. While we acknowledge the differences between the CMIP5 and the CMIP6 scenarios, CMIP5 still provide a useful set of climate scenarios that are still valuable and that can be associated to the CMIP6 ones.

REV2:

L 134: you did not investigate ecosystem responses?

ANSWER:

We rephrased the sentence, to make clear that we look at near-bed oxygen change.

REV2:

L130-132: I get the feeling here you used three models and then ran 3 ensembles within each model (namely by using slightly different CMIP5 forcings), can you clarify this already here?

ANSWER:

We have 3 downscaled ensemble members, each one is forced with one single CMIP5 ESM. We appreciate this may not have been clear in the text so we revised it accordingly (lines 176:181).

REV2:

Introduction: your final paragraph (lines 126-136) describes what your new contribution is. However, it is unclear at this stage what extra model variability you argue to have covered (and important to note that there are several sources of uncertainty, scenario, model variability, model uncertainty, see for example Fig. 3 in hdps://agupubs.onlinelibrary.wiley.com/doi/10.1002/2015Gti005338). Also, your research questions are not so clear to me. Why did you focus on the near-bed O2 specifically, why not the whole water column and then near-bed as a separate focus?

ANSWER:

We acknowledge the claim of addressing [comprehensively] some aspect of variability may be misplaced here. We improved the final paragraphs of the introduction where we describe the aim of the study so that we don't refer to this (lines 154:160). We also changed in the title and throughout the text the term "intra-scenario variability" to "multi-model comparison", which is perhaps more appropriate here. We also stated in the discussion which sources of variability our study fails to address.

We also improved the exposition of aim of the study that is further qualifying near-bed oxygen projections, their trends, and drivers in the NWES in a multi-model context, and assessing the impact of circulation changes on near-bed oxygen similarly in a multi-model context. All of which represent understudied questions.

We choose to focus on near-bed oxygen because the bottom layer is more vulnerable to deoxygenation than the surface, due to the seasonal decoupling from the atmosphere. Furthermore, near-bed oxygen dictates habitat suitability for benthic sessile or scarcely motile species and it is used as an indicator of eutrophication in the North Sea (Devlin et al. 2023).

We also added to the introduction some more explanation on why the focus on near-bed oxygen (lines 87:90).

REV2:

Line 140: is a 10-year spin-up enough? In ESMs a few hundred years is more common. What drift do you have in your variables during this spinup? If significant, drif should be subtracted from the data at the least (and a thorough discussion should be provided why you can still use the data).

ANSWER:

While a 10y spinup is not enough for Earth System Models, it is for the regional model used here and it is in fact common practice (e.g. Tinker et al. 2014, Holt et al. 2018, Ciavatta et al. 2018). The North Sea has a flushing time of 2-4 years, the Norwegian Trench about 100d (Blaas et al. 2001), depths are relatively shallow and seasonal mixing is intense throughout the water column in most locations; all of this concurs in making these relatively short spinup times acceptable for this domain. Indeed we did not observe appreciable drifts in the model during the spinup period. In addition to that the biogeochemical initial conditions, which include the slowest components of the system (i.e. the benthos) are initialised with the reanalysis from Ciavatta et al. 2018, which ran for an additional 10y, including its own spin-up. We added a couple of lines explaining this (lines 165:168).

REV2:

Sect. 2.1: I do not follow. There are 3 models which all are part of the NEMO-ERSEM model suite (so 3 times almost the same model?). Are these the 3 members then? Which you then forced with ESM data (from GFDL, IPSL and HADGEM)? What are the parent ESMs then (as its says that the boundary conditions of these 3 ESMS are taken from parent ESMs (line 152), these CMIP5 data are fully coupled ESMs without boundary conditions except towards space)? So, do you have 9 model runs in total (3*3)? When you write about ‘all models’ in line 155 you seem to be discussing the ESMs as if you have been running the ESMs, but you used this NEMO-ERSEM setup, right? Anyway, I do not follow. Alternating between the word member and model might be inconsistently done? Maybe a table? What happens in the forcing in the 21st century (e.g., wind/freshwater forcing changes?)

ANSWER:

We have 3 downscaled ensemble members that all use the NEMO-ERSEM suite, each one is forced with boundary conditions from one of three fully coupled CMIP5 ESMs, the "parent ESMs", which are GFDL-ESM2G, IPSL-CM5A-MR and HADGEM2-ES. So we have 3 downscaled ensemble members in total, which, for brevity, we call GFDL, IPSL and HADGEM (instead of “the downscaled ensemble member forced with boundary conditions from GFDL-ESM2G, IPSL-CM5A-MR or HADGEM2-ES”). We appreciate that our presentation might have been confusing, we revised it in order to make it clearer, and we took care of being consistent in the use of ensemble member instead of model (lines 163:220).

As for the forcings in the 21st century they are as described in the text: those available from the CMIP5 (e.g. wind, lateral boundary conditions, etc.) are from CMIP5 (both historical, up to 2005, and climate runs), those not available from CMIP5 are from other sources (e.g. river nutrient loads are from a reanalysis, Ciavatta et al. (2018), multiplied by river discharge). We also revised the

Ensemble description in the text in order to provide a clearer and more comprehensive list of all the forcing fields.

REV2:

L 150 and what are the ECS then of these models?

ANSWER:

We added to the text (lines 181:184) the estimates of the ECSs of the ESMs, thes are 4.59, 4.12 and 2.39K for HADGEM2-ES, IPSL-CM5A-MR and GFDL-ESM2G respectively (Andrews et al., 2012, Dufresne et al., 2013).

REV2:

Sect. 2.3: SS_t0 is not defined here? How is this approach different from AOU (Apparent Oxygen Utilization) or even better TOU (True Oxygen Utilization)? You open with that O2 change have 3 different components, but then you can only separate into 2, right? Namely the temperature effect through its effect on O2 saturation and then biology+circulation as the ‘other’ term (which is like in AOU and I am not aware of a method that can distuingish all 3). Calculating O2 sat and the contribution of O2 from circulation+bio is a simple calculation I would say, and I think the analysis should go beyond this and the correlations.

ANSWER:

Please note we reworked the methods section extensively, following Reviewer3's suggestions. The comment still applies though.

Our approach is indeed related to the classic AOU / O2sat decomposition, with the difference that our approach quantifies components of change at a point in space relative to a reference time period, whereas AOU measures the time-integrated amount of oxygen consumed since a water parcel has left the surface.

Whereas AOU implicitly assumes that O2sat of a water parcel doesn't change since its last contact with the atmosphere, our method explicitly accommodates changing O2sat (due to e.g. ocean warming). This way ΔO2 can be explicitly partitioned in two components, one related to changing O2sat and one related to changing SSO2.

ΔO2phy-ch and ΔO2other are directly comparable measures, being both components of the total ΔO2, as opposed to AOU (Δ concentration) and O2sat (concentration).

As for other metrics such as True oxygen utilisation (TOU Ito et al. 2004) and Evaluated Oxygen Utilisation (EOU, Duteil et al. 2013), the first must be evaluated explicitly at runtime to define the preformed oxygen, which unfortunately we haven’t implemented, and both address some known issues with AOU when O2 concentration and solubility are decoupled, which may happen e.g. when undersaturated water is subducted or when a water mass changes temperature away from the surface, or in the presence of sea ice. This is not quite the case in the NWES that is relatively shallow and well mixed, with intense ocean-atmosphere exchange. Which makes the assumptions behind AOU (and our method as well) a fair approximation. Overall, we felt the use of different metrics wouldn't have brought much improvement to the results. We added some lines in the discussion addressing the limitation of our method (lines 538:551).

While we agree our analysis method is fairly simple, we still believe that this is still informative, and indeed we successfully exploited it to diagnose oxygen dynamics and controls in our ensemble members.

REV2:

Sect. 3.1: Why don’t you bias-correct and only use the model/ensemble trends (like you actually do in e.g. Fig. 3), considering the significant biases? Then the absolute errors are less important and can go into an appendix or so. You seem to have done so anyway for (part of?) your analysis (mentioned in line 276 and caption Fig. 4 only…).

ANSWER:

We don't bias correct extensively (with the exception of the hypoxia estimation in fig. 4) because we almost exclusively look at delta concentrations and correlations and these are not influenced by bias. Instead we use bias correction when calculating hypoxia because when fixed low oxygen thresholds are considered, absolute values are relevant. We added some lines to the manuscript to make this clearer (lines 309:315).

REV2:

Fig. 2: based on the text units here are standard deviations? Maybe just use the full names instead of nurmsd and nbias? Or just call them Root mean square and bias and say that they are normalized? I think it would be good to get the equations from Jolliff et al. (2009) or to use more commonly used metrics like RSS?

ANSWER:

We don't indeed use normalised rmsd but normalised unbiased rmsd, which is normalised rmsd multiplied by the sign of the difference of model and data's stds, so that it can also assume negative values, which indicate that the model's std is smaller than that of the observations, and vice-versa for positive values. We explained this in the text (lines 232:239), we didn't add in the full equation because that is indeed quite trivial.

REV2:

L 259: here you for the first time use the word downscaling, this should be introduced in the methods section.

ANSWER:

We mentioned "downscaling" several times throughout the text.

REV2:

Fig. 3: If you would plot instead of a change over time a change at a certain global warming level (countering the differences in ECS, see Hausfather et al. (2022); 10.1038/d41586-022-01192-2), your model differences will likely be smaller? Would that be a more meaningful way of assessing model differences as showing differences in warming is inherent to choosing models with different ECS?

ANSWER:

We appreciate this could be an interesting angle to look at our projections, and we did run some additional analyses to look into it (see attached document, WNT_and_Warming_Level.pdf).

However, warming level, either global or regional (over the downscaled domain), doesn't seem to be relevant for the development of the change in circulation in the North Sea, whose effects on near-bed oxygen are an important focus of our manuscript. In particular, both IPSL and GFDL, which are exactly the same model with different forcings, show regional atmospheric warming in excess of 2K, but whilst IPSL develops the circulation change, GFDL doesn't.

In the manuscript we improved the statement of the aim of the study, shifting the focus away from the evaluation of model variability and differences. In short, we aim at testing the system's response (with a special focus on effects of circulation changes) at different climate change intensities, as this is at present poorly constrained for many biogeochemical variables (oxygen included) in the NWES. That is why we chose models covering a wide range of ECSs.

REV2:

Can Sect. 3.4-6 be merged?

ANSWER:

we merged 3.5-6 and also 3.7-8, the new titles are as follows:

3.5 Physical controls of oxygen change: temperature and stratification

3.6 Biogeochemical controls of oxygen change: primary production and respiration

REV2:

Sect. 3: I was actually a bit surprised about the section titles here, and it would be good to introduce the reader earlier what you will exactly cover in your results section to answer your research questions.

ANSWER:

We improved the methods section in order to introduce more detail about the analyses we performed and whose results are presented in section 3 (lines 285:304).

REV2:

Sect. 4: this mostly sounds like a conclusions/summarizing section except for the last paragraph. Please try to discuss limitations of your methods, implications, compare to other studies that may show something else? You find many confirmations/consistencies which is fine but makes your work sound less novel or complementary. What other stressors does the near-bed ecosystem experience (trawling/pollution?)?

ANSWER:

We thoroughly re-wrote the discussion focussing on the implication of our results and on the limitation of our methods, and on the evidence available from the literature. We also discussed the combined impacts with other stressors (lines 473:475).

REV2:

L 430: how does your study highlight this? Could you show your regional/downscaled model runs are superior to the ESM output? Same in line 441.

ANSWER:

As we don't indeed provide a direct comparison with the oxygen field in the parent ESMs, we see how these statements were problematic. Providing such comparison would be, we believe, certainly interesting but also out of the scope of this paper. We rephrased the two sentences referring to literature rather than this study (lines 512:515).

REV2:

You mention that you asses ‘ecosystem impacts’ throughout the manuscript, but I would say you mostly assessed a range of physical and biogeochemical changes and the possible drivers of the O2 changes.

ANSWER:

We rephrased all instances of 'ecosystem impacts' or similar in the text to make it clear that our assessment is limited to near-bed oxygen change.

REV2:

L450-451: reference?

ANSWER:

we rephrased the sentence (which is now moved to lines 87:90) and added a reference (Devlin et al. 2023, https://oap.ospar.org/en/ospar-assessments/quality-status-reports/qsr-2023/indicator-assessments/seafloor-dissolved-oxygen).

REV2:

Sect. 5: I do not see so well how this section connects to your results. Please quantify your results and focus your conclusions on the answers to your research question(s). E.g., your conclusions and abstract text are quite different while one would expect them to cover very similar statements. Sect. 3.9 is not discussed or concluded upon.

ANSWER:

We re-wrote the conclusions sections focussing on the answers to our research questions, we explicitly linked to what is stated in the abstract, and we improved the final recommendations. We also improved our discussion on how circulation change in the North Sea affects the development of deoxygenation hotspots.

REV2:

correlation is not causation

ANSWER:

We improved our results section by discussing in more detail the mechanisms that can observed correlations, including some that we failed to discuss earlier. This includes the discussion of covariances that determine correlations between variables also in the absence of a direct causal link. Here some examples:

[1] corr(SSO2, Tatm)>0 in southern coastal regions, all members, covariance explained by increasing NPP (lines 384:389).

[2] corr(SSO2, PEA)>0 in coastal regions, covariance mediated by the seasonality in NPP (lines 403:408).

[3] corr(SSO2, Tatm)<0 in the Trench and Eastern North Sea, all members, (new results without detrending under reviewer1's suggestion, covariation with increasing PEA, lines 390:392).

[4] corr(BResp, SSO2)>0 in the Norwegian Trench, IPSL, covariance explained by decreasing BResp, due to decreasing NPP, together with decreasing SSO2 due to increased stratification (no strong direct causal link, lines 440:443).

MINOR REMARKS

REV2:

Some spelling errors that can be captured by any spellchecker are still in the text

ANSWER:

We thoroughly revised the manuscript and corrected typos.

REV2:

L76: possibly? Sometimes? Regularly?

ANSWER:

frequently

REV2:

L185: limit validation?

ANSWER:

limit validation!

NB: All line number cited in the answers to comments refer to the edited manuscript without track-changes.

REVIEWER 3

MAJOR COMMENT

REV3:

The flawed general premise is that somehow in situ $[O_2]^{sat}$ controls in situ $[O_2]$, while other mechanisms drive the saturation state. However, in the ocean interior and particularly near the seabed, far away from the surface, changes in solubility alone (from changes in temperature or salinity) should have zero effect on in situ $[O_2]$, except in the case where the solubility is reduced below the in situ $[O_2]$. If a parcel of water with salinity S, temperature T, and oxygen concentration $[O_2]$ was artificially cooled down, its O_2 solubility would increase, its O_2 saturation state would decrease, but its O_2 content would remain unchanged. While $\Delta[O_2]^{sat}$ may correlate well with $\Delta[O_2]$ in the real ocean and in marine biogeochemistry models, there is no causation.

ANSWER:

We understand the concerns that Reviewer 3 raises and, while we agree Reviewer 3 is essentially correct, we also believe that, in the specific case of the NWES, that is the object of our study, the less rigorous approach adopted here and in Wakelin et al. (2020) still represents a good approximation that is useful to understand the processes involved. Clearly this will need additional clarification in the text. Our argument is as follows:

There is a main assumption under the method we used, that the water column equilibrates with atmospheric O_2 on timescales short enough (every winter in seasonally stratified regions, more frequently in regions that are well-mixed year-round) to establish a causal link between $[O_2]^{sat}$ and $[O_2]$. This is similar to what is assumed in other widely used metrics like AOU.

This clearly may be far from true in the ocean interior, close to sea-ice interface or in upwelling areas, when a water parcel has been separated from the atmosphere long enough to degrade such causal link.

However the North Western European Shelf has characteristics that do not preclude using the decomposition adopted here: it is a highly dynamic system, characterised by relatively shallow depths, short residence times (2-4y for the North Sea, 100d for the Norwegian Trench) and, crucially, intense mixing, both wind- and tidally-driven; many regions (Irish Sea, English Channel, Southern North Sea) are known to be well mixed year-round, and the regions that stratify do so only seasonally. This effectively resets $[O_2]$ towards equilibrium with atmospheric pO_2 every winter, i.e. towards $[O_2]^{sat}$ (and SS toward 1), hence the causal link.

See also Ito et al. 2004, where differences between simulated preformed $[O_2]$ and $[O_2]^{sat}$ are small over much of the global ocean up to depths of 500-1000m and away from the poles.

Our main (and only) conclusion that concerns causal links between $[O_2]^{sat}$ and $[O_2]$ is that there is a component of $[O_2]$ change that is solely warming driven and mediated by the effect of temperature on solubility. This is quite trivial as a result, and it is already well established (e.g. Kwiatkowski et al. 2020, section 3.3 and fig 3).

For these reasons, we believe that our approach is justified, despite some limitations that we will better highlight in the paper to respond to the concerns of reviewer 3.

In the manuscript we improved the methods section by explicitly stating the hypotheses behind our method (and other metrics such as AOU), its limitations and potential pitfalls, and the reason why the method is still valid in the case of the NWES (lines 247:255). Then, in the discussion, we mentioned again the method's limitations and cited some existing alternative metrics (TOU, Ito et al. 2004, EOU, Duteil et al. 2013) that overcome them (lines 542:551).

MINOR COMMENTS

REV3:

The authors notation is sometimes hard to parse and confusing, particularly when triple subscripts are used. I would recommend using a different notation, which I hope helps clarifying the comments presented here. (Note that the authors' notation is already different from the preceding work by Wakelin et al. (2020).) I recommend using a simpler symbol, such as f = $[O2] / [O2]^{sat}$ for the saturation state. Below I use the "0" subscript for "at t0", i.e., the 1990–2019 average, and the "1" subscript means "at t", i.e., the 2070–2099 average, so that for any quantity X, its 21st-century change is denoted by $\Delta X = X1 - X0$.

ANSWER:

While we recognise that other possible notations are in current use, the notation we use is at least partially consistent with published literature (e.g. Kwiatkowski et al. 2020, "ΔO_{2sat}", Ito et al. 2004, "$O_{2,sat}$", Duteil et al. 2013, "$O_{2\ pre}$").

We believe that SS for "oxygen saturation state" may be quite immediate for readers (we changed this from SS_{O2}), same for "t0" for "at time 0" and "t" for "at time t". As for the difference between our notation and that used in Wakelin et al. 2020, we changed it because we believe the notation in Wakelin et al. 2020 may have been confusing (e.g. DOs for "oxygen saturation state", dimensionless, and DO for "oxygen concentration", concentration).

REV3:

"comment on the product rule" see online

ANSWER:

Reviewer 3 is correct in pointing out that "$\Delta O_{2,other}$ arbitrarily combines a 1st-order term with the 2nd-order term" and that the second order contribution, $\Delta f \times \Delta[O2]^{sat}$, should be quantified separately. We modified the methods section, and the results, were relevant, to account for this more rigorous approach (lines 256:284). This includes computing the second order term.

However, the main body of results are not based on the ΔO_2 decomposition, but on correlations between $[O2]^{sat}$ and f with other variables, and it can be demonstrated how, given any variable X,

$\mathrm{corr}(X, f0 \times \Delta[O2]^{sat}) = \mathrm{corr}(X, [O2]^{sat})$, and

$\mathrm{corr}(X, \Delta f \times [O2]^{sat}_0) = \mathrm{corr}(X, f)$

hence the main bulk of our results still stands. We however amended the description of the decomposition of Δ[O2] in the methods and presented the separate contribution of the second order term in the results. The contribution of the second order term turned out to be negligible.