Reply on RC2
James R. Christian et al.

Author comment on "Ocean biogeochemistry in the Canadian Earth System Model version 5.0.3: CanESM5 and CanESM5-CaOe" by James R. Christian et al., Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2021-327-AC2, 2022

First of all, I would just want to make clear that I think that this sort of publication is valuable for documenting in detail the performance of major components of ESMs, especially where, as here, the authors make an effort to cover the model end-to-end, as well as compare with peer models. Given the broad span of tracers, processes and geographical / historical patterns, there is no natural end to where such analysis should begin or stop. That said, I have a number of major criticisms of the current draft of this manuscript:

While the focus is on marine biogeochemistry, it is remiss not to include information about the performance of the physical model that underpins this. The reader has no information on how well this latter model performs in terms of surface properties, mixing and ventilation and interior circulation. This need not be exhaustive, but some material on this (even if only summarising from another evaluation manuscript) seems important. Especially where this has a potential bearing on biogeochemistry performance.

We thank the reviewer for a thorough and constructive review.

A basic evaluation of the physical ocean model was presented in the overview paper by Swart et al. (2019). This includes comparisons to observations of SST, SSS, SSH, zonal mean T+S, sea ice extent and volume (seasonal cycle), and spatial distribution of sea ice in March and September. The MOC is shown as depth-latitude plots (Atlantic, Pacific, global) as is the integrated meridional heat transport (with observation based estimates at a few discrete latitudes). We believe that this analysis is adequate for the present purpose, but we did not do a good job of drawing the reader's attention to it. This has been corrected in the revised MS. In addition, we believe that some of our analyses (e.g., of oxygen distribution) are useful diagnostics of the performance of both the physical and the biogeochemical models. This was not presented clearly in the initial draft, and is addressed in the revised manuscript according to concerns raised by this and other reviewers.

The evaluation itself seems almost arbitrary in its choice of targets and, in particular, the order in which these are introduced and discussed. For me, oxygen and carbonate
chemistry parameters are essentially “downstream” of the main drivers of biogeochemistry in the ocean. Nutrient cycles, productivity and carbon / alkalinity seem much more important to first-order patterns. And this leads to oddities in the manuscript, for instance where patterns of oxygen biases, likely due to “upstream” biases in production, are discussed prior to anything about these likely sources. Also, the focus on export production instead of primary production is rather odd.

A wholesale restructuring of the Results would be difficult to achieve within the time frame available. We have tried to accommodate the reviewer's perspective as far as possible, and have added some text to more clearly explain why we chose to structure the paper in this way.

A key objective of the ESM intercomparisons is to evaluate the effects of climate change on the distributions of major tracers like oxygen, DIC, alkalinity and nitrate. And the major tracers are better observed: gridded data sets are available over the full ocean depth, which is important for evaluating models that take thousands of years to spin up. For biogenic particulates, satellite surface chlorophyll and POC are the only reliable global data sets, and even these have limited utility for validating coarse resolution global models (e.g., very high chlorophyll in coastal regions, associated with processes not resolved by the model). Realism with respect to plankton distributions and productivity are necessarily limited compared to ocean-only hindcast models or higher-resolution regional models. The ESMs are needed to provide boundary conditions for climate downscaling experiments with such models, and the requirement for such boundary conditions is limited to the slowly-evolving major tracers.

Latterly, the manuscript looks at historical trends in anthro CO2 uptake, export production and ocean anoxia. While these are valuable to look at, the manuscript offers little by way of explanation for them. These aren’t easy issues to tackle, and are probably beyond the scope of this manuscript, but it seems remiss not to include analysis. For instance, it might be informative to show whether there are spatial components to these trends, or links to physical phenomena. In the specific case of export production, the authors allude to long-known issues with increasing stratification and decreasing primary production, but show no evidence of either (further, I think the physics might be the same in both models).

We have tried to address this in concert with other reviewer comments, particularly with regard to the structuring of the various sections (I-M-R-D), and more discussion of the physical mechanisms underlying some of the results presented.

Finally, on the model naming side, while the manuscript eventually settles down, it is confusing at first about the identities and compositions of the CanESM models under consideration. The introduction and, especially, the abstract are messy on this front, and will likely confuse readers. A clear and simple statement

The ending of this comment seems to be missing, but it is fairly clear what the reviewer is trying to say. This issue was raised by several reviewers. We acknowledge our carelessness in this respect, and have tried to make the terminology consistent throughout the revised MS.

In addition, I have a number of specific comments on details of the manuscript and include these below. I should add that some of these reflect my own style / presentation preferences, and the authors should not feel obliged to address these if they disagree.

Overall, while I appreciated much of this manuscript, I judge that it requires major
revisions before it can be accepted.

Specific comments:

Ln. 19-22: This opening is confusing; which models are being examined here?; if there are more than one, and it sounds like there are, say this in the opening sentence

**We apologize for the vague wording. We have clarified it in the revised MS.**

Ln. 24: a brief mention of the aspects of the model being intercompared would be useful (e.g. nutrients, carbon, productivity, etc.)

*added*

Ln. 33: re: export decline - why?; is this explored in the manuscript?

**This result was shown in the paper (line 677-678), although the underlying mechanisms were not explored in detail.**

Ln. 33-35: re: plankton - is this worth mentioning in the abstract?; it's without context and isn't clear whether it's something that can be compared to observations

Ln. 35-36: re: phytoplankton - is this worth mentioning in the abstract?; again, there's just no context here - e.g. is this different between the different model versions considered?

**These results can not be directly compared to observations, but we think they are important in terms of understanding how the dynamics of the plankton community work in CanESM5 and CanESM5-CanOE. We have reworded the text slightly to clarify this.**

Ln. 39: if you're going to put the specific numbers in the abstract, you should also include some reference to the observational estimate (or range)

*done*

Ln. 39: re: anthro CO2 uptake differences - why?; is this explored in the manuscript?

**Again, this result was shown in the paper, although the underlying mechanisms were not explored in detail. As this is a key, and often cited, diagnostic of CMIP model performance, and is detailed for CanESM5 in the CanESM5 overview paper by Swart et al. (2019), we think the relative magnitude of uptake in CanESM5 and CanESM5-CanOE is of interest to the reader.**

Ln. 41: some of this material seems more appropriate for the methods section than the introduction

**For this type of paper, we believe that it is important for the reader to understand the historical context of the model development. We have moved some of the more detailed passages into the Methods.**

Ln. 41: more generally, the manuscript makes some specific choices on the model components and processes analysed, but does not articulate the science cases for these; the introduction is the place for doing this

**We added a section at the end of the Introduction to address the reviewer's**
general comment about the choice of data fields to analyze and the order in which they are presented.

Ln. 48: just out of interest, why not v3.6_stable?; that's a common CMIP6 configuration; (although having a different version isn't necessarily an issue - it's more variety in the CMIP6 ensemble)

We worked from what was available at the time when we began adapting NEMO for our ocean model. As some in-house parameterizations of processes in the physical ocean were implemented and coupling to the atmosphere was underway, there was not time to upgrade the ocean to NEMO3.6 when it became available.

Ln. 50: this is PISCES-v2?

No, we worked from PISCES-v1 which was what we took as a point of departure in NEMO3.4. The relevant processes did not change between PISCES-v1 and v2, and the description in Aumont et al, 2015 is an accurate description of the process model we used.

Ln. 66: how does tracer number relate to compute cost?

The total computational cost scales approximately linearly with the number of tracers which (as discussed below) is one of the reasons for implementing the additional OMIP-BGC tracers in CanESM5 rather than CanESM5-CanOE.

Ln. 66: the relationship between CMOC and the models here is unclear; how does this relate to CanESM1, 2 and 5?; this should be unambiguous

We apologize for the confusion in the terminology. We have tried to make it clear and consistent through the revised MS. We refer to CMOC in reference to CanESM1, 2 and 5 because the biological process models are identical in each case.

Ln. 69: re: CFCs - if the physics is the same between the models, you should only need to run the CFCs and SF6 in a single simulation anyway; they are non-interactive with the BGC

Yes, that is exactly what we meant. We ran these tracers in CanESM5 because the cost of CanESM5-CanOE with its much larger suite of tracers is already large. We have revised the wording slightly to make sure this is clear.

Ln. 71: re: “prohibitively expensive” - this isn't clear as many of these tracers are single tracers; or are you thinking that you might want to duplicate all of your BGC model tracers to have parallel reservoirs of natural C and C14?

No, we simply meant that we were trying to avoid the incremental cost of these additional tracers (even if there are only a few) on top of the already large cost of CanOE.

Ln. 74: does CMOC not have a version number?; seems odd as the version here is different (via O2) from earlier versions; not necessarily meaningfully (since O2 is a "downstream" tracer), but this does suggest at least a code difference

Ln. 74: there would be much less risk of confusion here if the different versions of CMOC here were identified with version numbers
Yes this would be good, but assigning version numbers (and placing the source code in the public domain) is a recent innovation. In the early years, our version control was rather ad hoc. All published simulations with CanESM5 (CMOC) have a version number because CMOC is the biogeochemistry model in CanESM5, but CanESM1 and CanESM2 did not.

Ln. 80-81: this specification of Fe limitation as being calculated from surface nitrate could do with a bit of justification; presumably this has been given before in a previous outline of the model; please include this here

Yes this is described in detail in Zahariev et al (2008), but a brief explanation has been added.

Ln. 95: re: 100% burial - is this true regardless of the saturation state of the seafloor?; i.e. above / below CCD

Yes this is the case, because when we originally developed CMOC we did not solve the carbon chemistry in the subsurface layers and were primarily concerned with global conservation of alkalinity. One of the things we have tried to achieve with CanOE is to make burial dependent on saturation state, although in the current version we implemented only a rather simplistic representation of this.

Ln. 99: maybe referring to CanESM5-CMOC might be better than CanESM5 on its own

Again, we apologize for the confusing terminology in the original draft. We sometimes use the terms as we use them among ourselves, and were a bit careless about how they were used in the initial submission. Referring to CanESM5-CMOC is not appropriate because CanESM5 (which has CMOC as its biogeochemistry) is an official CMIP6 name. We submitted data from two models to CMIP6: CanESM5 and CanESM5-CanOE. So we can not change these names at this stage.

Ln. 105: I would have expected some sort of summary about the physical ocean (and sea-ice) model used here; including resolution and major options selected; especially as the common NEMO configuration v3.6_stable has not been used here

As noted above, a basic evaluation of the physical ocean model was presented in the overview paper by Swart et al. (2019). We have revised the text to make this clear to the reader.

Ln. 117: modified from what?; do you mean from previous versions of this model, or from elements of PISCES?; the latter is implied, but it would be useful to have a "... modified from corresponding PISCES components to varying degrees."

Sorry, this was vague. It has been rewritten for clarity. What we meant was that we started with the basic code structure of PISCES and replaced the biology with our own (CMOC and CanOE), and the carbon chemistry and gas exchange with that specified for OMIP-BGC by Orr et al. (2017).

Ln. 118-119: and now we have another name for the model, NEMO-CMOC; some standardisation of naming would be helpful

Again, we apologize for the careless terminology. We refer to NEMO-CMOC to distinguish it from the original CMOC which had the same process parameterizations but an entirely different ocean model. The terminology has
been cleaned up and made consistent through the paper.

Ln. 121: does the model use the preferred carbonate chemistry of Orr et al. (2017)?; MOCSY

Yes. We do not use mocsy per se but all of the equilibrium constants etc. are identical to mocsy with the OMIP-BGC specified options. We took the original PISCESv1 carbon chemistry and replaced whatever was not consistent with Orr et al.

Ln. 126: this is ambiguous; previously it's implied that all calcite reaching the seafloor is buried, whereas this line implies otherwise

One refers to CMOC and one to CanOE. This has been clarified.

Ln. 131: re: phytoplankton functional types - ... which could be named here

Modified to specify large and small phytoplankton

Ln. 130-137: this summary paragraph of the model focuses on the phytoplankton, and doesn't mention other components, e.g. zooplankton; that tends to imply they're the same as before; anyway, this would be a good place to introduce other elements

Yes, good point. This paragraph has been modified to emphasize that CanOE has multiple size classes of zooplankton and detritus as well as phytoplankton.

Ln. 142: are we getting an explanation of what these size categories are meant to represent?; e.g. prokaryotic vs. eukaryotic

We do not think it is appropriate to associate the size classes with specific taxa with only two groups. There are places in the ocean where the small phytoplankton (strong grazer control and limited biomass range) are overwhelmingly (mostly prokaryotic) picophytoplankton, and regions where they are predominantly eukaryotic nanoplankton.

Figure 1: as already noted, I'd suggest properly introducing all of the elements of this model in section 2

Done as noted above

Figure 1: for ease of comparison, and especially as it is not described here, it might be an idea to include the corresponding schematic of CMOC;

The CMOC schematic is in at least two existing publications. In the revision we will include it in the Supplementary for ease of access.

Table 1: might it be worth noting where parameters have a corresponding parameter in CMOC, and whether the values are the same?; some model processes certainly overlap

Yes, but there are relatively few parameters that fit in this category, and it leads to a proliferation of footnotes (e.g., to indicate that K_NiX in CanOE is only approximately equivalent to K_DIN in CMOC). We believe that the interested reader can easily work this out for him/herself.

Table 1: re: parameter k_Ca - does CaCO3 dissolve above the CCD?
Yes. Sinking CaCO3 is subject to first-order dissolution at all depths. This is clarified in the revised MS. It is well documented that dissolution above the saturation horizon occurs (e.g., Milliman et al 1999, DSR I 46: 1653), although clearly there is a saturation-state dependence of the rate (e.g., 10.1002/2013GB004619), and the mechanisms are not entirely understood. The basic version of CanOE does not include saturation-state dependence of water-column dissolution because we wanted a 'base' version to branch out from to experiment with ocean acidification feedbacks to the carbon cycle.

Table 1: re: parameter K_NH4ox - might it be better to describe K_NH4ox as a maximum nitrification rate, which is then diminished by an irradiance function with a half-saturation, K_E

OK. Possibly just 'Nitrification rate constant' is more appropriate; it's really a rate constant, not a rate. (We also changed the K to lowercase to make it consistent with equation 20.)

Table 1: re: parameter K_NO3 - presumably N2-fixation occurs in ignorance of PO4 availability?

Yes. DNF is dependent on light, temperature, dissolved iron, and DIN. We did not include a parameterization of PO4 limitation (although we have developed one on an experimental basis).

Ln. 160: explain what's going on here with NH4 and NO3; does this functional form have a source describing it?

This is a commonly used formulation for NH4 inhibition of NO3 uptake, although possibly written in a unfamiliar way. It reduces to the formulation based on noncompetitive inhibition (Frost and Franzen 1992 MEPS 83: 291; Hood and Christian 2008 in Capone et al eds "Nitrogen in the marine environment").

Ln. 169: is E irradiance?

Yes. We added this to the text. Sorry.

Ln. 175: re: C_XS - this term needs expansion (or a reference to a later equation number if it appears below); I'm uncertain what you mean by this, or what ecological process it's meant to represent; equation 16a seems to be the right one

Added reference to equation 16a where the symbol first appears.

Ln. 180: "excessively low" in a model stability sense?; or is there an actual observed threshold here?

This is simply a device used to prevent biomass from declining to levels far below the 'seed' population required for realistic biomass to accumulate in spring-summer, even under the most favourable growth conditions. We do not know exactly what limits the losses in the real world, but we know that something does. In any case phytoplankton linear mortality terms are a very inexact representation of any real process. First-order mortality leads to biomass declining to levels that make it impossible for the population to meaningfully recover in the brief Arctic summer. We have expanded the text here a bit to make this clear.

Ln. 204-205: how equations 13a and 13b fit into equation 11b is unclear, especially as
equation 11b refers only to $G_L$, which seems to be calculated in equation 12b; the latter point also occurs for small zooplankton.

Large zooplankton graze on both small zooplankton and large phytoplankton, so it is necessary to partition large zooplankton grazing between the two groups in order to calculate the loss of small zooplankton to grazing. We have altered the text slightly to try to make this clear.

Ln. 208-216: might the clarity of this section be improved by the addition of a diagram that quantitatively illustrates the scale of excess C, N (and possibly Fe) over a span of intake C:N?

Table 2: this kind-of answers my point before about a diagram, although a diagram might still be better (if more difficult to create)

We prefer a table. It presents the critical information in a concise way.

Ln. 241: has the impact of forcing a common zooplankton C:N on detritus compared to dynamic C:N in phytoplankton been explored at all?; does this mean that the majority source of both detritus classes is zooplankton?; it seems odd to make a fuss about C:N in phytoplankton only to entirely overlook the C:N of the more heterogeneously-sourced detritus component.

No, detritus originates from both phytoplankton and zooplankton, and under some conditions it comes predominantly from phytoplankton. It is true that the recycling of 'excess' phytoplankton C or N into the dissolved pool is probably unrealistic in some cases, but it is necessary to maintain mass conservation. Not including variable C/N in detritus was a purely pragmatic choice that was made to limit the number of tracers.

Ln. 263: ah-ha; E = irradiance

addressed above

Ln. 265: does the absence of PO4 in the model cause any problems for this N2-fixation scheme?; low NO3 is often associated with low PO4

Yes but there was no realistic way to parameterize this. For carbon chemistry we followed Orr et al at assumed that the PO4 contribution to alkalinity can be estimated as DIN/16. But N2 fixation tends to be associated with large departures from the N/P Redfield ratio that occur under extreme oligotrophic conditions. N2 fixation models are still at a rather primitive stage of development. CanOE is a step forward over CMOC in that it at least includes Fe limitation, whereas in CMOC N2 fixation tends to grow without bound in a warming ocean (Riche and Christian 2018).

Ln. 284-285: does this scheme produce large-scale spatial patterns in calcite production that match the general high-equatorial, low-polar pattern?

Yes, calcification occurs predominantly in the low latitudes in CanOE. But we believe that most CMIP5 models (including CMOC) overestimated the degree to which rain ratios decline with latitude or temperature (see Eq 13 and Figure 1 of Zahariev et al (2008)). Honjo et al (2010, Progr. Oceanogr. 85: 137) show that Arctic rain ratios are similar to the global mean, although in some cases they are very low (probably associated with diatom blooms). Some of the worst misfits of CMOC with the regional mean rain ratios estimated by Sarmiento et al (2002) are
due to the (probably excessively) strong dependence of rain ratio on SST (e.g., the subarctic Pacific).

Ln. 290: again, it's implied earlier that 100% of calcite is buried, but this suggests otherwise

**No. 100% of calcite is buried in CMOC, but not in CanOE. The loss of alkalinity to burial is treated in the same way (reintroduced at surface in the same vertical column).**

Ln. 291-292: is this localised spatially?; i.e. loss at the seafloor is added at the surface immediately above

**see previous point**

Ln. 296: some expansion here on the precise links between processes would be helpful; e.g. NO3 vs. NH4

**added**

Ln. 317-318: this is a little paradoxical; the closer a seafloor tile is to sources of O2 (surface productivity and the atmosphere), the less oxygenated the sediments; this presumably reflects the supply of organic matter to the seafloor and the resulting oxygen demand; if this is the logic, make this clear

**Yes, it is based on the greater organic deposition and presence of reducing sediments at shelf depths. The text has been expanded a bit to make this clear.**

Ln. 328: this also implies that particles can scavenge iron continuously without saturation; I don't imagine this is a problem, but it might make the model's behaviour in areas dominated by slow or fast sinking detritus interestingly different

**Interesting point. No we did not consider such 'saturation'. But it is unlikely to be a major factor. Possibly a topic for a future experiment, although on the list of oversimplifications in our scavenging model it probably ranks fairly low.**

Ln. 332-340: sensible; I like this

**Thanks**

Ln. 342-352: this could be clearer and sourced to relevant work on the topic; Wolf-Gladrow et al. (2007) (which you cite earlier) suggest +1 ALK for N2-fixation to NH4+, -2 ALK for NH4+ to NO3-, and +1 ALK for denitrification of NO3-; here, assuming N2-fixation goes to NO3-, this implies -1 ALK for N2-fixation and +1 ALK for denitrification; anyway, the text here is ambiguous, and should be straightened out and sourced

**All of the sources and sinks of alkalinity associated with N cycle processes are detailed in Table S2. In an earlier draft this table was included in the main text, but we were afraid that reviewers would dismiss it as a reiteration of well-known information. This paragraph is important to make clear to the reader how alkalinity is conserved globally given that both N2 fixation and denitrification are prognostic.**

Ln. 354: links for the data?; and access dates; some of these products are revised periodically
The versions of each data product used are stated. As new versions of these data products are released periodically, existing versions are unlikely to be altered.

Ln. 360: this is GLODAPv2

added

Ln. 360: which offline carbonate chemistry calculations are needed?; GLODAPv2 includes pH

We do all carbon chemistry calculations offline to make sure that it is done in a consistent way across models and observations. We have tested this many times (e.g., offline Ω_A vs Ω_A output from the model) and the differences are minor.

Ln. 363: rephrase to "... were used for the absent tracers, phosphate ...

done

Ln. 382: why 2x2 degree?; the 33 levels is more understandable

We regridded all data to a uniform grid for easy comparison across models. For the global scale patterns that we are primarily concerned with here, there is no meaningful difference between gridding at 2x2 or 1x1 degree. For example, when a 1x1 grid is used, none of the correlation coefficients for CanESM5 vs observed oxygen on the six depth levels shown in Figure 4 changes by more than 0.005 (max 0.0028, mean 0.0011).

Ln. 382: how regridded?; linear, nearest neighbour, etc.?

This is specified in the revised MS.

Ln. 383: technically, GLODAP follows WOA (which did this vertical grid first)

Clarified in the revised MS.

Ln. 386: I think ignoring variability across the CanESM5 ensemble is not an unreasonable assumption, but it might be useful to support it for this particular model with some evidence; e.g. a plot of some key property (e.g. NPP, CO2 flux, SST, etc.) across the ensemble for, say, the decades of interest here; this could be put in supplementary if it breaks the flow

This is a good idea. We will include a Supplementary table that illustrates the magnitude of historical trends relative to internal variability.

Ln. 394: it would be helpful to name (and source to descriptions / evaluations) the CMIP6 models used in this analysis within this section

added

Ln. 396: while it may have been done elsewhere, some sort of outline of the performance of the physics model seems necessary to me; even if it's cursory and largely points to this other work; if there is no other work, some expansion would be useful; things like surface physics (incl. mixing), sea-ice, major circulation (AMOC, Drake), MOC would be of interest

This is addressed in the response to the reviewer's general comments above.
please be clear why you're starting with oxygen; in most models it's largely
slaved to other more dynamic model processes and tracers (which are, in turn, strongly
influenced by physics processes); if there's a good reason why you're looking at it first,
make it clear

The reasons for structuring the paper in this way are now clearly stated.

why these depths?; including a more abyssal depth might hint at circulation issues

The depths were chosen to allow the reader to assess the models' representation
of the spatial distribution of low-oxygen waters, particularly in the Pacific. These
regions and depths contribute disproportionately to the global model-data misfit,
and we think it is important for readers to understand how well, or poorly, the
models are representing the underlying processes.

Abyssal depths are much less diagnostic of these specific processes (formation
and maintenance of the oxygen minimum zones) although, as the reviewer
notes, there are issues with ventilation of the deep ocean in some models. This
issue is discussed briefly in the revised MS, but really deserves a paper all by
itself.

for a number of reasons, I would not expect to MEM to be a good comparison; do
you know how it compares to observations relative to the performance of its component
models?

The idea that a MEM outperforms individual models (Lambert and Boer, 2001)
has proved over time to be quite robust, and it is generally true for all sorts of
fields including ocean biogeochemistry fields (see e.g., Figure 22 of Chapter 5 of
the AR6 (WG1) report). It can also be deduced from the data shown in the paper
itself as noted by the reviewer below (comment on Figure 8).

Figure 2: with fewer colours in this scale, it would be easier to discern differences between
the models

Agreed, but the range of concentrations is quite large at these depths. It is
possible to shrink the range of the colorbar slightly, and we have done so.

you're inferring these as "circulation features" but haven't reported on your
model's circulation at all (e.g. MOC)

As noted above and in the responses to the other reviewers, we have tried to be
more clear and specific in our references to ocean circulation processes and refer
specifically to the analysis presented by Swart et al (2019).

"the ensemble mean" = "MEM"

We were careless in our terminology here and have made it consistent
throughout the revised MS.

it's old-fashioned of me, but would a profile of O2 further assist here?; possibly
not given its spatial heterogeneity, but a series of vertical Taylor diagram slices is a little
hard to take in!

We accept that there is a bit of an 'information overload' factor here, but we
believe that the multiple Taylor diagrams for different depths are a quite
powerful way of visualizing model skill. Clearly it only tells you the magnitude of the bias and not its spatial distribution. But it contains a lot of information about how the model is performing that is lost if we do e.g., a single Taylor diagram over 3D space. Clearly we are assuming that the reader has a certain level of familiarity with ocean circulation and the three-dimensional distribution of major tracers; we have tried to make it a bit more clear what we diagnose about model performance from these figures. We believe that Figures 2 and 3 give the reader adequate information to visualize the vertical distribution (although as the reviewer notes above, it excludes the abyssal depths).

Ln. 444-450: how does this relate to any hard-wired limits in models?; the model I use, for instance, is prevented from consuming oxygen below a limit

Our model does not have such a hardwired limit, but shifts respiration from O2 to NO3 below 6 uM. We will review the other model description papers for discussion of such limits.

Ln. 461: re: "much deeper" - why?; and does this relate to the abyssal issue I raised re: oxygen?; i.e. this is an interesting depth

We chose the depths for O2 because we believe that the processes that create and maintain the OMZs are of interest to readers. We chose a more abyssal depth for saturation because there is quite a lot of variability among models in terms of transporting DIC and alkalinity to the deep ocean.

Figure 5: might a table be better for this information?; maybe combined with other measures of model performance?

We think that the bar graph provides a compelling visual illustration of the key points we are trying to convey here.

Figure 6: too many colours here makes it more difficult to discern differences between the panels

Again, the ranges are large. In this case we used separate colour scales for the different depths in order to constrict the ranges as much as possible. We will reassess whether it is possible to shrink them further but the change is unlikely to be large.

Figure 6: might the depth at which omega aragonite hits some threshold (value 1 would be most obvious) be better?

Maps of the depth of the saturation horizon have been added to the Supplementary material.

Ln. 471-472: again, remineralisation is mentioned in the context of biases before anything about production and export is introduced; omega is a downstream variable, so the ordering of the analysis here is perplexing

A few comments about the productivity of the overlying waters have been added here. The question of the overall ordering of the presentation of Results was addressed above.

Figure 7: how reliable are the observations here?; GLODAP is much less data-rich than WOA
For the kind of global-scale analyses we are primarily concerned with here, the gridded data products are quite reliable. It is instructive to consider how little the gridded data product changed between GLODAP 1 and 2 (except in the Arctic of course). This can be in large part attributed to the foresight of the people who designed the first global survey (e.g., Feely et al., 2001, 10.5670/oceanog.2001.03).

Figure 7: geographical plots of surface omega, seafloor omega, and the depth at which omega hits some threshold would seem more valuable to me; and easier to compare between models - these are very similar looking plots whose differences are not easy to discriminate

We think these plots give the reader an indication of whether the models are doing a good job of representing the large-scale distribution of DIC and alkalinity, and their effect on the saturation state, which is a commonly used diagnostic of biological and geochemical impacts of anthropogenic CO2. We agree that there is a certain amount of redundancy here as OmegaA, OmegaC and [CO3\-\-] are determined by similar processes. Note that in response to other reviewer comments we will be replacing these plots with ones that show only the observed distributions and the model anomalies relative to it, which probably brings out the differences a bit more starkly.

As for the additional maps suggested, we are evaluating these for inclusion in the Supplemental.

Figure 7: do you need both aragonite and calcite?; the model seems to use calcite only

The model assumes calcite for purposes of calculating burial/dissolution at the sediment/water interface, to avoid introducing unnecessary and unconstrained complexity. The saturation states of both minerals are determined by the distribution of DIC and alkalinity, regardless of what assumptions the model makes about the solid phases, and are of interest from the perspective of biological impacts and climate feedbacks.

Figure 8: actually, I take my earlier comment back, the MEM is pretty much always better than the individual models

See above

Ln. 496: as the N and Fe cycles regulate productivity and therefore ocean interior remineralisation and DIC/ALK, it would perhaps make more sense to discuss these ahead of the more downstream oxygen and carbonate chemistry properties

This is addressed at the beginning in response to the reviewer's general comments.

Ln. 502: HadGEM2-ES's marine BGC included a prognostic Fe cycle; see the full description of Totterdell (GMD, 2019)

Yes we discovered this error on our own, after submission. Sorry about that. This text has been deleted.

Figure 11a-11b: it seems overkill to have both 11a and 11b in the manuscript; I'd suggest deleting 11b

Possibly we have a bit of a tropical bias, but we think that the seasonal cycle of
Equatorial upwelling and the associated HNLC condition is of interest to readers, and it is not readily discernible from Figure 11a. But more than one reviewer mentioned this, so Figure 11b has been moved to Supplementary.

Figure 12: is this scale running across three orders of magnitude?; so is it 1 nmol/m3 to 1000 nmol/m3?; if so, the labelling of this log scale differs from that of the nitrate plots above.

We altered the colorbar of Figure 12 so that it is done in the same way as Figure 11 (actual data on a logarithmic scale rather than log(X) on a linear scale).

Figure 14: why not geographical plots of DIN?

These were included in an earlier draft and left out in the interest of space.

Figure 18: worth plotting some regressions on here?; the data density means that the shape of the curves might be easier to discern then; also, why does the plot's chlorophyll appear "capped" at 1 mg / mg?

Observational data > 1 mg m^-3 were excluded because the vast majority of these occur in coastal waters and are associated with processes not resolved by coarse resolution global models. In the open ocean, concentrations > 1 are very rare. This should have been stated in the caption and the Methods. This has been corrected.

Figure 18: re: 17 mg / m3 - this seems a bad idea; why do this?; it looks like you're trying to maximise the appearance of fit

The offset is clearly stated in both the caption and the legend, and the rationale for it is clearly explained in the text. Given the processes that are not considered at all in the model, its existence is unsurprising.

Figure 19a: why crop the scale?; it's not helpful with bar charts

This was based on the assumption that starting the scale at zero makes it harder to visualize the difference among models, as the range is fairly narrow. But having created a new version with the y axis starting at 0, we think it looks good and will substitute it as per the reviewer's suggestion.

Ln. 666-673: a plot that might be helpful here is the geographical map of cumulative CO2 uptake; for instance, to identify whether the uptake pattern is the same but the magnitude different, or that there are actual differences in the spatial pattern of uptake

Ln. 666-673: another plot which might be useful here is the geographical inventory of anthro CO2 in the models (the CanESM5 ones); again to identify whether there are patterns in the differences between the models

This is a good idea. We have conducted some preliminary analysis and will include it as Supplemental in the revision. It does shed some light on the reasons for the difference in cumulative uptake between the two CanESM models, but the differences are not large.

Ln. 681: is this decline in response to stratification happening here?; I thought the models were physically identical?

The comment was simply that there has been a general consensus in the existing
literature that the trend in global total export is likely to be downward in a warming ocean, but that the trends shown in the paper are difficult or impossible to verify using observations. The wording has been revised to make sure there is no ambiguity.

Ln. 684: geographical plots of export production in the models, and how it changes between, say, 1980 and 2014 would be helpful here

We showed zonal means because we thought the plot conveys the key message that the trend is fairly consistent across regions, and the processes differ between CMOC and CanOE in a somewhat consistent way, especially in the Southern Ocean. We could include difference maps as well, but we do not think they add much.

Ln. 684: more generally, it seems strange to include these trends in export production without (a) talking about primary production, and (b) trying to dissect what the source(s) of the trends are

Figure 21: why focus on export ahead of production?; and would it be more interesting to consider export / production over this time period?

We consider export production to be a more robust diagnostic of biological impacts on biogeochemical cycles than primary production. For example, a net change in export production will affect both ocean net CO2 uptake and subsurface oxygen concentration, whereas a change in primary production does not necessarily affect either. We agree that the discussion of the underlying processes was superficial. We have expanded the analysis slightly to address both of these concerns.

Ln. 711-713: broken word here; "depend ... ent"

fixed

Ln. 768: are there runtime figures on how more costly it is?; you might expect cost to scale with complexity; e.g. Kwiatkowski et al. (2014) found ~linear relationships with tracer count

Yes, as noted above the scaling is quite linear as discussed by Kwiatkowski et al.

Ln. 798: my personal preference is to conclude a paper with a set of bulletpoint conclusions of the main findings

The Discussion has been extensively restructured as per the comments of this reviewer and Reviewer #1.

Ln. 807-809: maybe include the source ID for the model on the ESGF system together with the variant labels for the specific ensemble members included in the analysis

done