Authors' response to 2nd round reviewers' comments on "*Resolving ecological feedbacks on the ocean carbon sink in Earth system models*" (Manuscript reference: esd-2020-41)

Dear Prof. Crucifix and Reviewers,

Thank you for your second round of reviews of our manuscript "*Resolving ecological feedbacks on the ocean carbon sink in Earth system models*". Please find below our detailed responses to the reviews. We include the original comments and our response under each point in **bold** text and with line numbers referencing the revised manuscript with changes marked.

We hope we have sufficiently answered your queries in our response.

Yours sincerely,

David I. Armstrong M^{c}Kay & co-authors

Response to Reviewer 1, Anonymous (R1):

I'd like to thank the authors for addressing the majority of the points identified. I have a few follow-up points on some of the responses, and a couple of suggestions the authors might like to consider that may help improve the manuscript.

Follow-up:

- A significant outstanding omission remains the contextualising with CMIP5 (I accept the authors' stated issues with CMIP6). The authors assert that there is little information on this, but omit Fu et al. (2016)'s quite comprehensive study on this specific subject. This should provide helpful relevant context that the manuscript is currently weak on.

The headline result of Fu et al 2016 for export production closely reflects those of Bopp et al (2013) already cited (both give 7-18% EP decline in 2090s vs. 1990s, to which we add ~2% for inter-comparison with a preindustrial baseline), but we have included Fu et al here as well for completeness (lines 344-346).

- While the manuscript is now clearer on how CaCO3 production is parameterised in the model, I was a little confused about how it dissolves down the water column. Line 218-219 states …

"BIOGEM by default uses a fixed remineralisation profile similar to the Martin curve for the sinking labile fractions of both POC and PIC"

… which tends to suggest that PIC dissolution follows the Martin curve. More typically it is assumed that it has a relationship with the CCD, beneath which carbonate is undersaturated and dissolution occurs.

PIC remineralisation/dissolution follows a similar procedure to POC in this version of cGEnIE, in that it's parameterised to fit a curve according to a characteristic length scale. This is indeed not entirely typical, although the original wording wrongly implied that both POC and PIC both followed the same Martin-like Curve rather than different ones, which has now been clarified (line 221)

- I previously asked about mixing in the study's model …

"Ln. 169: does the model resolve seasonal mixing of different ecological regimes?

No, but insolation and light attenuation is seasonal, and ecoGEnIE does reproduce seasonal variation in primary production. (pg.8 ln.249 – pg.9 ln.263)"

This suggests to me that there is no seasonal mixing. However, I'm a little confused because the new manuscript revision explicitly mentions (ln. 229-230) …

"photosynthesis … subject to light limitation, photoacclimation, and seasonal light attenuation within a variable mixed layer depth"

Figure S48 also tends to suggest that there is a representation of the mixed layer, so there's perhaps just a mix-up in the description that needs clearing up.

Apologies, the original response was mistaken – cGEnIE has a mixed layer scheme based on Kraus & Turner 1967, which of course is a seasonal thermocline model. This has been made clear in the preceding section to avoid confusion (lines 215-216).

Additional specific points:

- As an aside, it would probably increase reader interest in this manuscript if the headline changes in carbon cycle processes were quantified in the abstract. This block of text could really benefit from having numbers (e.g. from Table 2) added where suggested:

We have added headline figures to the abstract as suggested (lines 18-22).

- Supplementary figures that show OBS and then comparative model deltas could be better arranged to allow readers to directly compare the plots. At present, readers must flip back and forth to compare plots making something that should be easy needlessly difficult. For example, Figures S1, S3 and S5 (and S2, S4 and S6) should be rearranged onto the same page so that readers can easily compare patterns of bias. This should be repeated throughout the supplementary material. Also, Figure S13 shows the model-obs difference, but without showing the observations, so the reader cannot tell how large the biases really are.

We have reorganised the relevant supplementary figures to group maps and depth plot model-observation comparisons together. However, fitting 3 plots per page would significantly reduce the figure resolution and make some features difficult to see, and the reorganisation significantly reduces the need to flip back and forth, so we have kept the plots at full size. We have also added the SeaWiFS observational data in a new figure prior to Figure S13.

Response to Reviewer 2, Jamie Wilson (R2):

Armstrong McKay et al., have put a lot of effort into revising the manuscript and responding to the reviewer comments, especially given the varied pressures of working during a pandemic. Overall, I think the manuscript is improved with the addition of adequate details about the experimental set up, new analysis of results and improved general presentation. The overall concept and findings are interesting to the Earth system community and are currently very relevant to a wider discussion of resolving plankton ecosystems in Earth system models.

Unfortunately I think there are still some outstanding issues with the experimental set-up and analysis that need to be resolved before I can recommend the manuscript for publication. Overall, I think the problems described in the comment by Kvale et al., combined with my own comments highlight a broader issue: can a common baseline be defined for different models and does this limit the level of quantitative analysis? This seems like a tricky issue to resolve but I have tried to outline two ways in which I think it could be achieved below as I think this is worth pursuing. I have also included a more detailed criticism of the recalibration process now that the process has been fully described, but this forms a part of the broader issue.

It is indeed a tricky issue, and as discussed in this and the previous revision round neither recalibrating nor directly comparing different configurations fully gets round the issue. We agree that it needs ameliorating and discussing, and outline how we do so now below.

Experimental Set-up

Based on the previous reviewer comments and the clarifications in the revised manuscript, I think there is a fundamental issue in trying to define a common baseline for all the experiments. It seems essentially impossible to get a common baseline that is *exactly* the same using different biogeochemical and ecosystem models despite them sharing the same physical model. In matching one variable, such as POC export, there will always be a subsequent trade-off in another (POC remineralisation, surface carbonate chemistry) that will have an impact on the results (nutrient delivery timescales, carbon sink). A key question here is whether the findings are robust to this issue. Figures S61 and S64 show some insight into this question. Figure S61 shows some level of agreement in the trends of POC export across different baselines but disagreement in transient behaviour and magnitudes. Figure S64 shows disagreement in both magnitude and sign of the trends in the ocean carbon sink. However, trends in S64 are relative to the BIO+FPR run not the individual baseline as in S61 so it’s hard to tell whether the presentation choice is a factor here.

Unfortunately I think this is a difficult issue to deal with. I have two suggestions for resolving this:

1) The issue could partially be resolved by presenting changes relative to the corresponding baselines as is now done for POC but not for the ocean carbon sink. Then, at least, there is a clear distinction between experiments. There needs to be additional discussion that details that the response of models with varying complexity has two components: a dynamical response to environmental change driven by the model itself, and a dependence on the initial state that is inherent in using different models. The downside of this approach is that it really limits the findings to more semi-quantitative comparative descriptions because it’s very difficult to separate out the impacts of the dynamical-responses from the initial state. It’s not obvious even that this would be consistent

across the experiments. In my opinion, the recalibration process used adds additional biases (see comment below) and is arguably not necessary if the experiments are compared to their own baseline anyway. I would therefore strongly suggest presenting the results using the default versions of the model. This actually facilitates a broader discussion that has more relevance to the wider modelling community given that models are replaced by newer versions and assessed against a broad range of metrics, e.g., Seferian et al., (2020).

2) To resolve the baseline issue completely the experiments need to be run using the same model set up for each experiment. The ECO+TDR model should be used to create a single preindustrial spin-up. The impact of temperature-dependent remineralisation, size-dependent partitioning of DOC:POC and non-Redfield stoichiometry can then be quantified by controlling each element. For example, the FPR experiment can be replicated by forcing the remineralisation to “see” the preindustrial temperature field, thereby causing it to behave as the FPR experiment but not deviate as a baseline. Similarly, one could control for elements of the ecosystem model such as the size-dependent POC:DOC export or stoichiometry. This would allow the authors to quantify the influence of each component to a much greater extent and reliability. However, this approach requires some adjustments to the model to enable this and extensive revision of the text.

Approach 2) would indeed give a thorough experimental basis, but there is the issue that ecoGEnIE itself (i.e. ECO+FPR) is yet to be fully recalibrated (and is still missing some processes). Because of this, using the ECO+TDR spin-up as the basis for all runs as suggested would still have limitations affecting the robustness of specific numbers.

We have chosen to follow approach 1), and have updated figures and the text throughout the manuscript to reflect this (e.g. lines 289-310, 359-404, 435-459, 472-483, 510-538, 545-550, 607; Table 2; Figures 2-5, S48, S62, & S67). As the reviewer notes, this means our approach is indeed more semi-quantitative – a point also relevant to the missing calcifier/silicifier classes and other important processes mentioned in the Discussion. Nevertheless, the trends that emerge from across the different calibrations despite these limitations (relative biological pump strengthening and weakening for activating TDR and ECO respectively, translating to relative ocean carbon sink weakening and strengthening in turn) is a useful outcome that demonstrates the relevance of including ecological and metabolic dynamics in Earth system models.

We have also replotted the absolute ocean carbon sink capacity as requested, however we have presented it alongside the original relative plot rather than instead of. This is because unlike for POC export we cannot plot Air-to-Sea CO2 flux as a % relative to preindustrial baseline (as by definition the ASG flux at preindustrial was close to zero), and in the absolute cumulative ocean carbon plot it is hard to tell apart some of the curves. As a result, we now show the absolute cumulative ocean carbon plot as context as well as the relative cumulative ocean carbon plot to show the detailed differences between the scenarios.

Biological Pump - Recalibration and Interpretation Issues

The authors have now fully detailed the recalibration process involving POC remineralisation. I understand the justification for recalibration but I think this adds additional biases to the findings that are not quantified or even acknowledged in places. The crucial issue here is that the authors achieve the same global POC export production across the different model set-ups by altering the fraction of export that is remineralised as refractory POC. In GENIE, export is divided into “labile” POC (~95%) that attenuates strongly across the upper 1000m and “refractory” POC (~5%) that attenuates minimally with most POC remineralising in the grid-boxes overlying the seafloor boundary. The authors defend the calibration by stating that “biological pump perturbations on sub-overturning timescales (<500-1000 years) will not significantly affect surface DIC…” (lines 295 - 300).

I strongly disagree with the author’s defence. There is an average characteristic lifetime of regenerated DIC (and correspondingly, nutrients) that is a function of ocean ventilation times (First Passage Time: Primeau 2005) and remineralisation rates. By lowering the labile:refractory export partitioning the authors are increasing the average lifetime of regenerated DIC and nutrients in the ocean. I agree that the ventilation time of DIC from the deepest ocean to the surface is predominantly longer than the timescales analysed, but, this is compensated by reducing the amount of regenerated DIC entering the intermediate ocean where the ventilation timescales are relevant to the timescales analysed.

To demonstrate this as an issue, I have run some idealised experiments in an offline transport-matrix based version of GENIE (in-preparation for publication based on earlier work described in Wilson et al., 2015). The circulation is diagnosed from the equilibrium annual-mean circulation in GENIE at the native resolution. The biogeochemistry model is the same as reported in Ridgwell et al., 2007 and Cao et al., (2009) which is the same as the BIO+FPR set-up used by the authors. I created two spin-ups with a simple phosphorus cycle: one using the same set-up as the BIO+FPR experiment and one where I increase the refractory export of POC to 35% as per the author’s re-calibration. Each run is then continued for 500 years with an immediate cessation of biological uptake and DOP remineralisation. The surface ocean is also subject to a zero boundary condition, i.e., supplied PO4 is removed at each timestep to isolate the ventilation of PO4 from the interior ocean. Because the circulation is static any differences in the transient response of interior-to-surface PO4 supply results from the difference in initial distribution of PO4 associated with the re-calibration of refractory export. The spin-up global mean concentrations of PO4 in the ocean interior are 2.19 µmol kg-1 and 2.21 µmol kg-1 for the default and recalibration set-ups respectively.

Figure 1 shows that the re-calibrated model does have a different transient behaviour well within the timescales explored in the manuscript. Both the supply rate (Fig. 1A) and cumulative supply (Fig. 1B) of PO4 to the surface ocean are correspondingly lower for calibrated run with deeper remineralisation. Whilst this is a simplified scenario, it demonstrates that the transient adjustment of nutrients (and carbon) in the ocean interior in response to a decrease in export production is impacted by this re-calibration. As such, the recalibration has some impact on the transient features of export production and the air-sea gas exchange of carbon that may in-part explain the differences between set-ups seen in Figures S61 and S64.

This is a good point, and we thank the reviewer for providing a model demonstration clearly illustrating the issue. Although recalcitrant POC reaching the seafloor wouldn’t impact surface waters on millennial timescales, of course a higher recalcitrant fraction would also have the effect of depriving intermediate waters of remineralised nutrients as well with consequences emerging on nearer-term timescales. We have added discussion of this issue as a confounding factor for the recalibrated configurations (which have also now been adjusted to be supplementary rather than main results) in both the Methods and Results section.

As discussed previously, either using the default or recalibrated configurations leads to different problematic confounding factors (that could be partly but not wholly dealt with by using one configuration), but the broad trends that emerge regardless and the dynamics they reveal remain as useful results that demonstrate the role of ecological complexity in Earth system models.

It is notable that changes in remineralisation are generally not considered in the analysis and discussion of the results. It is more complex with the TDR model as the transient behaviour is driven by the rate of warming across the ocean water column but it is likely that it will have some impact.

We have added more specific discussion of the impact of remineralisation changes in our discussion of the carbon sink capacity results (lines 485-508), which supplements existing discussion of remineralisation in the biological pump results (lines 362-383 & 452-456).

Specific Comments:

Line 65: “weakening of the biological pump” - this phrasing is used throughout the manuscript. Weakening and strengthening are used in various ways by the wider community from referring to export production and the total sequestered carbon (Csoft). Because you are not quantifying Csoft, these terms need to clearly defined.

We have clarified here at the first mention of biological pump weakening that we define it as reduced POC export (line 68), and have also clarified this POC export focus elsewhere (e.g. in the abstract).

Line 84: “follows a power law distribution” - this is somewhat pedantic but Cael & Bisson (2018) showed that a power law is no better a description (statistically) of the Martin Curve sediment trap data than other functions.

Edited to “power law-like” for accuracy, but keeping power law as a commonly understood function shape (line 86).

Line 140: “global deepening of 24m” - I feel like “of the e-folding depth” is missing in this sentence.

Clarified (line 142).

Lines 142 - 144 - I appreciate this was a point from another reviewer and I agree that at steady state the pump is neither a source or sink. But all things being equal (and assuming a closed system w.r.t. $CaCO_3$ sediments) a “stronger” pump, either through higher export or deeper remin), will be associated with lower atm. CO_2, e.g., the relationship between CO_2 and Cbio in Goodwin et al., (2008). I think there is a conflation between source/sink of carbon in a transient sense and equilibrium states of atm CO_2 here.

We agree that a change in the biological pump will lead to a transient carbon sink/source even if this is not the case for the equilibrium state, and that in the long run a system with a stronger biological pump would store more carbon in the ocean, but in responding to the short comment the initial point was not so clear in the manuscript. We have further adjusted this and at the start of section 4.3 to make this clearer (lines 143-146).

Lines 348 - 350: “more POC is remineralised within the surface layer” - this is a misunderstanding of what is happening in GENIE. For the FPR runs the exponential remineralisation curve is normalised to the base of the surface grid-boxes, i.e., no POC remineralisation occurs within the surface boxes. I believe this is the same for the temperature-dependent remineralisation scheme - particles sink explicitly from the base of the surface layer. There is remineralisation of *dissolved* organic carbon in the surface layer. As such, it’s tricky and potentially misleading to define new and regenerated production in this way. This section needs to be reanalysed and presented using the correct understanding of what is happening in GENIE.

You are of course correct here, and this section was poorly phrased and constructed. We have edited and rearranged the text here to focus on the effect of the remin. depth shoaling towards cGEnIE’s subsurface layers, which is what actually drives the relative increase in export in cGEnIE (lines 364-366).

Line 374 - I am struggling to follow the logic of mean cell size becoming smaller and extending the number of trophic levels. In this model trophic levels are primarily initiated by the presence of size-dependent grazing.

This explanation was unclear, in that it referred to a hypothesis by Riebesell et al 2009 that has a parallel to but in fact doesn’t directly apply to cGEnIE. We have edited this sentence to make it clearer that the primary driver of reduced biomass is stratification-induced nutrient decline, and removed the trophic level discussion as being tangential (lines 393-396).

Lines 383 - 384: “…the amount of carbon exported for every unit of phosphorus increases with warming in response to stratification, reducing surface phosphorus loss…” - PO4 is the model currency here not DIC so C is changing relative to P.

“reducing surface phosphorus loss” has been removed as you are correct in that P is not being changed here, but the overall effect ameliorating carbon export decline is still the case (as more C is lost per unit export) (line 404).

Line 387: It’s worth noting that Wilson et al., (2018) is showing equilibrium results which, though related, are not directly comparable to the transient results here.

Paragraph edited to make this clearer (lines 406-420).

Line 393: “allowed” instead of “made”?

Changed (line 413).

Line 397: there is a problem with the sentence structure.

Adjusted and split into two sentences for clarity (lines 418-420).

Lines 431 - 436: see the comment for lines 142 - 144. There is maybe a conflation between transient source/sinks and equilibrium CO2.

Extra detail added to clarify that this is the case for comparing equilibrium states such as with warm palaeoclimates, and that our critique here is targeted at misapplication to the modern transient case (lines 461-46).

Lines 450 - 463: There is no discussion of remineralisation changes here!

Lines 450 - 463: It would be useful to state that the circulation response (temperature and stratification) are the same across the experiments here.

Remineralisation changes were implicit in the description of the initial production changes impacting DIC and ALK, but could have been spelled out better as a key factor in surface DIC/pCO2 increase – this has now been clarified, along with the supplementary schematic updated to include more mechanisms and the parallels to Kwon et al (2009)'s mechanisms made explicit. We have also clarified in the previous paragraph and in the 4.1 physical climate response section at the start of the Results that the climate response is almost identical across the experiments, and so the observed differences must be biogeochemically driven (lines 485-508).

Line 461: "…and so adding TDR results in a synergistic interaction with ocean acidification" - to me this does not follow logically and the details of how TDR interacts is not well described.

We simply mean here that both warming+TDR and OA result in increasing surface pCO2 and decreasing saturation state, and so adding TDR slightly worsens OA. We have edited this sentence to make this clearer, and remove any implication that more complex interactions are being described (lines 498-501).