Review of “Resolving ecological feedbacks on the ocean carbon sink in Earth system models” by Armstrong McKay et al.

Armstrong McKay et al., have put a lot of 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.

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.

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 PO_4 is removed at each timestep to isolate the ventilation of PO_4 from the interior ocean. Because the circulation is static any differences in the transient response of interior-to-surface PO_4 supply results from the difference in initial distribution of PO_4 associated with the re-calibration of refractory export. The spin-up global mean concentrations of PO_4 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 PO_4 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.

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.

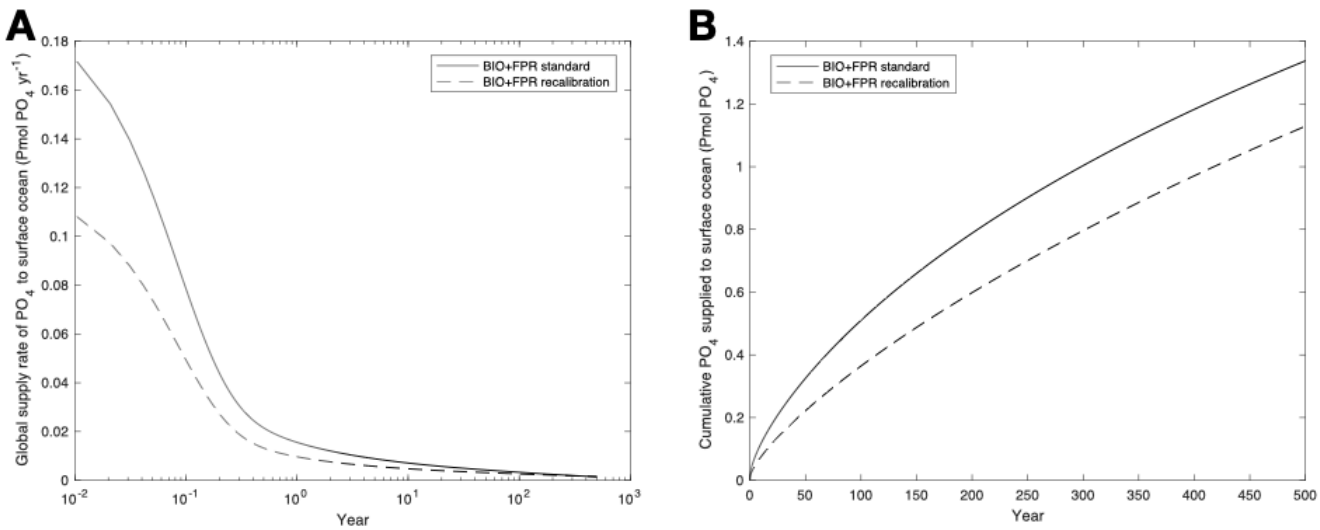

Figure 1. Transient changes in global PO_4 supply to the surface ocean over 500 years following a complete cessation of the biological activity for the default (solid line) and recalibrated (dashed line) BIO+FPR spin-ups. Panel A shows the supply rate of PO4 (Pmol $year^{-1}$). Time is shown on a log scale to show the initial rapid change. Panel B shows the cumulative supply of PO_4 to the surface (Pmol). All experiments are run from a spun-up initial state using an offline version of GEnIE.

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.

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.

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

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.

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.

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.

Lines 383 - 384: “…the amount of carbon exported for every unit of phosphorus increases with warming in response to stratification, reducing surface phosphorus loss…” - PO_4 is the model currency here not DIC so C is changing relative to P.

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.

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

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

Lines 431 - 436: see the comment for lines 142 - 144. There is maybe a conflation between transient source/sinks and equilibrium CO_2.

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.

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.

References

Cao et al., (2009) The role of ocean transport in the uptake of anthropogenic CO2. *Biogeosciences.*

Goodwin et al., (2008) Analytical relationships between atmospheric carbon dioxide, carbon emissions, and ocean processes. *Global Biogeochemical Cycles.*

Primeau (2005) Characterizing Transport between the Surface Mixed Layer and the Ocean Interior with a Forward and Adjoint Global Ocean Transport Model. *Journal of Physical Oceanography.*

Ridgwell et al., (2007) Marine geochemical data assimilation in an efficient Earth System Model of global biogeochemical cycling. *Biogeosciences.*

Seferian et al., (2020) Tracking Improvement in Simulated Marine Biogeochemistry Between CMIP5 and CMIP6. *Current climate change reports*

Wilson et al., (2015) Can organic matter flux profiles be diagnosed using remineralisation rates derived from observed tracers and modelled ocean transport rates? *Biogeosciences.*

Reviewed by Jamie Wilson