REVIEWER 1

Review: The open uncertainty window in ocean biogeochemical responses to fossil carbon emissions

Overall:

This study investigates how automated parametrisation may help reduce the gap between observations, models, and transient climate simulations. The authors employ three versions of a physical circulation model with different vertical mixing profiles (BL, TID, and CON) and conduct offline automated optimisation to calibrate six biogeochemical parameters against contemporary oxygen, phosphate, and nitrate observations for each physical model. The optimised parameter sets are then applied to the fully coupled model in transient CO2-forced simulations. The authors find that despite rigorous calibration, substantial uncertainty remains in the ocean response to climate warming, with each model showing different internal biogeochemical behaviour depending on the associated mixing regime.

Overall, this is an interesting study addressing central questions in Earth system modelling. The experimental design needs some clarification in terms of what is being interpreted and whether this is consistent with the analysis. I also have a few comments for consideration regarding the presentation of several figures and clarification of aspects of the methodology.

We thank the Reviewer for their careful assessment of our manuscript. Responses are found below in blue. Modifications to the manuscript text are shown in red.

General Comments:

Objective one states to 'define a window of model uncertainty for projections of marine biogeochemical cycling under climate change'. The authors do quantify this but it would interesting, and more complete, to include a comparison with an untuned biogeochemical configuration to demonstrate the extent to which tuning reduces this uncertainty window (or not). In other words, I don't think the authors fully "demonstrate the potential to reduce, but not to close the open uncertainty window in ocean biogeochemical responses to CO2 emissions." (lines 21-22) without this counterfactual.

This is an interesting point. The difficulty in providing counterfactuals is how to choose them- whether hand tuning of biogeochemical parameters to the three circulations would be the counterfactuals, or would the biogeochemical parameters need to be the same for the three circulations? We can make any arrangement of poorly tuned models, and all models are hand-tuned.

The sentence is modified to avoid mention of reduction:

However, an uncertainty window remains open in ocean biogeochemical responses to CO2 emissions, despite systematic tuning

Objective two explores correlations between physical and biogeochemical metrics and

their ability to inform about future changes. I think this analysis is very limited by:
- the small number of experiments (typically n=3) which I don't think can capture the likely complex or non-linear relationships explored.

We acknowledge the limitation of using only 3 models in the conclusions, and on the basis of this comment and that of Reviewer 2 have revised the paragraph where the correlation coefficient is first defined:

Line 240 Figure 6 examines whether the optimal misfit and optimal biogeochemical parameter values systematically vary with the models' physical circulation metrics using calculation of the square of the Pearson product moment correlation coefficient, R2. The sign of the correlation indicates the slope of the relationship. It is important to note that our small sample size of only three points increases the uncertainty in our analysis, and we furthermore do not account for covariance, nor do we attempt to identify causal links. These correlations are provided as a means to explore systematic relationships between optimal biogeochemical parameter sets and minimum misfit with the models' physical circulation metrics, which could suggest some physical-biogeochemical parameter compensation. Optimal minimum misfit and biogeochemical parameter values are also plotted against physical circulation metrics, with linear regression slopes in Appendix Figure A1. Minimum misfit demonstrates strong correlation with AMOC and ACC strength (R2 values over 0.9, Fig. 6), where weaker overturning is associated with a lower misfit. This association suggests a compensation between physical model uncertainty and biogeochemical parameter uncertainty.

- the lack of multi-variate analysis or dimension reduction when processes are interlinked, e.g., is a strong correlation between AMOC and POC fluxes actually representing some relationship with ACC and POC ? Holden et al., (2013) is in my view an example of a similar approach to identifying key controls on ∂13C but which uses a 500 member ensemble with emulators and principle component analysis to tackle the complex interactions between physics and biogeochemistry in a more comprehensive way (www.biogeosciences.net/10/1815/2013/).

We agree that a larger model ensemble and rigorous statistical analysis is required for robustly establishing causal relationships between variables. We caveat our conclusions, stating that the correlations are not mechanistic links. Our manuscript is mainly about model calibration and reducing model spread. The correlations help to identify potential physical-biogeochemical compensation, as described in the point above and now clarified in the manuscript.

The authors need to be careful interpreting and discussing correlations. For example, a negative correlation between NPP and ACC strength across the 3 tuned steady-steady model outputs (Figure 9 left most panel) strictly means NPP decreases with increasing ACC for a given similar fit to observations, e.g., NPP=f(ACC; J) where J is the misfit function in equation 1 and biogeochemical parameters are altered to achieve a similar J for each estimate of ACC. It's not obvious that this is interchangeable with the mechanistic interpretation of NPP as a function of ACC for a given set of biogeochemical parameters, e.g., the resulting NPP when ACC is perturbed NPP=f(ACC; μ) where μ is a fixed set of biogeochemical parameters. Notably, the change over the

21st century is consistent with NPP=f(ACC; μ) but is compared to NPP=f(ACC; J), adding more need to clarify this. The authors need to be clear what interpretation they are using and if the mechanistic interpretation demonstrate it can be interpreted as intended.
We agree that we provide no mechanistic explanation with our correlations- this is stated in the conclusions/caveat section and clarified in the revised paragraph reproduced above. We have removed discussion of drivers from the discussion sections. We have also now clarified throughout the results sections how correlations are meant to be interpreted:

Line 341:...but note that it is not the circulation per se so much as the compensation of biogeochemical parameters to the circulation contributing this association.
Line 380: …optimisation of parameters such as R−O:N and wD,0 controlling the…
Line 394: features to relatively small variations in Southern Ocean circulation metrics after AMOC, oxygen and nutrient distributions are constrained to be as similar as possible across the model ensemble. Dissimilar pre-industrial biogeochemical rates resulting from biogeochemical parameter compensation for residual physical circulation differences across the model ensemble…
Line 510: …reflecting calibration of wD,0 impacting remineralisation length scales, oxygen consumption and nutrient export.

Summary and Discussion Line 522: We find using simple regression correlations that response of a biogeochemical feature to warming may be predicted by a metric that showed no significant relation in the pre-industrial ocean, e.g. the response of shallow POC fluxes being strongly correlated to pre-industrial AMOC strength, but not to the change in AMOC strength, and with no strong correlation of pre-industrial shallow POC flux with pre-industrial AMOC strength. These shifting relationships between biogeochemistry and physics arise due to the constraints imposed by parameter calibration and inhibit identification of causality.
Conclusions Line 571:…and whether they are useful for identifying mechanistic linkages.

Heatmaps are used to summarise relationships across physical and biogeochemical parameters as measured by the square of the correlation coefficient, e.g., Figure 6. I think some caution is needed when interpreting the R2 values:
- I think the sample size is typically n=3, e.g., 3 physical models, which may not be big enough to reliably characterise the relationship between variables, for example, many of these relationships appear quite extreme with R2 close to or equal to -1 or 1. This suggests ~100% of the variability is explained which seems surprising – is this a caveat of 3 samples?
In no correlation did the R2=1 or -1. The strongest correlations were close to 0.98, which could be an artefact of the small sample size.
We agree the sample size is too small to draw robust conclusions on causal links. We have made this more clear now in the text, as described above.

- Is it reasonable to interpret R2 for the variables vs. AMOC seems if the 3 AMOC values are hand-tuned to be within 0.4 Sv of each other (Table 1)?

Perhaps not at steady-state, but the R2 does become relevant for the transient forcing, where AMOC response varies across the three models. Providing the R2 for the steady-state gives some context for the transient analysis.

- I wondered whether an alternative visualisation may better communicate these relationships and provide insight into the underlying patterns. For example, given the above, it may be illustrative to have the plots of variables against each other shown in supplementary.
We have added these supplementary figures in an Appendix.

- The authors discuss the heatmaps in terms of proportion of variance explained (R2) (see lines 240 -242). The heatmaps however show R2 from -1 to 1 not 0 to 1, so there should be some clarification that they also show information about the slope of the relationship.
This information is now added

- When using heatmaps to interpret relationships between variables (rather than parameter values) are single correlations appropriate when multiple parts of the system are changing? i.e., is there covariance that has not been accounted for?
We do not attempt to account for covariance. We clarify this now in the text, as provided above.

I would suggest considering whether some figures could be combined for better clarity and comparison of panels. For example Figures 13 and 14 are interesting together rather than separately; Figures 15 and 17 (maybe 8 & 10) are all components of organic carbon cycling so useful to have in one.
Figures 13 and 14 are now combined (new Fig. 12). Figures 8 and 10 are now combined (Fig. 8). Figures 15 and 17 are now combined (new Fig. 13).

Specific Comments:
Lines 104 – 108: It would help to very briefly explain the basis of the algorithm.
The basis of the algorithm is described in lines 93-103. The misfit is calculated for models using 10 parameter combinations, from which a probability distribution is constructed around the minimum misfit combination. A new generation (10 more parameter combinations) are selected from within this probability distribution.

Table 2: how does the misfit break down across the different tracers? Is one tracer always contributing a significant misfit? Is there a tradeoff between tracer misfits on the total?
The misfit is dominated by nitrate in a near-mirror inverse relationship with phosphate. This tradeoff limits total potential for improvement. This is now mentioned in the text:
L238: Nitrate fitness comprised the dominant fraction of the misfit, in an inverse relationship to phosphate (improvement to nitrate fitness resulted in reduction in phosphate fitness), a tradeoff that limited the total minimum misfit.

Line 257: I think this sentence refers to the R-O:N ratio but I feel this could benefit from explicit clarification.
This is now clarified

Line 286: I found this section difficult to follow. The manuscript describes the R-O:N ratio as being robust, but it is then discussed alongside parameters that are described as being highly sensitive to model circulation. It may help to clarify the distinction between parameter robustness, and its correlation with circulation metrics.
This section is rephrased to:
Line 293: In summary, while the phytoplankton growth parameter and R_-O:N ratio were robust across model calibrations, their optimal values were strongly anti-correlated with average ocean age and correlated with AABW strength. This is also the case for wdd, which showed a larger sensitivity to circulation differences.

Figure 1: BL, TID, and CON could be explained in the figure legend for better clarity
This information is now added to the figure caption.

Figure 7: Can there be some distinction between DIC and ALK for parameters that weren't used for optimisation on the top x axis, to distinguish them from parameters that were used for optimisation (PO43-, NO3-, O2).
A red box is added to show which profiles were used for calibration.

Figure 7: Where does the variability in alkalinity come from? It's quite striking versus DIC which is also not calibrated against.
Alkalinity is doubly affected by CaCO3 dissolution with respect to DIC, making it more sensitive to differences in CaCO3 export. There is high variation in the CaCO3 export across calibrated models because this is a fixed ratio of POC production, which is not tuned. This is now made explicit in the text:
Line 327 These differences arise due to differences in particulate organic carbon (POC) export production, from which CaCO3 export production (and subsequent dissolution) is calculated.

Figure 8: It would be useful to see a comparison with observations to in these spatial figures.
Observations are now added to Fig. 8. Text is also added to the document:
Line 336:
The greatest spatial variation in NPP is in the low latitudes (Fig. 8), which is a region where all models over-estimate NPP compared to satellite-derived observations (NASA 2018)
Line 348:
All models show an over-estimate of deep POC fluxes in the Southern Ocean compared to Honjo et al, 2008 (Fig.8).

Figure 16: colorbar labels are skewed from correct locations in positive part of colorbar
This is corrected (now Fig. 14)

Figure 17: it's notable that the spatial patterns have very similar structure despite the divergence of global properties in Figure 14. It suggests that the context of the question is important for the significance of the uncertainty window, i.e., if the question is how much do tuned models diverge in the broad regional responses then this figure suggests a much smaller impact compared to the question of globally integrated fluxes. This will have different implications such as carbon storage from globally integrated POC fluxes.
This is correct. We do not examine carbon storage here, which would expand the scope of an already lengthy paper. The carbon storage starting points are different at pre-industrial for each model.

We now note the broad similarity of POC deep flux trends:
L456: ", although spatial patterns are broadly similar across all model configurations"

Line 522: "substantial uncertainty still remains…" – it would be helpful to characterise or contextualise the changes found in order to demonstrate they are substantial. Are these bigger or small than the spread of CMIP6 model predictions?
"substantial" is now removed since we do not compare to a counterfactual, which would be the most useful way to examine how uncertainty might change with calibration. A new sentence is added to the conclusions on the topic of CMIP:
L565
Relative to the CMIP6 and CMIP5 model ensembles, our calibrated models have a smaller model response spread in NPP (Ryan-Keogh et al.,2025), POC fluxes (Wilson et al., 2022), and nitrogen fixation (Wrightson and Tagliabue, 2020), which is encouraging.

REVIEWER 2

Kvale et al. – The open uncertainty window in ocean biogeochemical responses to fossil carbon emissions.

Summary.

Kvale and colleagues explore how slight variations in how an ocean model is calibrated under pre-industrial or historical conditions can lead to a divergence in the transient response under greenhouse warming.

To do so, they use the UVIC ESM. They set up three independent physical states by using 3 different vertical mixing regimes - BL, TID and CON – which they tune manually so that each is as similar as possible. Similarity is measured in terms of AMOC, AABW, ACC and CDW overturning strengths and these 3 independent physical states are run for 10,000 years. They feed these equilibrated physical circulations into a transport tracer matrix (TTM) and follow a useful method for simultaneous parameter optimisation of 6 key ocean BGC model parameters. Thus, they arrive at 3 different physical and biogeochemical ocean states that nonetheless near-equally represent the ocean and BGC state relative to modern day observations. Following this achievement, the authors then force the ESM with these 3 ocean states with increasing atmospheric CO2 and describe the differences and similarities in the transient response.

The authors find divergent responses in key biogeochemical properties of interest, like net primary production, showing that how we set-up our ocean models can have a real influence on our projections. Their results highlight how even very similar model states tuned to reproduce modern day conditions can result in different projected states in the future and this is a major challenge in our work.

Comments.

The authors do a good job of assessing their model states and addressing biases at each stage of their approach. The paper is well presented and I find no major problems or concerns with it. I have some minor comments below that I hope the authors can address.

We thank the Reviewer for their careful assessment of our manuscript. Please find our responses provided in blue. Modifications to the manuscript text are shown in red.

Line 149 – sentence needs rephrasing.

Thank you for the correction- “later” was supposed to read “alter”

Line 297 – the authors state: “All three model configurations retain similar biases in pre-industrial nutrient and carbon distributions in all ocean basins (Fig. 7) despite calibration, which result from shared biases in the general circulation.” Can the authors

show this? At present, this statement has no clear evidence behind it and appears to be an assertion. Isn't it possible that the BGC model, which is simple as far as BGC models go, has an upper ceiling in its process representation that sets a limit to how well the model can represent nutrient and oxygen fields?

Figure 7 shows these shared biases, in the sense that e.g. DIC, PO4, and NO3 concentrations are too low at mid-depth and too high in the deep ocean, despite model calibration. This assertion is based on the circulation sensitivity study of Kriest et al. 2020, which is cited in the next sentence. For clarity, this citation is also moved to the sentence in question.

Correlation metrics - Isn't it a bit dishonest to present the correlation statistics when these are actually based on only 3 points? Noting that you've done a lot of work to get those three points, I think a different way of presenting this information would be good. Instead of large monotone boxes, couldn't you have individual panels with the three points and a line, which would be honest about the fact that these correlations are created from 3 points.

The application of the color matrices was a decision based on the need to show a lot of interrelated and complex information in a compact way. We acknowledge the limitation of using only 3 models in the conclusions. Individual panels of line plots are more difficult to assess holistically, especially for e.g. Fig. 9 showing how absolute metric vs. a change in metric affects the relationships. However, we have now added these as a supplement. The text has now been edited to emphasize the limited data points and the motivation for showing correlations:

Line 243 It is important to note that our small sample size of only three points increases the uncertainty in our analysis, and we furthermore do not account for covariance, nor do we attempt to identify causal links.These correlations are provided as a means to explore systematic relationships between optimal biogeochemical parameter sets and minimum misfit with the models' physical circulation metrics, which could suggest some physical-biogeochemical parameter compensation.

Discussion Lines 503 - 518: I was unsure of your point. Consider rephrasing or separating this paragraph into distinct arguments.

The paragraph has been separated into 2 to clarify that two distinct points are made. The language is also edited to clarify the points and to reflect concerns of Reviewer 1.

Line 520 - I disagree that the parameter optimisation is rigorous if it can only address 6 parameters. Suggest rephrasing.

"rigorous" is changed to "objective"

Line 525 - Which processes? Can you give a concrete example? Without one it becomes too abstract and hard to follow.

The sentence is somewhat redundant and is now removed. The following sentence is revised for greater clarity.

The discussion in general feels a bit thin while the results section is overwhelming with detail. Consider re-prioritizing.

We agree that there is a lot of detail in the results section. Calibration analyses are always difficult to write clearly and in a way that does not exhaust or lose the reader, which is why there is quite a bit of discussion interwoven into the results sections, making them long. In earlier versions we had more strictly separated results and discussion and found this was confusing readers. Internal feedback also showed that flow benefitted from additional "summary and discussion" sections to provide a succinct overview of the section results, which may be what the Reviewer is judging to be thin?

Thank you for considering my input to your research.