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 CO_2-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.

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.

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.
- 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/).

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.

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 R^2 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 R^2 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?
- 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)?
- 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.
- 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.
- 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?

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.

Specific Comments:

Lines 104 – 108: It would help to very briefly explain the basis of the algorithm.

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?

Line 257: I think this sentence refers to the $R_{-O:N}$ ratio but I feel this could benefit from explicit clarification.

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.

Figure 1: BL, TID, and CON could be explained in the figure legend for better clarity

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 (PO_4^{3-}, NO_3^-, O_2).

Figure 7: Where does the variability in alkalinity come from? It's quite striking versus DIC which is also not calibrated against.

Figure 8: It would be useful to see a comparison with observations to in these spatial figures.

Figure 16: colorbar labels are skewed from correct locations in positive part of colorbar

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.

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?