Review of “Exploring the role of different data types and timescales for the quality of marine biogeochemical model calibration” by Kriest et al., for Biogeosciences

Kriest et al., set out to quantify the impact of using different observational constraints when calibrating predominantly surface-based parameters in a biogeochemical model. They also quantify the impact of nominally short or long spin-up times when doing these calibrations. This is achieved through a series of biogeochemical model experiments using the same ocean circulation model. The authors find that the inclusion of organic tracers such as dissolved organic phosphorus led to improved calibration in terms of computational time, from both faster calibration and through use of shorter spin-ups, and a reduced misfit to observations. The authors also provide a detailed examination of the impact of spin-up time on various tracers.

Overall, this is a really interesting and useful contribution. Approaches to spin-ups are very varied and this provides practically-relevant evidence for the biogeochemical model community that biogeochemical models can be reliably calibrated with relatively short spin-ups. The analysis also contributes valuable scientific understanding about the different behaviour of tracers in spin-ups. Generally, the experimental design and analysis is detailed and robust but would benefit from some additional clarifications concerning one particular experiment and the issue of misfit generated from the ocean circulation model. The authors have done an impressive job to tackle what is a large topic but the downside is a manuscript that can be hard to follow in places. I think some additional definition of the quantitative concepts up front would provide a useful reference for readers to help navigate the manuscript better.

General Comments

- *Design of L4-SO experiment*

The experiment design is very logical overall and isolates the various differences as best as possible but the last experiment (L4-SO) adds two changes that should be isolated and aren’t: the 3000 spin-up time and the switch to using global tracers as constraints. How can you be sure that the results from this experiment solely represent the inclusion of deep tracer information rather than the longer spin-up? This experiment is used for a large part of the oxygen discussion so this needs addressing with an additional “L4-SO-Surface” experiment to isolate the longer spin-up time from the use of global tracers. If this additional experiment is not too different from the L4-SO experiment then I would be happy if the authors documented this briefly in the supplementary and kept the existing experimental set-up and manuscript text with a brief note stating it is not a problem.

I can see that the two factors are potentially inter-linked because the longer spin-up will primarily allow for the deep tracers to equilibrate so it may be that the impact of one is implicit in the other. It would be really interesting to know if just extending the spin-up time whilst constraining against the same surface observations would have a different outcome.

- *Influence of errors from ocean circulation and interpretation of calibration*

One of the challenges the authors identify is that the ocean circulation model contributes a large part of the misfit which cannot be improved by the biogeochemical model calibration. This introduces a potential issue that the inclusion of organic tracers in the calibration is accounting for some misfit due to the ocean circulation, i.e., overfitting the biogeochemical model, rather than representing the fidelity of the biogeochemical model. In particular, the interplay between the organic tracers and the particle flux exponent *b* relates to nutrient cycling in the upper ocean which is also strongly related to the ocean circulation model.

One way to remove the circulation error would be to calibrate against an existing model set-up, such as the ECCO* experiment, that acts as a set of pseudo-observations. Such an approach could completely demonstrate that the inclusion of organic tracers improves the calibration. However, I appreciate this is could be a big piece of work! I think the approach used in the manuscript is valid because it best relates to the practical reality of calibrating biogeochemical models. If the authors are able to explore this alternative option, even if just as a briefer complimentary set of experiments, then it would help strengthen their key findings. Otherwise, some additional discussion of this potential issue would be helpful.

- *Clarification of quantitative concepts used*

The methods section describes the RMSE misfit function used in the study, However there are a number of other quantities used in the study (bias, bias-normalised RMSE etc…) that are not described which made it hard to keep track of how they all relate to the interpretation and to each other. For example, this is particularly the case when discussing pattern-matching statistics and Taylor diagrams later in Section 3.2 and beyond. It felt like the authors are introducing new concepts for interpreting the data in these sections which makes the text harder to follow. It would really help with fully understanding what the authors are showing and describing if they can expand the existing section on RMSE to include an overview of the additional quantities, particularly how they relate to each other (e.g., Jolliff et al., 2009) and their use in interpreting the model performance.

Specific Comments

Lines 49 - 50: how many studies is this out of interest? A brief list of the studies with some details on what observations are used would be a great resource for the community - perhaps this could be added to the supplementary if not too onerous for the authors!

Lines 65 - 70: it seems worth mentioning how equilibrium can be defined, such as some dXdt quantity that is smaller than a defined threshold?

Lines 65 - 70: spin-up time will also depend on the initial conditions used - do the models you consider here all initialise from observations? Also, the spin-up time for a tracer like PO4 will be different to other tracers that involve additional processes like gas exchange, e.g., DIC.

Lines 85 - 86: This statement is a little unclear without having read the rest of the manuscript. I'm not quite sure whether this is referring to the way the misfit function is set up to balance the different constraints or whether this is referring to the focus on the surface ocean (in which case, it would help to add a sentence of clarification as to this assumption).

Line 114: "ECCO*" lead me looking for a footnote! It would help to clarify this is an abbreviation.

Line 127: "half of the model's zooplankton" - is this literally 0.5 * biomass?

Lines 132 - 137: are all the observations compared as annual averages?

Line 154: I think that 3000 years is an appropriate time for the model to reach equilibrium given the transport matrix circulation but it would help to confirm this is the case.

Lines 233 - 235: Does the convergence occur faster simply because there are less parameters to optimise?

Lines 235 - 236: Is it the spin-up time or the deep tracer constraints that drive the faster convergence? (See general comments above)

Line 278: "that targets only at dissolved" doesn't read particularly right to me, possibly there's a typo or grammar issue?

Figure 2: It would help to have an additional marker legend for the constraints

Figure 4: It's notable that although the EP fluxes are all very similar across the experiments, the flux to 2000m varies considerably! This makes me wonder whether the experiments have very different regenerated (and preformed) tracer inventories? For example, the S4-Org experiment seems like it would have to have lower regenerated inventories if the export flux is similar but so

little is getting delivered to the oldest ocean depths. Could this evidence for an additional constraint in calibrating the models?

Figure 4: It would be possible to add a shaded area for model predictions in panel D using the transfer efficiency values at 1000m from CMIP6 runs in Wilson et al., (2022). This would require changing the reference depth from 2000m to 1000m but at least for the Henson observations this could be calculated from the published fit to SSTs?

Figure 8: The information shown on spin-up time here would be useful in the introduction!

Lines 467 - 470: The upper/lower left/right description seems the wrong way round to the figure? I may be wrong, I found Figure 9 generally quite a challenging figure to interpret.

Figure 9: Overall, this is a challenging figure to interpret! I think part of the reason for this is that you have parametric and temporal ranges as axes with ranges also depicted by the rectangles, which may be leading to me misreading the figure and related text?

Lines 489 - 491: is the smaller temporal variation related to the shallower b? Does the shallower remineralisation mean that more of the temporal variation in tracers is weighted towards the faster-to-equilibrate upper ocean rather than the slower-to-equilibrate deep ocean?

Lines 505 - 510: this answer is somewhat specific to the tracers explored in this study. DIC and alkalinity may have different responses for example. This caveat should be mentioned.

Lines 546 - 547: the suggestion of an “early-criterion” may depend on what the initial conditions of the spin-up are? Would this still be the case if you spin the model up from uniform initial conditions? A clarification about initial conditions would be useful to make throughout the manuscript generally.

References

Jolliff et al., (2009) Summary diagrams for coupled hydrodynamic-ecosystem model skill assessment. Journal of Marine Systems. 76 (1-2), pp. 64 - 82

Wilson et al., (2022) The biological carbon pump in CMIP6 models: 21st century trends and uncertainties. PNAS. 119 (29)