We thank the referee for his / her encouraging and constructive comments! We address the suggestions and questions (highlighted in blue) in detail below (our answers in black).

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.

Indeed, optimisation L4-SO combines two changes to the overall misfit at the same time, namely the spin-up time and the consideration of deep inorganic tracers. For the latter to yield useful results, the longer spin-up time is essential. Thus, as correctly noted by the reviewer, the two aspects are inter-linked. A short spin-up time with deep tracers included the misfit would reflect mainly the initial conditions of deep inorganic tracers. Thus, deep tracer concentrations only become informative after a long (enough) spin-up time.

The reviewer's suggestion is to investigate whether the extended spin-up time introduces different results compared to S4-SO. According to our results we readily find solutions for the surface ocean

after short periods that agree with the solution after a much longer spin-up time, although deep inorganic tracers were regarded in one case. As shown in the discussion paper, the a posteriori evaluation of the (surface) misfit after 3000 years results in almost the same misfit values and its components as those after 10 years (see, for example, the upper right panel of Fig. 1). Likewise, biogeochemical fluxes (Figure 4) and other metrics (Figure 2) of L4-SO are very similar at the two time slices. Hence, we do not expect an optimisation against surface tracers after a spin up time of 3000 years to result in a fundamentally different set of parameters or biogeochemical turnover. Given this potential insensitivity of the surface misfit and other metrics to length of spin-up time, we have refrained from carrying out such a computationally expensive optimisation. **We will comment on this in the discussion of a revised version of the paper.**

- 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.

We thank the reviewer for this very constructive and helpful suggestion. Indeed, an identical twin experiment with pseudo-data or respective subsets (with different subsampling strategies) would be a great opportunity to explore this issue further. So far, and within the time frame available for our response, we are not able to devise such an experimental setup for additional optimisations, but we would like to follow up on this in the future. We note that the pseudo-data is typically representative for one particular model solution, which means that we should now have sufficient prior knowledge about the model behaviour to come up with a meaningful twin. **We will comment on the effect of circulation and the potential future directions in a revised version of the manuscript.**

- 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.

We agree that the presentation and discussion of the different metrics (RMSE, normalised RMSE, unbiased RMSE, bias, Pearson correlation coefficient) can be confusing. The individual metrics looked at are essentially all part of the RMSE and normalised RMSE. There are no additional quantities. How the individual parts relate to each other has been nicely addressed in Jolliff et al. (2009) and in Taylor (2001). **We suggest adding a brief, more formal description in the Methods section, which also refers to the previous works of Jolliff et al. (2009) and Taylor (2001).**

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!

The sentence in lines 49-50 relates to the statement by Arhonditsis & Brett (2004): *"Less than 5% of the modeling studies assessed included information (statistics or time-series plots) for all the state variables predicted, thus we were not able to evaluate overall model performance."* We discussed this reviewer's comment and realised that the reference to the work of Arhonditsis and Brett (2004), which included local, regional and large scale models, may not be representative for the current state of model assessments, almost twenty years after their study was published. To highlight this, we will rephrase the sentence as **"Almost two decades ago, far less than half of the studies reviewed by Arhonditsis and Brett (2004) reported performance statistics for all simulated state variables."**

We agree that a comprehensive and updated overview on model-data comparison would be very informative and helpful. Coming up with some detailed update is not straightforward and would be a study of its own, even if restricted to global biogeochemical model applications. For example, while the initial presentation and evaluation of global BGC models often focuses on model skill only with regard to inorganic tracers, often subsequent studies look more deeply into organic components (e.g., Gehlen et al., 2006; Petrik et al., 2022). We will mention that the situation with regard to model evaluation seems to be improving (for CMIP5 and CMIP6 models). We will also add the range of zooplankton metrics from the study by Petrik et al. (2022) in the revised manuscript (lines 305ff).

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

There may be different criteria to define equilibrium, or near steady state, e.g. by an Euclidean norm (e.g., Priess et al., 2013). In general, as different quantities (phosphorus, nitrogen, carbon …) are of different magnitude, one may rather define the requirement of a maximum relative change for each tracer over the steady annual cycle (as suggested by the reviewer). **We will explicitly mention and present the term "equilibrated" in the introduction.**

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.

The inorganic tracers are initialised from observed distributions, and the organic tracers from globally homogenous concentrations of 0.0001 mmol P/m3. **We will clarify this in Section 2.4.**

So far, our model does not include any DIC, but only phosphate, nitrate and oxygen. Indeed the spin-up times for oxygen and nitrate (for example) can be quite different: because nitrate concentration and its global inventory are affected by the spatial distribution of, and distances

between, denitrification and nitrogen fixation.. In the model predominant regions for denitrification and nitrogen fixation are the OMZ in the eastern tropical Pacific and the subtropics in the Atlantic and Pacific respectively. Because these distant regions of fixed nitrogen loss and gain are connected through circulation, the nitrate inventory equilibrates only after several millennia. Oxygen, in contrast, adjusts on shorter time scales, depending on the biogeochemical model parameter values (see also Kriest and Oschlies, 2015, Fig. 2). **We will comment on the different transient behaviour of different tracers, and their dependence on biogeochemical parameters in the introduction.**

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).

We agree, the statement could be misinterpreted. The statement should relate to the "calibration bias", in reference to the work by Arhonditsis and Brett (2004). For example, changes in model parameters could lead to a reduction in surface phosphate bias, at the cost of a larger bias in dissolved organic phosphorus (Kriest, 2017). The bias in DOP will, however, remain overlooked if we excluded thus tracer from the misfit analysis (or optimisation). **We will rephrase this more clearly in a revised version of the paper.**

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

The name "ECCO*" relates to the name in Kriest et al. (2020). **We will clarify this in the revision of the paper.**

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

Yes, we will exchange the current phrase by **"half of the zooplankton's biomass".**

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

Yes, we will change this to **"simulated and observed annual mean tracers"**. We will also **add a sentence that clarifies that we neglect any temporal variation.**

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.

As noted above, 3000 years may even be too short - this depends indeed on the biogeochemical constants applied (see above). A perfect optimisation setup would derive the spin-up time depending on, e.g., a Euclidean norm; however, this so far does not seem feasible in the current set-up of optimisation, where 10 model simulations with different sets of parameters run in parallel - here we may end up with very different simulation times (depending on parameter value) when aiming at a common criterion for steady state, which could result in a large potential computational overhead. **We will add a comment on this in a revised version of the revision.**

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

In general, the convergence depends on the characteristic of the parameter-cost/misfit function manifold and on how the optimisation algorithm copes with it. Faster convergence can be achieved with a reduction of the dimensionality of the problem (smaller number of parameters), by the introduction of additional (informative) data constraints, or by both combined. Since in this

particular case (S6-All vs S6-DOP) we only reduced the number of parameters to be optimised, we assume that the faster convergence indeed arises because of this reduction in the dimensionality of the problem. **To clarify this, we will add a comment in a revised version of the paper.**

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

The distribution of deeper tracer concentrations results from the combination of biogeochemical processes in conjunction with ocean circulation processes. Because the latter processes include millennial timescales, effects on deep tracers on the misfit are only informative after long spin-up times. We therefore cannot provide a conclusive answer to this question, as already explained above, in the reply to the general comments. However, we do know that large-scale mismatches in deep tracer concentrations, which only occur after a long enough spin-up time, are highly informative for model parameters such as *b* (e.g., Kwon et al., 2006, Kriest et al., 2012, Kriest et al., 2017).

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

We will replace "targets only at" by "considers only the misfit to".

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

We will provide a marker legend for the constraints (tracer types).

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?

We agree, an optimisation against a preformed and/or regenerated observational counterpart would be very useful. We so far have not implemented such "synthetic" tracers directly into the model. It would be possible to derive this information from AOU and stoichiometric assumptions - but, in cases where these assumptions (e.g., the value of $R_{O2:P}$) are simultaneously affected during the course of optimisation, the situation may become complicated: for example, should we apply respective values of the parameter $R_{O2:P}$ (that is itself subject to optimisation) for deriving regenerated nutrients from observations? Also, the circulation error of the models may become even more important, as it will affect the saturation concentration of oxygen that enters the calculation of AOU. Overall, the reviewer's comment is inspiring for interesting and more elaborate approaches to optimisation.

We would like to note that in S4-Org because of the very large *b* (=very shallow remineralisation) a large fraction of the EP will be recycled in the upper water column (e.g., about 30 % of EP will be recycled between 100 and 200 m). As EP is typically largest in the productive high latitudes (with deep winter mixing) or in upwelling areas (see Fig S2) this fraction of EP will likely be returned to the euphotic zone on seasonal time scales. Hence, when relying on annual tracer concentrations for model assessment, the definition of regenerated and preformed may be somewhat ambiguous (in contrast to the flux in 2000 m).

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?

We thank the referee to point to the publication by Wilson et al. (2022), but we would prefer to keep the reference depth of Fig. 4 at 2000 m, as many observational data sets refer to that particular depth. **We will additionally compare our accomplished (diagnosed from flux at 1000m divided by EP) transport efficiency TE with those reported in Wilson et al. (2022).** Wilson et al. report a global mean TE between 0.03 (UKESM1-0-LL) and 0.25 (IPSL-CM5A2-INCA), and a global observed TE of about 20 %. The full range of our model setups includes values as low as 6 % (S4-Org) up to 23 % (S6*-All), which agrees with the range reported by Wilson et al. (2022). Our model simulations with b=1, after a spin-up of 10 years exhibit a global mean TE of 0.17-0.18, which agrees with the observed range. Furthermore, we would like to note that there is a discrepancy between TE diagnosed from simulated fluxes and the nominal TE that can be calculated directly from b, via $(100\ m/1000\ m)^b$. For example, b=1 results in a TE of 0.1. This discrepancy arises from many facts, for example the (additional) transport of particulate organic matter through mixing, as well as numerical diffusion (Kriest and Oschlies, 2011). This should be kept in mind when parameterising and analysing the models.

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

We will extend and detail our presentation on model spin-up times in the introduction by the range shown in Figure 8.

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.

Yes, we agree, this figure is challenging, and we will try to describe this figure in a more comprehensible way. Figure 9 condenses and illustrates the outcome of our cross-validation experiment (i.e., variation among parametric setups vs variation due to the time at which the models are evaluated), and serves as a graphic illustration of the results depicted in detail in Table S4. Typically (except for particle flux at 2000 m) the left boundary of each rectangle shows the spread of model diagnostics after 10 years of simulation, and the right boundary after 3000 years. The lower and upper boundaries shows the difference between 10 and 3000 years, with the lower boundary indicating the model with the minimum difference (of the entire ensemble of six model setups) and the upper boundary the model with the maximum difference. Hence, all parametric and temporal differences fall within each rectangle.

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?

See above: the rectangles illustrates the domain of uncertainty (or variation) due to spin-up time and model parameters. **We will try to explain this figure and its interpretation better. In particular, the first paragraph of 3.5 will be revised, to clarify the link between Figure 9 and the results of the cross-validation experiments.**

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?

As can be seen from Figure 4, biogeochemical fluxes of S4-Org (the model configuration with the largest b or shallowest remineralisation) shows the largest differences between year 10 and year 3000. This can be explained with the complex feedbacks that occur at a global scale, where for

example, a "shallow" *b* eventually (on long timescales) increases subsurface nutrients (Figure S3), primary and production, grazing and ultimately deep particle flux in the tropics and subtropics (Figure S2). Removing this member from the ensemble leads to a lower maximum temporal variation of most biogeochemical fluxes (compare upper boundaries of left and right panels of Figure 9). Hence, the very high *b* of S4-Org triggers the largest temporal variation, whereas low temporal variation is achieved by models with *b*=1 and less. For OMZ volume, S4-SO and S6-DOP show the largest temporal variation (Figure 8). This likely arises from the complex processes (physical and biogeochemical) that determine OMZ extent, in conjunction with the high oxygen demand of remineralisation ($R_{O2:P}$ = 200 mol O_2:mol P) and *b*=1. So ultimately, the large temporal effects can be traced back to large-scale (remote) effects on long time scales that redistribute subsurface nutrients. **To clarify this, we will add a sentence in the following paragraph (currently line 493), that emphasises this again.**

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.

We agree, the spin-up times of DIC and alkalinity may be quite different, as also indicated by Seferian et al. (2013). **We will comment on this in the revised version.**

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.

We agree, and will restrict this statement to models that are started from observed inorganic tracer distributions.

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)

Arhonditsis and Brett (2004) Evaluation of the current state of mechanistic aquatic biogeochemical modeling. Mar. Ecol., Prog. Ser., 271, doi:10.3354/meps271013

Kriest and Oschlies (2011) Numerical effects on organic matter sedimentation and remineralization in biogeochemical ocean models. Ocean Modell., 39, doi:10.1016/j.ocemod.2011.05.001

Kriest et al. (2012) Sensitivity analysis of simple global marine biogeochemical models. Glob. Biogeochem. Cy., 26, GB2029, doi:10.1029/2011GB004072

Kriest and Oschlies (2015) MOPS-1.0: towards a model for the regulation of the global oceanic nitrogen budget by marine biogeochemical processes. Geosci. Mod. Dev., 8, doi:10.5194/gmd-8-2929-2015

Kriest et al. (2017) Calibrating a global three-dimensional biogeochemical ocean model (MOPS-1.0). Geosci. Mod. Dev., 10, doi:10.5194/gmd-10-127-2017

Kriest et al. (2020) One size fits all? Calibrating an ocean biogeochemistry model for different circulations. Biogeosciences, 17(12), doi:10.5194/bg-17-3057-2020

Kwon and Primeau (2006) Optimization and sensitivity study of a biogeochemistry ocean model using an implicit solver and in situ phosphate data. Glob. Biogeochem. Cy., 20, GB4009, doi:10.1029/2005GB002631

Leles et al. (2016) Evaluation of the complexity and performance of marine planktonic trophic models. Annals of the Brazilian Academy of Sciences, 88, doi:10.1590/0001-3765201620150588

Moriarty & O'Brian (2013) Global distributions of mesozooplankton abundance and biomass - Gridded data product (NetCDF) - Contribution to the MAREDAT World Ocean Atlas of Plankton Functional Types. Earth System Science Data, 5, doi:10.5194/essd-5-45-2013

Petrik et al. (2022) Assessment and Constraint of Mesozooplankton in CMIP6 Earth System Models. Glob. Biogeochem. Cy., 36(11), e2022GB007367, doi:10.1029/2022GB007367

Priess et al. (2013) Accelerated parameter identification in a 3D marine biogeochemical model using surrogate-based optimization. Ocean Modell., 68, doi:10.1016/j.ocemod.2013.04.003

Seferian et al. (2013) Inconsistent strategies to spin up models in CMIP5: implications for ocean biogeochemical model performance assessment. Geosci. Mod, Dev., 9, doi:10.5194/gmd-9-1827-2016

Taylor (2001) Summarizing multiple aspects of model performance in a single diagram. J. Geophys. Res., 1006, doi:10.1029/2000JD900719