We here briefly list a brief point-by-point reply (in black) as to how we have responded in the revised manuscript to the reviewers comments (blue). A more detailed reply is given in the discussion forum. Line numbers refer to the markup file submitted with this list, that highlights the changes made. Beside our changes made in response to the reviewers we have also spotted an error in table 2, a wrong reference, and modified a sentence to be more specific. These are listed at the end of this file.

Reviewer 1:

- 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.

We now discuss the potential effect of the combination of these two changes (length of spin-up time and consideration of deep tracers) in section 3.2, in particular lines 360-372 in the markup file.

- 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 now discuss the potential effects of circulation, and the possible solution via a twin experiment in section 4 (Conclusions), in lines 688-694 in the markup file.

- 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 have extended section 2.3 (Misfit function) by explaining different metrics and their potential relation more clearly, with references to Taylor (2001) and Jolliff et al., (2009). We also refer throughout the paper to the different metric concepts as presented in 2.3, and distinguish more clearly between the misfit function applied during optimisation (J_{RMSE}^{opt}) and the metric evaluated a posteriori (J_{RMSE}^{post}).

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!

As stated in our reply, we agree that a comprehensive and updated overview on model-data comparison would be very informative and helpful, but would probably be a study of its own, even if restricted to global biogeochemical model applications. However, we now added some examples of more comprehensive model assessment studies (such as the one by Petrik et al., 2022) and reworded our reference to the study by Arhonditsis & Brett (2004) to be more specific (see lines 51-65 in the markup file).

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

We now comment on the concept and criteria for equilibrium in the introduction (lines 74-100 of the markup file) and also address our applied spin up times in section 2.4 (lines 224-230).

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.

We have will clarified this in Section 2.4 (lines 215-216), and comment on spin up in the introduction and in lines 224-230 in the markup file.

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).

See above, we have reworded and clarified out statement regarding the reference to the study by Arhonditsis & Brett (2004). At the same time we have skipped the reference to Leles et al. (lines 51-64 in the markup file).

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

We try to clarify this by hyphens around the name (line 144 in the markup file).

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

We changed the phrase to “half of the zooplankton’s biomass” (line 159 in markup file).

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

We now describe more explicitly that we refer to the annual means in lines 178-180, and comment on the reasoning and possible effects of this assumption in the last paragraph of section 2.2 (lines 165-174 in markup file).

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.

We now comment on this in section 2.4, lines 224-230.

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

We comment on the reason for faster convergence in lines 315-316.

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

See above, additionally we discuss the potential effect of the combination of these two changes (length of spin up time and consideration of deep tracers) now in section 3.2, in particular in lines 360-372.

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

We replaced “targets only at” by “considers only the misfit to” (line 375).

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

We now 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 do not comment in the revised manuscript on the extensive topic of regenerated vs preformed nutrients, but more discuss the feedback effect by shallow remineralisation in S4-Org, for example in the revised section 3.5.

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?

As stated in our detailed reply to reviewer 1 we would prefer to keep the reference depth of Fig. 4 at 2000 m. In lines 482-497 we compare our accomplished transport efficiency TE (diagnosed from flux at 1000m divided by EP) with the ones reported in Wilson et al. (2022).

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

We extended our presentation on model spin-up times in the introduction, including a reference to the source of the range (Seferian et al., 2020) 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.

We have tried to describe this figure in the caption and text in a more comprehensible way (section 3.5).

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: We have tried to describe this figure in the caption and text in a more comprehensible way (section 3.5).

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?

In the revised section 3.5 we now try no express more clearly that a large temporal variation arises from shallow remineralisation because of complex, large scale feedbacks on longer time scales.

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 now note this caveat in lines 661-666 .

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 have clarified the initial conditions in section 2.4, and now comment more explicitly on the interplay between the model’s fit to observed global and local particle flux and the inventory of non-conserved tracers such as oxygen or nitrate, as well as the initial conditions in general.

Reviewer 2:

1) It is a bit confusing that the term J_RSME (equation 1) is used to denote the misfit (cost) function used in the optimization procedure, but also for the a-posteori (after optimization) quantification of model misfit. Although both misfit functions take the same form (eq. 1), they are different in which data and regions are considered. This is not explained in the methods section, rather there are only some hints scattered in the text. For example, the reader can guess what is shown in Fig. 1 based on the fact that the J_RMSE values are the same as in table 2, where there is a note in the caption. To make this more transparent, I would suggest to describe this two-fold use of the misfit function in the methods section. I also would prefer to use a (slighly) different notation for the cost function and the a-posteori misfit function (e.g. \hat{J} for the cost function).

We now extended the methods section (2.3) and included more description of the other metrics applied throughout this paper (see comments by Reviewer 1). These descriptions now include the RMSE, unbiased RMSE, bias and Pearson correlation coefficient. We also refer throughout the

paper to the different metric concepts as presented in 2.3, and distinguish more clearly between the misfit function applied during optimisation (J_{RMSE}^{opt}) and the metric evaluated a posteriori (J_{RMSE}^{post}).

2) The representativeness of the data products used for optimisation is discussed several places in the manuscript, but it would be helpful to gather a short description of this aspect in the methods section. The WOA climatologies for nutrients and oxygen should provide to first order a like-to-like comparison with the coarse model and the climatological ocean circulation. The same is probably true for the chlorophyl data? The discussions that are found later in the manuscript point towards the fact that the Martiny data is too sparse to be representative of the simulated model counterpart. It would then also be useful to frame the later discussions more consistently as a problem of representativeness, for example lines 262-266: Is this really a problem of the coarse resolution, or more the problem that the available data are not representative of a climatological average over a 1x1 degree gridcell? The same comment applies for lines 286-287, and also for lines 321-328 (where representativeness is finally mentioned). Also, is it really plausible that the lack of correlation (for large scale global patterns) can be explained by errors in the (data assimilated) ocean circulation? Doesn't this potentially also point towards a too low model complexity, i.e. only one phytoplankton and zooplankton type?

We skipped the occasional references in section 3.2 to the potential causes of large RMSE and RMSE' as well as low correlation. Instead we extended the presentation of data sparsity in the Methods section (2.2) and discuss the lack of improvement in pattern metrics at the end of section 3.2 (lines 417-426 in the markup file).

-line 2: "...state of the ocean biogeochemistry..." I would find it justified to delete the word biogeochemistry here. The models tell us something about the ocean in general.

We have deleted this (line 2 in markup file)

-line 15-16: "mainly located in surface layers" is a bit unclear, please consider rewording.

We replaced this sentence by: "Following the optimisation procedure we evaluated the RMSE for all tracers located in the upper 100 m (except for POP, for which we considered the entire vertical domain), regardless of their consideration during optimisation. " (lines 16-18 in markup file).

-line 32: Consider adding "combined with data assimilation techniques" or similar after "Global biogeochemical ocean models"

We rephrased this to: "Global biogeochemical ocean models, especially when combined with data assimilation techniques, …" (line 34 in markup file).

-line 49: "Far less than half of the studies...". Unclear which studies this refers to. Please consider rewording to clarify this.

We have rephrased parts of this paragraph to clarify this. We now also highlight the fact that the study by Arhonditsis and Brett is almost two decades old, and that the situation may have improved. (lines 51-64; see also our reply to comment by Rev 1.)

-line 82: "one of the simulated compartments." It is unclear to me what "compartments" refers to (inorganic/organic? tracers/fluxes? nitrogen/phosphourous?)

We have replaced this by "at least one of the biogeochemical state variables simulated by the model." (line 107 in markup file).

-line 83: "basic optimisation procedure". Is "basic" a good word here? Maybe better "reference"?

We have replaced “basic” by “initial” (line 107)

-line 79-90: It is confusing that it reads "three further experiments" and "these five optimisations". The fact that the "basic" optimisation is actualy two different optimisations (one with range of b more constrained) is difficult to understand. Please consider explaining this better.

We now mention that the initial optimisation setup against the full data set is carried out with two different boundary ranges for b, and then highlight that all further optimisations consider a reduced data set (see lines 107-117).

-line 101: consider changing "a circulation" to "a circulation field"

We have changed this (line 131).

-line 170: instead of saying "the DOP parameters", the two parameters could be spelled out for clarity.

We now spell out the two DOP parameters in the descriptions of the experiments.

-line 169-174: I don't understand the logic behind this experiment: If the objective is "to analyse whether the neglect of iron limitation in MOPS yields a bias in parameter estimates" then why is the number of paramters to be optimized changed at the same time? This way it is not clear whether changes in the optimized model performance are due to the change in data coverage or due to different set of parameters to be optimized? Could the authors please comment on this?

As noted in our detailed reply, we have also carried out an optimisation with all phytoplankton data included, but decided not to show this additional experiment, which yielded results similar to those of S4-SO. We mention this fact in the description of S4-SO (lines 246-255), and also provide an extended version of Table 2 in the supplement, where we also indicate the misfit, iterations and optimal parameters (and their ranges) for this additional optimisation.

-Table 1: instead of using italic font for fixed parameters, why not write "fixed at 0" and "fixed at 0.1848" for the two cases where this is relevant? Would be easier in my oppinion.

We have changed this.

-line 204: "introduced" sounds odd to me. Maybe just "used"?

We have changed this. (line 285)

-line 231: Why "a likely larger number"? L is given in the table, or does this refer to something else? Please clarify.

We deleted the word. (line 312)

-line 252-253: "...in contrast to a reduction of JRMSE by about 25% for the consideration of DOP measurements." I don't understand this. In Fig 1a I don't see a very significant difference between S6-all and S6-DOP?

We rephrased the sentence by “in contrast to a reduction of J_{RMSE}^{post} of DOP by 25% (Table S2).” (line 335)

-Figure 4, caption: For panel b (grazing), I can't see any short or long horizontal bar?

We deleted the sentence.

-Figure 7, caption: "Numbers on top of the panels..." I can't see any numbers on top of panels?

We deleted the sentence.

-Figure 9, caption: I don't think "sources of variability" is a good wording, please consider rewording. Also, "which is different for the individual model setups": It is actually only the L4-SO setup that is different, right? So maybe "which is different for L4-SO compared to the other setups"

We replaced this by “sources of variability in model results, caused by model setup (parameter sets) and spin-up time before model analysis (…)”. This comes along with a hopefully better description of this figure in the caption as well as in the text.

-line 28: non -> not

We have changed this. (line 30)

-line 273: expose -> show (?)

We have changed this. (line 356)

-line 278: organics -> organic tracer data

We have changed this. (line 374)

-line 370: agrees with -> falls within

We have changed this. (line 476)

-line 373: too low -> below all other estimates

We have changed this. (line 479)

Additional changes:

Table 2: The upper value for the good range of I_c of S4-SO is 31.8 instead of 39.9 (as wrongly given in the discussion paper). This has now been corrected.

The reference to Seferian et al. (2013) in lines 69, 71, 76, 462 and 556 of the discussion paper was wrong, and should have been a reference to Seferian et al. (2016). This has now been corrected. (See lines 91, 95, 100, 585 and 731 in the markup file).

We have reordered the sentence in line 555-557 (conclusions) of the discussion paper to clarify that we are referring to the extrapolation of trends in general, not to that of global OMZ volume in particular (lines 731-732 in the markup file).

The DOP data by Landolfi listed as “unpubl.” in the discussion paper are now referenced to Landolfi et al. (2008).