We thank the referee Jörg Schwinger for his insightful and very constructive comments! Our replies to his questions, comments, and proposed changes (highlighted in blue) are given below (our answers in black).

The authors present results from an extensive parameter optimization study for a global scale ocean biogeochemistry model. They evaluate the impact of different optimisation set-ups, in terms of observation based data used, parameters to be optimized, and model spin-up length. Particularly, they focus on the benefit that assimilation of organic tracer information (phyto- and zooplankton, POP and DOP) might have for the optimization of ocean biogeochemical models. The authors generally find that the inclusion of organic tracer data has a strong impact on the representation of particle sinking in the model and improves model fidelity with respect to global oxygen biases and the represenation of oxygen minimum zones. Different spin-up times are shown to have a considerable influence on global oxygen and OMZ volume biase, even for optimized parameter sets that perform similarly with respect to surface nutrient and oxygen data.

Parameter otimization of global biogeochemical models is understudied due to its complexity and technical challenges, and this study is a very welcome addition to the field. The results are useful (beyond the technical aspects of optimization) for a wide community of ocean biogechemical and Earth system modellers since they tell us about model sensitivity in general. The paper is well written and the method is sound. I didn't find any problem with this manuscript, except a few rather minor points where the manuscript would benefit from clarifications. These points along with technical corrections are listed below.

General:

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 agree that the distinction between the misfit function applied during optimisation and the one used for a posteriori evaluation should be described more clearly. **We will extend the Methods section related to the misfit function to clarify this, and also include more description on the other metrics applied throughout this paper (see comments by Rev. 1), such as (unweighted) RMSE, unbiased RMSE, bias and Pearson correlation correlation coefficient.**

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 suggest addressing the potential problems that arise from sparsity and episodic nature of the organic observations (except phytoplankton) in the Methods section (2.2) more extensively. This sparsity and episodicity, in conjunction with the global model setup (in short: a coarse, climatological circulation), and the neglect of temporal variability in the misfit function may introduce the following causes for the lack in pattern matching: (1) There can be a mismatch between the climatological circulation in comparison to the hydrographic conditions that prevailed during sampling of the observations; (2) The observations, which usually provide only snapshots of the biogeochemical state, may not be representative for the annual average that we apply in our misfit function; (3) the biogeochemical complexity may be too low (i.e. the model is too simple).

(1) and (2) are somewhat intertwined, and could be addressed with a physical model that more realistically resolves the physical environment at high spatial and temporal scales. However, it would be difficult to run such a model at a global scale, and over long time scale. Even if we had a model that resolves eddies and filaments well, it may still be possible that an eddy would occur at a slightly different location (i.e., a few kilometers further north), or with a temporal delay. In this case, any misfit function that aims at an exact spatial and temporal match would create a large error. One possible solution would be to use difference metrics such as the Hellinger distance to assess model performance, and we briefly discuss this at the end of section 3.2. (3) Too low model complexity is addressed briefly in the discussion about data-model comparison with regard to zooplankton, where at least for large zooplankton quite comprehensive direct (Moriarty and O'Brian, 2013) and indirect (e.g., Petrik et al., 2022) data sets exist. The situation is worse for microzooplankton, a common component in many global models, and for other model components such as DOP and POP, where data sets are indeed very sparse. Hence, even with a more complex model we may have to face the same problem of data sparsity in time and space.

To address these issues more concisely and comprehensively, we suggest to skip the occasional references to the potential causes of RMSE, RMSE' and correlation mismatch in lines 262-266, 286-287 and elsewhere, and rather combine these in a more exhaustive discussion at the end of this section, with reference to the (sparse) data coverage presented more extensively in section 2.2

More specific comments:

-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 will delete this.

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

We will replace this sentence by : When evaluating the RMSE of tracers located in the upper 0-100 m (except for particulate organic matter, for which we consider the entire vertical domain) we find similar values for the different model setups, with a range of 14% …"

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

We will add this: "Global biogeochemical ocean models, especially when combined with data assimilation techniques, …"

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

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."* **To highlight also the fact that this study is almost two decades old 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 will also briefly comment on a few recent advance in model assessment for global CMIP5/CMIP6 models.** (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?)

This refers to all biogeochemical state variables simulated by the model. **We will rephrase this by "at least one of the biogeochemical state variables simulated by the model."**

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

We would prefer to reserve the word "reference" to the simulation ECCO* by Kriest et al. (2020), **but will change "basic" to "initial".**

-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 will mention the fact 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.

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

We will change this.

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

We will spell out the two DOP parameters.

-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?

In fact, the (short-term) optimisations presented in this paper are only a subset of a larger ensemble of optimisations. We have also carried out an experiment similar to S4-SO, where we include the phytoplankton data in the Southern Ocean in the misfit (as in S6-DOP), but only optimise four model parameters (as in S4-SO). The results of this optimisation (the four optimal model parameters) were very similar to to those of S4-SO, and resulted in an optimal Ic = 28.45 W/m2/d (compared to 28.84 of S4-SO), and optimal zooplankton grazing rate of 2.967 1/d (compared to 2.807 of S4-SO), a stoichiometry of 199.5 (compared to 200 of S4-SO) and a *b* of 1 (the same as in

S4-SO). We note that these values lie well within the 0.1% range of misfit for S4-SO. To keep the plots and analysis simple, we have not added this optimisation in the tables and figures. **However, as this optimisation may provide valuable information and clarify the (small) impact of Southern Ocean data, we suggest to mention these results briefly in the Methods section as follows (lines 174ff):**

"We note that in an intermediate step between S6-DOP and S4-SO we carried out an optimisation similar to S4-SO, but with the Southern Ocean data included. This optimisation resulted in parameters which were very similar to those obtained from S4-SO, and hence produced very similar model results. We therefore do not include these in our model analysis."

If regarded helpful, we can also provide an extended version of Tables 1 and 2 in a supplement, where we indicate the setup (Table 1) and misfit, iterations and optimal parameters (and their ranges) for this optimisation. We would, however, prefer to not include its results in the figures and analysis of the main part, as this may reduce the visibility of the plots and does not seem to add to the overall insights and conclusions.

-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.

Yes, we agree and will follow this suggestion.

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

We will change this.

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

We are sorry for this confusion: "Likely" survived from an earlier version of this manuscript, and we will delete the word.

-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 are sorry that this was expressed ambiguously. What we meant to say was that JRMSE of DOP improved by 25% (Table S1), and **we will rephrase the sentence by "in contrast to a reduction of JRMSE of DOP by 25% (Table S2)."** As "particulate organic tracer concentrations" can be ambiguous (one may think of POP, which, in fact, deteriorates slightly compared to ECCO*), **we will also replace this term by "plankton concentrations".**

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

We are sorry for this confusion: The sentence relating to short and long horizontal bars relates to an earlier version of this figure, and we will delete the sentence.

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

We are sorry for this confusion: When preparing the final version of this Figure we skipped the numbers (as these are already in Figure 8). We will delete 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"

"Sources of variability": **We suggest to replace this by "sources of variability in model results"**. "which is different for the individual model setups": We agree, this expression is not clear. We here do not want to refer to the spin-up times applied to the different optimisations, but to the spin-up times after which each model was analysed (a posteriori). **We will replace this expression by "due to the spin-up time, after which each model was analysed" and skip the phrase in parentheses.**

Technical/typos:

-line 28: non -> not

We will change this.

-line 273: expose -> show (?)

We will change this.

-line 278: organics -> organic tracer data

We will change this.

-line 370: agrees with -> falls within

We will change this.

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

We will change this.

References:

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

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

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