The authors present the use of a machine learning-based approach to emulate the output, or more specifically, the model-data misfit as quantified by a cost function of a NPZD-style global ocean biogeochemical model to estimate model parameters. The approach is interesting, appears to work well in a relatively high-dimensional parameter space; however, some of the implementation details are not described well, and a key model shortcoming is not emphasized enough.

We thank the reviewer for their thoughtful suggestions and have attended to them in our answers below, and we feel that these comments have improved the manuscript.

**general comments**

One aspect of the study that may and is not emphasized enough in the current version of the manuscript is the use of a single phytoplankton and a single zooplankton tracer/variable in a global ocean model. Phytoplankton growth rates and other key parameters vary greatly between different phytoplankton groups, for example those that dominate coastal ecosystems in comparison to phytoplankton more commonly found in the open ocean. Thus, in some way, this study aims to do the impossible: fit the parameters of a one phytoplankton variable to a global dataset which is based on a complex spatially and temporally varying phytoplankton population. Here, I am not arguing that such as study should not be conducted -- the use of a machine learning surrogate model is very interesting -- but many of the shortcomings of the model identified in the study may be due to the very simple plankton representation in the model. For example, the manuscript describes underestimates of net primary production and issues in bloom phenology, and provides possible reasons for these model shortcomings, but the use of a single phytoplankton variable is not among them. Yet, a single phytoplankton model is unlikely to provide good primary production and bloom phenology estimates globally.

We agree wholeheartedly that a single phytoplankton and a single zooplankton functional group is not sufficient to represent the global marine ecosystem, particularly the transition from coastal to open ocean biomes. The reviewer is correct to assert that the model is simple, and that this simplicity limits its potential to represent the processes of primary production, remineralisation, carbon export and air-sea gas exchange, among others.

However, we note that extending the model to many functional types of phytoplankton and zooplankton is beyond the scope of this work. The biogeochemical model itself is called WOMBAT-lite precisely because of its computational efficiency, which relies on having less tracers and process complexity. That said, we would also hasten to point the reviewer to the

extensive developments that were made during this work to the model architecture (Appendix A), which form a considerable jump in model complexity compared to the previous WOMBAT model, while still maintaining its "lite" tracer number.

To acknowledge and address the reviewers concern, we have:

- Acknowledged in the introduction that the ocean-biogeochemical model is simple compared to others, but nonetheless is complex enough to demand optimization with surrogate techniques, and this optimization is required to realize the full potential of the model.

  Lines 93:99 - *In this study, we optimise a relatively simple ocean-biogeochemical model designed to represent open-ocean biomes using surrogate machine learning techniques: version "lite" of the World Ocean Model of Biogeochemistry And Trophic dynamics (WOMBAT-lite) (Fig. 1). This surrogate approach is crucial. Although WOMBAT-lite has few tracers and is computationally efficient, making it viable for high resolution configurations (Kiss et al., 2020; Matear et al., 2015; Menviel and Spence, 2024; Oke et al., 2013) and large ensembles (Mackallah et al., 2022; Rashid, 2022; Ziehn et al., 2020), it is nonetheless a global, three-dimensional, biogeochemical model with complex non-linear process interactions. This makes it computationally demanding enough to prevent parameter calibration via traditional techniques.*

- Acknowledged in section 3.4.3 (Phytoplankton bloom phenology) that

  Lines 635:637 - *"Alternatively, it is possible that representing the global marine phytoplankton community with only one functional type limits the model's potential to realize the full variation."*

- Acknowledged in the Summary that

  Lines 797:803 - *"Alternatively, spatial variations in ε that capture transitions from nano- to meso-zooplankton from oligotrophic to eutrophic regimes (Rohr et al., 2024) may serve to accelerate the phytoplankton bloom at the beginning of the growth season. Furthermore, the succession of different types of phytoplankton is important for the biological carbon pump (Tréguer et al., 2018). Therefore, representing these shifts in community with additional functional types of plankton beyond that explored herein might be important for the phenology of the annual spring bloom, and by extension Southern Ocean $CO_2$ fluxes."*

One more example: the authors highlight the "difficulty in reproducing the seasonality of air-sea CO2 exchange in the Southern Ocean" (l. 654) as a chief issue of the model. There are

many studies that emphasize the unique phytoplankton composition in the Southern Ocean and the role of Southern Ocean diatoms in modulating carbon export, for example "Influence of diatom diversity on the ocean biological carbon pump" (Tréguer et al., 2018, DOI: 10.1038/s41561-017-0028-x). Thus, it does not seem surprising that a single phytoplankton model, optimized on a global dataset, does not perform particularly well in the Southern Ocean.

In the revised manuscript, we have acknowledged and addressed that the model has limits in its potential to more completely represent the succession of different forms of phytoplankton due to only having one functional type, and that this may feed directly or indirectly into error in other features of the model, such as air-sea exchange of $CO_2$. (See above).

A related aspect to the comments above is that model performance can easily be biased based on the observation locations and the amount of data from each location. That is, if more open-ocean than nearshore observations are included in the RMSE-based cost function used for parameter estimation, then the phytoplankton variable will be parameterized more like a general open-ocean species (likely adapted to low-nutrient conditions) whereas the parameter estimates may be quite different if mostly nearshore observations are used. In summary, I suggest emphasizing the drawbacks of simple biogeochemical models more. Yes, parameter estimation becomes easier with fewer plankton variables, but even optimized parameters cannot capture the complex plankton ecosystem on a global scale for simple NPZD-style models.

We agree with the reviewer that the simplicity of the model will necessitate some acceptance of model-data misfit. We accept that this is the case, and we chose to optimize towards observational target fields with a bias towards open-ocean environments. Thus, our global ocean biogeochemical model is optimised towards representing open-ocean biomes.

We acknowledge this directly in the introduction:

Lines 93:95 – *"In this study, we optimise a relatively simple ocean-biogeochemical model designed to represent open-ocean biomes using surrogate machine learning techniques: version "lite" of the World Ocean Model of Biogeochemistry And Trophic dynamics (WOMBAT-lite) (Fig. 1)."*

Building suitable and balanced cost functions for parameter estimation experiments is not easy and the authors include 8 data products in their cost function (Eq. 2), making it quite complex and interesting. Here modelers and readers like me might be interested in a bit more detail that can be obtained without additional model experiments: One question if

interest is how "orthogonal" the different cost function parts are, i.e. which cost function parts/data products provide additional constraints on the parameters. Calculating the linear correlation of the cost functions parts for the 512 simulations or visualizing them using a "scatter plot matrix" could be of interest and would show which cost function parts can be optimized together and which move the parameter estimates in different directions.

This is a great idea. We show in the figure below the linear correlations between all cost functions for the 512 sensitivity experiments. Some are poorly correlated, and others are relatively well correlated. Among those most well correlated ($r \geq 0.7$) are:

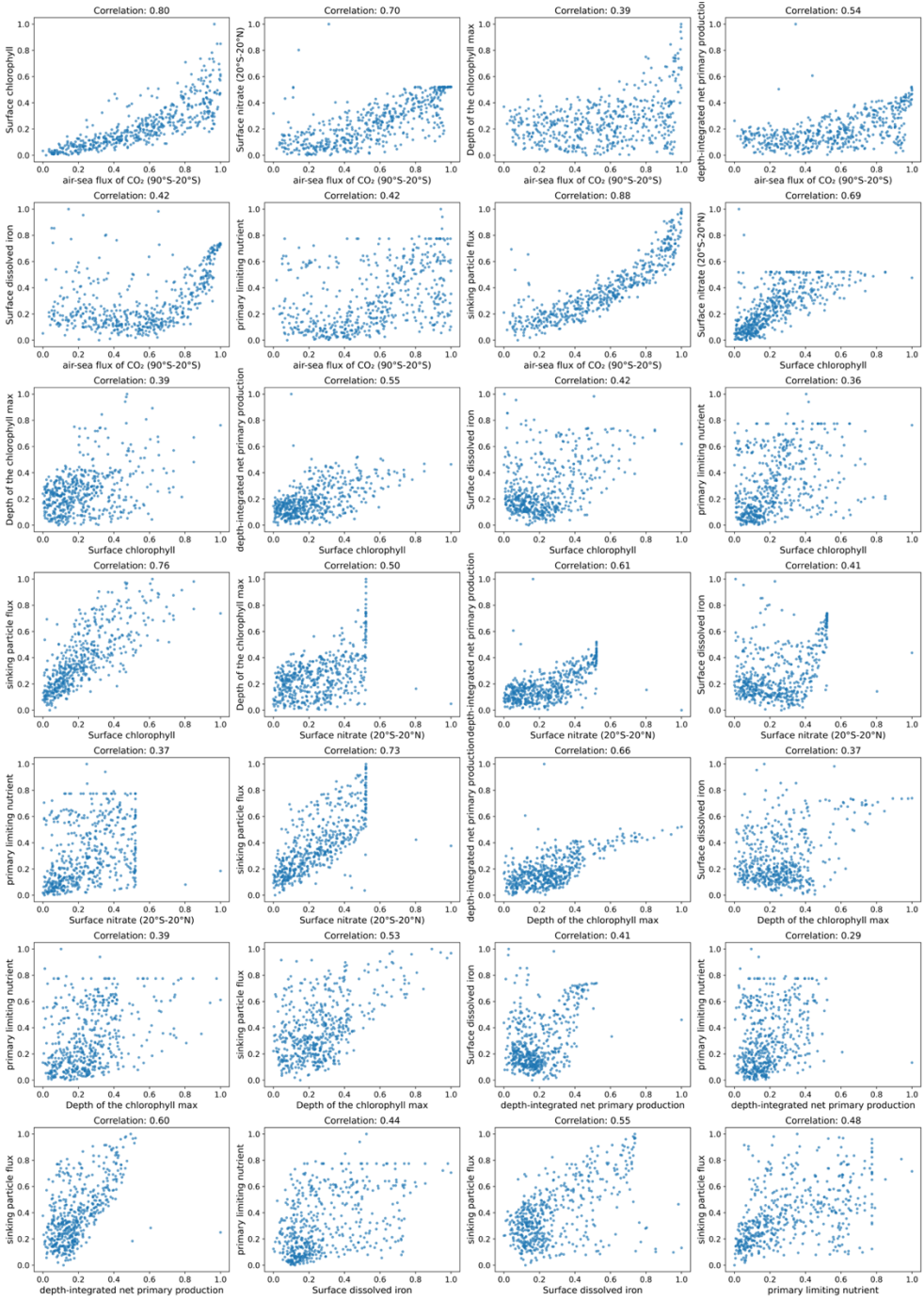
- Air-sea flux of $CO_2$ (20°S-90°S) – sinking particle flux
- Air-sea flux of $CO_2$ (20°S-90°S) – surface chlorophyll concentration
- Sinking particle flux – surface chlorophyll concentration
- Air-sea flux of $CO_2$ (20°S-90°S) – surface $NO_3$ concentration (20°S-20°N)
- Sinking particle flux – surface $NO_3$ concentration (20°S-20°N)

This suggests that these target fields are not providing completely orthogonal constraints on the parameter set, and that optimising for one is optimising for another. Mechanistically, it suggests that how we represent the biological pump has important effects on the air-sea flux of $CO_2$.

All other associations between target fields are moderately to poorly correlated, meaning that they are more orthogonal and that their optimisation can proceed without always affecting the other fields.

We can include this figure as a supplementary figure 5 for the manuscript and reference it here:

Lines 344:346 - *"Although expected, we note that the predicted cost functions using the GPR models are not completely orthogonal, with some being well correlated (Fig. S5), indicating that the optimization of parameters towards one target field will affect other target fields."*



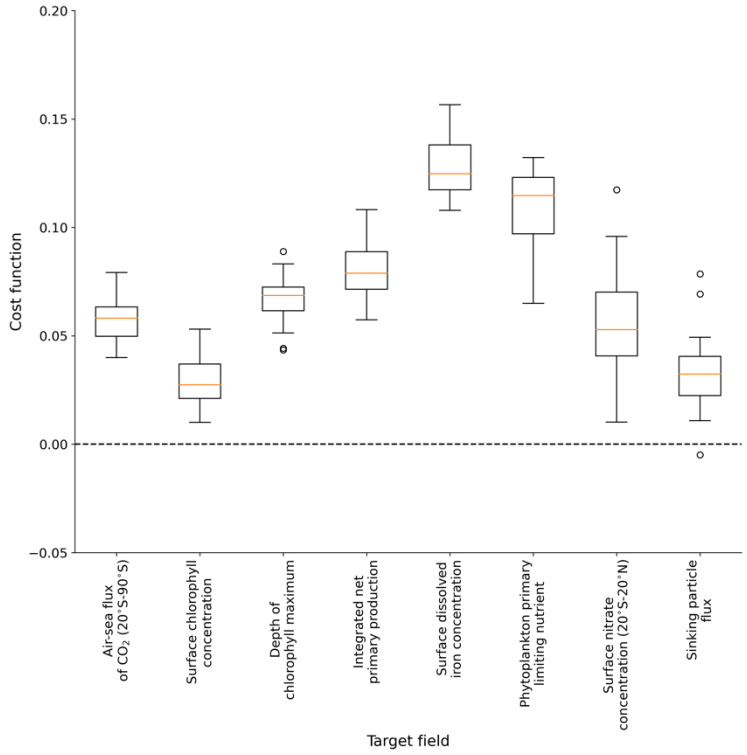

125

***Figure S5. Pair-wise correlations between cost functions of the 8 target fields.*** *If positive and significant, the gains in the skill of the model to reproduce one target field will positively affect the models skill in reproducing the other.*

Relatedly, it would be good to know if the optimal parameter values that were obtained perform well (have low values) for all cost function parts or if they perhaps do not fit certain datasets very well at all.

The optimal sample sets (sampling of the posterior distributions in Figure 6) performed well, as shown in Figure 7 in terms of the global cost function. However, what the reviewer is asking for is how the optimal parameter sets behaved in terms of the individual cost functions for all 8 target fields. This is presented below in the box and whisker plot.

Some fields are optimised more effectively than others. The air-sea flux of CO2, surface chlorophyll concentrations, sinking particle flux and surface NO3 concentrations are well reproduced. Surface dissolved iron and the primary limiting nutrient for phytoplankton growth are not as well reproduced.




*The performance of each of the 20 "optimal" parameter sets sampled from the posterior distributions of Figure 6 in the main text against all 8 of the target fields.*

Although not made explicit, the simulations in this study appear to rely on a 9-year spin up (output from year 10 was used for model data-comparison). The authors

state that "Continuing to run the model forward for 100 years post initialization showed some degradation in the performance" (l. 469). Parameter estimation relies on not having too much drift in the model and for the parameter values to have taken effect (there is no longer drift introduced by the change in parameter values). How did the authors ensure that the 9-year spin up was sufficient?

We explicitly discuss the choice of 10-year simulations in the final summary paragraph, acknowledging that this is a key caveat of our work. The degradation in the performance of the model that occurs by running the model forward for 100 years, rather than just 10 years, is an unfortunate symptom of only running our 512 training experiments for only 10 years.

We stress that this choice was unavoidable because of computational demands of running the ocean model. Moreover, because the ocean model was run online with the physics, our choice of using only 10 years was also informed by a desire to optimise the biogeochemical ecosystem component, which overturns quickly, while limiting physical biases, which grow as the simulation progresses. Optimising the biogeochemical component of an ocean model should be done in a physical setting that is as close to the observations as possible. If we optimised the model using simulations that extended for centuries or even millennia, then it is likely that we would be correcting for mismatches that are generated due to physical error, rather than biogeochemical error.

We fully acknowledge and discuss this caveat at several points in the manuscript and suggest future fixes:

Lines 253:263 – *"We chose to run the experiments for only 10 years, making a total of 5120 model years and at a nominal horizontal resolution of 1º. This short timescale was sufficient to assess the skill of the biogeochemical model, at least regarding its ecosystem component. Marine phytoplankton contribute half of all primary production in the Earth system (Field, 1998) but represent less than 1% of photosynthetic biomass (Friedlingstein et al., 2023; Le Quéré et al., 2005), meaning that they turn over quickly. Changes to key parameters within the ecosystem component therefore result in a rapid realisation of different patterns in biological states (e.g., chlorophyll and net primary production, among others). Our analyses and optimisation thus focus on the ecosystem component using 10-year model runs. We do acknowledge that longer-term, low frequency modes of variation exist in biogeochemical models, and to partially address this we completed 100-year simulations with optimal parameter sets. However, we also note that longer integrations risk the compounding of physical and biogeochemical model errors as the physical state drifts further from the observations. Optimisation of the biogeochemistry under these conditions can cause over-tuning of the biogeochemical parameter set that compensate for physical errors (e.g., Singh et al., 2025)."*

Lines 559:564 – *"Continuing to run the model forward for 100 years post initialization showed some degradation in the performance (red bars in Fig. 7). This is expected,*

*since our optimisation procedure was trained on model output only 10 years post initialization due to computational constraints. Model outcomes drift further away from the target fields with longer integrations. Lower frequency variability and trends are thus missed by the optimisation that are nonetheless present in the biogeochemical model, and these play out as the model is integrated forward for longer."*
*Lines 807:816: "Even with its optimal parameters, WOMBAT-lite suffered a loss in performance when run over 100 years compared to when run over only 10 years. Future iterations of surrogate-based optimisation would therefore benefit from extending the length of simulations done by initial set of sensitivity experiments. That said, significant savings in computation efficiency would be needed before this is possible with computationally demanding models, such as ocean biogeochemical models, but could be feasible by running the biogeochemical model offline from the ocean physics (e.g., Séférian et al., 2013). This approach would also eliminate any confounding errors caused by an evolution of the ocean's physical state since the physical state would not be allowed to evolve. Future versions of WOMBAT, including WOMBAT-lite, WOMBAT-mid and WOMBAT-full, and their deployment into different configurations (e.g., higher resolution versions) would benefit from this methodology of optimisation."*

One of the metrics used in the cost function is primary limiting nutrient data, but there is not much information about its use and how model-data misfit is quantified. The model seems to include carbon and iron, all other elemental quotas/ratios are considered to be fixed. Fig. 3g shows nitrogen, iron and phosphorus as limiting nutrients in the data, how are these compared to model output? How is the RMSE computed? Some more information is needed.

We have included a greater discussion of how this comparison is made between the observations and the model.

Lines 204-209: *"For the primary limiting nutrient dataset, we only consider nitrogen and iron as limiting (excluding phosphorus) and ascribe nitrogen limitation, iron-nitrogen co-limitation and iron limitation as equal to 1.0, 1.5 and 2.0, respectively. This is compared directly to the degree of limitation by the model, which varies continuously from strong nitrogen limitation to strong iron limitation (1.0 to 2.0). Although the model does not include co-limitation as a process, simulated values between 1.0 and 2.0 can represent seasonal variations between nitrogen and iron limitation over an annual timescale."*

Several aspects of study design and implementation details are not described well and can lead to confusion. For example, the generation of 512 parameter samples is mentioned in several places independently: "We undertook 512 simulations that each sampled randomly from predefined ranges of 24 key parameters related to the ecosystem component of the model" (l. 189). Then, later: "Initially, a Quasi-Monte Carlo Sobol sequence is applied to generate 512 parameter samples using the Uncertainty Quantification Python Laboratory package" (l. 263). Here, it is not clear if the 512 parameter samples are the same ones mentioned before; these methods sections need to be connected better. Furthermore, the text mentions that

512 parameter samples are drawn initially and then a sample size of 512 was selected based on the sample size sensitivity experiments (l. 267). Was it sheer luck that the initial number of experiments was also the right number of experiments confirmed in later tests?

We have added extra text to explain why 512 experiments were done for both steps: the sensitivity analysis and the optimisation. This information is already in the supplementary material, but we have been more explicit in pointing the reader towards it and have also added more sentences to explain why two sets of 512 experiments were done.

Lines 220:222 – "*A total of 512 experiments was selected because it was enough for training the surrogate machine learning model to an acceptable standard (Fig. S1).*"

Lines 301:302 – "*The analysis focuses on the target fields detailed in Fig. 3 and section 2.3.1. To generate the 512 sensitivity experiments required to train the surrogate machine learning model,*"

Lines 322:332 – *"Following sensitivity analysis, which identified the most important parameters for model performance (RMSE) of the eight target fields, we performed parameter optimisation (Fig. 1). This study uses Gaussian Process Regression-based Bayesian Optimisation (G-BO) (Reddy et al., 2024a) to identify the optimal parameter distributions so that WOMBAT-lite can best reproduce the eight target fields simultaneously. The process begins by generating another 512 parameter samples via the same Quasi Monte-Carlo (QMC) Sobol sequence design implemented through the Uncertainty Quantification Python Laboratory (UQ-PyL) package (Wang et al., 2020). This set of 512 sensitivity experiments are different from those generated during the sensitivity analysis because we now use a reduced set of only the most important parameters, which also provides a denser sampling of the parameter space essential for the G-BO process (Reddy et al., 2024a). These 512 sample parameter sets are used as input to WOMBAT-lite, and the model is run forward for 10 years (see above). A sample size of 512 is selected based on the sample size sensitivity experiments (Fig. S3)."*

**specific comments**

L 12: "Optimisation of the model parameters is crucial to ensure model performance based on process representation, rather than poor parameter values.": This sentence is not very clear, please rephrase.

Rephrased for clarity:

Lines 12-13 - *"Optimisation of the model parameters is crucial to ensure that model performance is based on process representation (i.e., functional forms), rather than poor choices of input parameter values."*

L 19: Isn't the size of the training dataset (512) somewhat in conflict with the previous claim that "(tens of) thousands of simulations are required to accurately

estimate optimal parameter values" (l 14). Perhaps it would be useful to add some qualifiers that this large number of simulations would be required by naive approaches like grid search and for a certain number of estimated parameters.

We have added a clarifying clause to this sentence:
Lines 18:20 - "A computationally inexpensive surrogate machine learning model based on Gaussian Process Regression was trained on a set of 512 simulations with WOMBAT-lite *and was used to produce synthetic results emulating tens of thousands of simulations*."

L 22: What are "optimal posterior distributions"? From a Bayesian technique, one would expect to obtain (samples from) the posterior distribution, or perhaps maximum a posteriori (MAP) point estimates. But there is no "optimal" posterior distribution.

Replaced "optimal" with "constrained"

L 66: "However, even if our understanding and observational network were complete, there exist many tuneable and potentially inter-dependent parameters that control many target outcomes...": True, but with complete understanding, we would know the relevant rates that these parameters aim to represent. We would not even need models if our understanding and observational network were complete. I would suggest rephrasing the issue and avoiding the fictional scenario of "complete" knowledge.

Rephrased.

Lines 69:71 – *"However, even with greater understanding and observations, there exist many tuneable and potentially inter-dependent parameters that control many target outcomes (air-sea $CO_2$ fluxes, nutrient fields, chlorophyll concentrations, etc.) that must be reproduced simultaneously."*

L 74: Does the "a priori ranges" imply that the prior distributions are uniform? As a first-time reader familiar with Bayesian techniques, I would have expected the term "prior estimates" here or a brief explanation of the "a priori ranges".

The prior distributions are indeed uniform over the prescribed range. We have added "uniform" as an adjective.

L 118: "We have developed a new ocean biogeochemical model called WOMBAT-lite": This statement is a bit confusing to the reader, as a previous sentence appears to suggest that WOMBAT-lite has been applied in previous studies: "Although WOMBAT-lite has few tracers and is computationally efficient, making it viable for high-resolution configurations (Kiss et al., 2020; Matear et al., 2015; Menviel and Spence, 2024; Oke et al., 2013)" (l 90). I would suggest clarifying.

Changed to "updated".

L 118: Fig. 2 is referenced before Fig. 1.

Figure 1 is referenced in the Introduction.

L 119: It would be useful to many readers to add a brief description of Sobol sensitivity analyses or at least a reference.

Reference added.

L 145: How is cell size included in the model using just one phytoplankton tracer?

See the Appendix for this explanation. Variable cell size is emulated by considering a relationship between the density of phytoplankton and the average cell size. This affects half-saturation coefficients for nutrient uptake, the packaging effect on light transmission and also on the sinking rates of detritus.

L 174: The connection between CbPM, VGPM and the data used for model evaluation needs a better explanation.

Added a sentence to explain why we use CbPM-based NPP rather than chlorophyll-based NPP.

Lines 200:202 – *"Net primary production (NPP) built from the CbPM products should therefore provide more independent constraints on model assessment than an NPP product built from chlorophyll concentrations."*

L 189: Where samples obtained "randomly" or via a more sophisticated sampling strategy like Latin hypercube sampling? How were the parameter ranges selected?

Additional information given:
Lines 218:220 - "We undertook 512 simulations that each sampled randomly from predefined ranges of 24 key parameters related to the ecosystem component of the model (Table 1; Fig. 1) *using a Quasi-Monte Carlo Sobol sequence (see section 2.4).*"

L 288: How does the max-min scaling work? Are max and min values taken from the 512 samples, are outliers considered in the normalization (for example those in the depth of the DCM, Fig. S2 d)?

Clarified.

Lines 334:335 - "is the normalized root mean square error (scaled by the max-min*, such that the worst experiment has an NRMSE of one and the best is zero*)"

Eq. 2: I would recommend placing the superscript into parentheses "(n)" to avoid confusing it for an exponent.

Done.

L 301: The "p(z)" here is not very helpful without providing additional context.

Removed.

Fig. 4: It is unclear how the primary limiting nutrient is turned into a number and how the misfit is quantified.

This information has been added earlier in the manuscript within section 2.3.1 Observational target fields for assessment.

Fig. 5: What kind, if any, normalization was applied to the RMSE values here? Not using any would make the values dependent on the scale and units of the properties/target field shown. Using max-min (previously denoted as NRMSE) would make it dependent on the variability that can be introduced by any changes in the parameter values -- which might vary greatly between properties.

The RMSE values are presented here as normalised by the max-min scaling. We have added an extra note of this in the figure legend.

Lines 441:445 – *"Performance is measured by the root mean square error (RMSE), which is here normalised by the full range (max-min scaling). First-order indices are direct individual effect of a parameter on the target field, while higher-order indices indicate that interaction effects with other parameters are important. First-order indices sum to at most one for a given target field, while first-order + higher-order indices sum to at least one. If there are no interaction effects, then higher-order effects are nil."*

L 444: The units here would raise fewer eyebrows by changing the "m^6 (mmol C)^-2" to "(mmol C m^-3)^-2".

Agreed!

L 450: "The fact that our optimisation always chose the lowest values (near 0.01 day-1) suggests that the proportion of the community that is stressed is considerably lower than we assumed.": I would suggest being careful in generalizing from the "optimal" value of a single global phytoplankton tracer to properties of the full plankton community.
Agreed. Rephrased.

Lines 535:537 - *"The fact that our optimisation always chose the lowest values (near 0.01 day$^{-1}$) perhaps suggests that, at least with our simple model, the proportion of the community that is stressed is lower than initially assumed."*