Response to Reviewer #2

We greatly appreciate the time taken by the reviewer to read our manuscript. We have taken into consideration and addressed all comments, questions, and suggestions from the reviewer, and we feel that the revised manuscript is now substantially stronger as a result. Additional changes made to the text at the request of the reviewer have been highlighted in red in the newly revised manuscript. In the following, reviewer comments are repeated in blue italics and our responses are provided in the bulleted sections.

The authors have mostly addressed my comments, and I am happy about the addition of appendix B. However, results from the additional twin simulation experiment (TSE) that was conducted, highlight differences between parameter estimation experiments including 5 compared to 17 observation types. Yet, this aspect is not examined adequately in the manuscript.

 # general comments

The introduction provides a nice overview of recent parameter estimation studies for BGC ocean models. However, it focuses mostly on the computational difficulties associated with the estimation of a large number of uncertain parameters. Other big problems that are not mentioned are that of parameter dependency and limited data availability (not in the spatial or temporal sense, but in the type of data commonly observed). For example, low phytoplankton growth and mortality rates may yield similar phytoplankton model estimates to high growth and mortality rates -- the values of phytoplankton growth and mortality rates are difficult to determine with phytoplankton data alone. In terms of data availability, chlorophyll data (the most commonly used data type for BGC data assimilation) may be useful for constraining phytoplankton parameters -- with the caveats just mentioned above -- but almost useless in estimating a parameter related to detritus, e.g., a decomposition rate. Both of these problems are more likely to occur in BGC models with large numbers of uncertain parameters. In fact, they may explain some of the differences seen in the TSE in which all state variables are observed (the 30-day TSE) compared to that in which only 5 variables are observed (the annual TSE). Fewer data types constrain fewer parameters and, since the objective function is based on fewer data types, sensitivity values can differ greatly.

Here, it is unfortunate that the two TSEs in question differ in two aspects (number of observed variables/data types and length of the observations), so it becomes difficult to estimate to what degree each aspect causes the difference in the results. I would suggest a new experiment in which only one of the two aspects is changed from the 30-day TSE, but I can understand if this is not possible due to computational constraints.

Based on my comment above, and even without additional experiments, the authors should mention and discuss parameter dependency and limited data availability as important parameter estimation problems in the introduction, which is currently very much focused on computational issues.

Furthermore, the difference in the number of observed variables should be emphasized more in section 4.1 when interpreting the results of the TSEs. The expanded interpretation of results may also warrant a new paragraph in the Conclusions section.

We thank the reviewer for their continued valuable comments, which we summarize here: (i) Additional consideration should be paid to other challenges that limit our ability to predict parameter values; and (ii) The two TSEs presented in the paper change multiple aspects simultaneously.

Regarding the first point, we have added a new paragraph near the beginning of the Introduction that addresses the issues of data sparsity and parameter dependence raised by the reviewer, which are indeed valid and pressing concerns. This additional text has also allowed us to restate the purpose of this paper more clearly for the reader. That is, we are primarily focused on presenting and demonstrating a novel calibration methodology that is computationally efficient for large sets of parameters. This methodology is sufficiently flexible that different models with fewer parameter dependencies and

different reference data that is less sparse can be examined in future studies. These points are additionally reiterated in Section 4.1 in the revised paper where we discuss the results of the TSEs.

To address the reviewer's second point, we have run two additional TSEs that use all 17 state variables in the objective function with the full parameter bounds from Table C1. In the first case we simulated 30 days only, like the TSE shown in Fig. 5 where we used tighter parameter bounds. In the second, we calculated monthly averages for the last year of three-year model runs, such that the resulting TSE can be compared to the results in Fig. 6 where only five state variables were used in the objective function. The resulting recovery of baseline parameter values for the two additional TSEs is qualitatively like those shown in Figs. 5 and 6, respectively. Because of this similarity, we do not show these additional TSE results in the revised paper, but these results are now discussed in a new paragraph near the end of Section 4.1. In particular, the additional TSEs suggests that the duration of the model runs, more so than the parameter bounds or number of state variables, may be responsible for much of the difference between Figs. 5 and 6. We also reiterate the possibility that the previously discussed issues of data sparsity and parameter dependence may contribute to a decrease in the number of fully recovered parameters.

 # specific comments (line numbers correspond to tracked-changes document)

L 4: "ocean high-dimensional (in parameter space) BGC models" → "high-dimensional (in parameter space) BGC ocean models"

- The recommended change has been made.

L 9: "objective functions": some readers may be more familiar with the terms "cost function" or "loss function" and I would suggest adding a brief description (such as "which quantifies the error between observations and the model estimate") to make the abstract more accessible.

- These changes have been made to improve clarity.

L 27: "Subsequent to verification using the TSE, we use the method to estimate parameters for the two sites, both individually and together.": Mention here that "real" data are used for this.

- We have clarified this point as requested at the end of the third paragraph in the Introduction of the revised paper.

L 102: "these weights can be adjusted as desired, for example [...] to provide greater weight to state variables for which observational uncertainties are smaller": While true, this statement makes it appear as if this objective function does not provide greater weight to state variables for which observational uncertainties are smaller, when in fact it does so via the \sigmas. I would suggest changing the statement to "these weights can be adjusted as desired, for example [...] to provide greater weight to select ocean sites".

- This is an interesting point and touches on the differences between observational uncertainties in the data and physical variability in the data (which are both often represented using the symbol sigma). To clarify our intentions, we have made the sentence reference by the reviewer more general and provided a comment after Eq. (2) specifying that our sigma represents the latter type of variability referenced above.

Fig 1. caption: "calculates the error": I would suggest using "calculates the current value of the objective function" just because the diagram shows the "Objective Function Calculator" and no mention of "error".

- The caption has been updated accordingly.

Fig 5. caption: Mention what S-hat is. Consider adding a shaded area or dotted blue lines to indicate the +/-5% range around the baseline value, indicating that a parameter is considered "recovered".

- The captions for Figures 5 and 6 have been updated to describe the meaning of S-hat and refer the reader to Eq. (6) where it is defined. We have also added horizontal dashed blue lines in Figures 5 indicating +/-5% around the baseline value, as suggested by the reviewer. We have not done the same to Figure 6 since the parameter range is not a constant percentage of the nominal value so the +/- 5% is not a constant normalized distance from the baseline values for all parameters, and therefore cannot cleanly be represented using single continuous lines.

Fig 5. caption: Consider using "p-hat_i - p-hat_O, the difference between the test and baseline normalized parameter values" instead of "the difference between the test and baseline normalized parameter values, p-hat_i and p-hat_O, respectively", it directs the reader's eyes directly to "p-hat_i - p-hat_O" on the right y-axis.

- This change has been made in the captions for Figures 5 and 6.

Fig 5. caption: To me personally, the sentence "The parameter values are normalized between 0 and 1 based on the upper and lower bounds for each parameter." is not helpful here.

- We have removed this sentence from the captions of Figures 5 and 6.

L 230: "Five parameters fell into this category": Due to changes in the text preceding this statement, it is no longer clear what "this category" is referring to, I would suggest rephrasing it.

- We now clarify that we are referring here to the fact that five parameters are perturbed in only one direction.

L 231: "maximum objective function evaluation for each parameter between the positive and negative 5% perturbation cases": The "between" makes it sound like all parameter values in the +/- 5% range were tested, which I do not think is the case. I would further suggest using "value" instead of "evaluation".

- We have now updated this sentence to make it clear that we are referring to the maximum of the two objective function values corresponding to the 5% positive and negative perturbations.

L 480: "This shows that the water is clearer than initially estimated.": I would be a bit more careful in the wording when interpreting estimated parameter values; my suggestion: "This result suggests that the water is clearer than initially estimated."

- Thank you for pointing this out and we have updated the wording as suggested by the reviewer.

L 636: "The run time for a single model evaluation is approximately 5 min.": Run on a single node, single core, or what configuration?

- The text has been updated to indicate that the cost of running the model is 5 min on a single core.

Fig. B1: The apparent lack of correlation between initial sampling and final objective function values may suggest that it is not important to start the gradient-based optimization at a low value, but rather to start the optimization at different locations in parameter space to avoid local minima. But any additional experiments would be beyond the scope of this study.

- The reviewer makes a very good point, since we are subject to the local nature of the gradient based method. Falling into local minima is a primary concern which suggests that it may be worth considering relative locations of the initial parameter sets for the optimization runs. The improvement in the baseline cases emphasizes the fact that we cannot rule out the possibility that a parameter set with a higher initial error but in a different region of the parameter space could result in better agreement than that achieved by the 20 best randomly sampled parameter sets in our results. Using the best cases was assumed to be the best approach for identifying regions of relatively low error to proceed with an optimization runs.

  That being said, this methodology could be modified to use a different criterion for selecting the initial parameter sets. For example, instead of only using the error based objective function, one could

incorporate some measure of distance to the selection criteria to ensure that we are getting reasonable coverage of the parameter space. As mentioned, this was not done in this work but it is worth highlighting and should be considered in the future. The major issue that will be faced in taking on this will be determining a way to ensure sufficient coverage of the parameter space for such a high-dimensional parameter space, since this is what motivated the random sampling in the first place. This a discussion of this important point raised by the reviewer has been added to Appendix B.

We thank the reviewer for these useful comments, and the paper has now been revised to address all the above points. Sincerely, the authors.