Response to Reviewer #1

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. Changes made to the text at the request of the reviewer have been highlighted in red in the revised manuscript. In the following, reviewer comments are repeated in blue italics and our responses are provided in the bulleted sections of text.

The article presents a hybrid optimization method which is implemented using the optimization toolkit DAKOTA. The method is applied in order to optimize the parameters of the recent BGC model configuration BFM17 developed by some of the papers' authors (Smith et al., 2021).

The proposed method is a combination of (many) Latin hypercube samples (LHS) of the parameter space and a gradient-based optimization (gradient search, (GS)) using some best-ranked LHS samples as initial search points.

Although both, LHS and GS have been examined before in combination with population-based search methods like genetic algorithms (GAs) the proposed combination of LHS and GS seems to be new.

The application to the BFM17 BGC model is validated in a twin experiment setup where it is able to detect at least the misfit-sensitive parameters. A drawback of the twin experiment is that only a time window of 30 days is considered (according to line 294 of the paper) which does not represent the full annual cycle.

- In response to this helpful comment, as well as a similar comment from the other reviewer, we have performed an additional Twin Simulation Experiment (TSE) that more closely matches the annual, five-field optimization that is the focus of the parameter estimation at the BATS and HOTS locations. The results of this TSE are included as a new Figure 6 in the revised paper, with substantial additional text included at the end of Section 4.1 to discuss the results of this TSE. Briefly, we found in the annual TSE that more parameters had a relative importance greater than 0.01, as compared to the 30-day tests. This is indicative of a more complex optimization problem where many more parameters can affect the results. Consequently, although we recover the baseline parameter values across the range of relative importance values, there are still some parameters that we do not fully recover. However, most parameter values that do not reach the baseline value do at least approach the value. As we now discuss at greater length at the end of Section 4.1, the comparison of the 30-day and annual TSE results demonstrates the challenge of estimating many sensitive parameters in a complex objective function space, even when using a gradient based approach.

A calibration of all model parameters against 1D observations yields a new and justified parameter set for BFM17 (only a manually tuned parameter set has been provided before):

An RMSD type model-data misfit w.r.t. five tracers does significantly improve in comparison to the manually determined parameter set and, consequently, model simulation results do better agree with their observational counterparts (average mismatches drop from several standard deviations to less than one standard deviation).

However, even more convincing would be a calibration of BFM17 coupled to a global circulation model, and against global 3D observations, since (according to Smith et al., 2021) BFM17 is designated to be used in global 3D model configurations.

- We completely agree that this would be the ideal approach. However, even with the cost savings enabled by a smaller BGC model such as BFM17, global circulation models (GCMs) would still be extremely expensive to evaluate many tens of thousands of times, as is required even when using a gradient-based approach. It is common in optimization approaches to use surrogate or lower-fidelity models to accelerate the optimization process, even when the intended application of the optimized parameters is a higher-fidelity simulation. In this sense, the current approach effectively employs

POM1D as a physics-based, low-cost replacement for a GCM. Because this is such an important point, we now re-emphasize this point in the final paragraph of the conclusions section of the paper.

It would also be interesting to see, how a population-based search method (alternatively incorporating LHS and GS) would perform in comparison to the proposed method. I notice that the computational demand of 25,000 function evaluations for LHS would be comparable with many iterations of, e.g., an evolutionary algorithm.

- We agree that a genetic or evolutionary algorithm could be a promising alternative strategy to the optimization problem. We did in fact consider such an approach and estimated the smallest possible population size to be roughly 360. With such a large population size, the number of iterations required to meet reasonable convergence criteria was prohibitive; initial testing with this approach did not produce converged solutions. By contrast, a major advantage of the initial LHS step in the present method is that we can run many model evaluations in an embarrassingly parallel fashion. So, while the computational cost of running many LHS samples may be comparable to running a genetic algorithm (GA) to convergence, the real time taken to perform the optimization would be much less since there are limits to what can be parallelized when implementing a GA. It is conceivable that truncated GAs could be used instead of the LHS, and we have mentioned this approach in the last paragraph of Section 2.2. To address the broader point from the reviewer, we have also added a paragraph to the Conclusions.

Summarizing, I suggest a minor revision to mention hybrid population-based search methods using LHS on the one hand and GS on the other hand. Also, though the main focus of the paper is on presenting and testing the proposed optimization method and providing a suitable parameter set for BFM17, I wonder if the calibration experiments reveal some general relationships between parameter values and model skills, one could elaborate on?

- As outlined in our response to the previous reviewer comment, we agree that it is important to mention hybrid population-based approaches and have added text to this effect in Section 2.2 and the Conclusions. We also agree that it would be interesting to determine if there are general relationships between parameter values. As mentioned by the reviewer, the primary focus of this paper was on the calibration and, due to the number of parameters we are optimizing simultaneously, it is difficult to succinctly interpret the calibrated values for the full set of 51 parameters. Nevertheless, we have added some text in Section 4.3 discussing the relationship between two specific parameters, as an example of what this type of analysis would look like.

Some technical corrections:
- The DOI link of Vichy et al., 1998, does not work in the paper PDF.

- Thank you for finding this error. The link is split across two lines, which is why it does not work. We will ensure that the link is on one line only in the final manuscript (we were unfortunately not able to control this in the revised paper).

- Line 176: "A detailed description ... has been outlined in detail ..." (2 x "detail", drop one)

- This has now been corrected.

- Line 305 ff: the parameter disturbance seems to be 20% instead of the stated 10%

- In the initial twin simulation experiment, the 10% perturbation is from the nominal value. The range of values is then set to +/- 25% of the nominal value. So, the perturbation will not have a 0.1 difference initially, because the normalization is not based exactly on the nominal value but the upper and lower bounds. In the initial TSE experiment we get a constant 0.2 shift because the parameter range is 50% of the nominal parameter value as we are using +/- 25% of the nominal value (10% of 1 is 20% of 0.5). In the additional TSE, the bounds are not a percentage of the nominal parameter value so the initial displacement is not a constant proportion of the parameter range.

- Line 305: The definition of S^ is difficult to read. Perhaps replace "sensitivity metric S(pi)" by "sensitivity factor S(pi)" and explicitly write "S^(pi)=S(pi)/Smax with Smax=max_i(S(pi))"

- Thank you for this recommendation. Based on this comment and one from the other reviewer, we have added Eq. (6) to make the calculation of the sensitivity factor clearer, as well as updated the text to add clarity.

- Lines 299 and 303 and Table 1: what does "nominal parameter value" mean? Is it the "baseline parameter value" from the manual calibration by Smith et al.?

- Yes, this is correct, the nominal parameters are the baseline parameter values taken from Smith et al. The wording has been changed to consistently use baseline to avoid ambiguity.

- Figure 5: what is pi^ and po^? Is it the parameter values, normalized w.r.t. their credible interval length?

- This is correct, $\hat{p}_i$ is the normalized parameter value being tested by the optimization algorithm, while $\hat{p}_o$ is the normalized baseline parameter value. The parameters are normalized between 0 and 1 with respect to their upper and lower bounds. We have added a comment to the caption of the relevant figures clarifying these parameters and to the main text. The clarification is included with the addition to address the concern with respect to Line 305 above.

- Table 1: normalized RMSD values? It is according to Equation (2), i guess?

- This is correct, the normalized RMSD values are calculated according to Eq. (2). The table caption now specifies this to avoid ambiguity.

- Line 364: "year round" => "all year round"

- This has now been corrected.

- Line 366: "maxima" => "maximum"

- This has now been corrected.

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