Geoscientific Model Development Discussions

Geosci. Model Dev. Discuss., referee comment RC3
https://doi.org/10.5194/gmd-2021-175-RC3, 2021

Comment on gmd-2021-175

Benoit Pasquier (Referee)

Referee comment on "A derivative-free optimisation method for global ocean biogeochemical models" by Sophy Oliver et al., Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2021-175-RC3, 2021

This study presents a novel, efficient, and robust optimization algorithm, named DFO-LS, for improving the skill of global biogeochemical models. Despite the potential of the proposed methodology, I believe major revisions are required before this manuscript can be published.

Given the complexity of biogeochemistry in the real ocean, parameterization of its unresolved processes is imperative for models of global marine biogeochemical cycles. In order to improve the ability of these models to fit the growing observational data, modellers have traditionally relied on common physical sense, marine biogeochemical expertise, or simply manual tuning and experimentation. Indeed, automated (programmatic) parameter fitting is generally problematic because of the prohibitive computational costs associated with global marine biogeochemical cycles, the most advanced of which can require months of computation time on large high-performance computing clusters. Oliver et al. thus present us with an inexpensive alternative that could potentially allow for the efficient optimization of parameters even in the case of sparse or noisy observations to fit. Hence on the face of it, the methodology presented offers a welcome tool on the way to improve the accuracy of wide swathes of oceanographic and climate models.

In order to illustrate the performance of DFO-LS, the authors set up a benchmark of optimization runs. A control simulation of the Model of Oceanic Pelagic Stoichiometry (MOPS-2.0) with known parameter values serves as a target reference for the benchmarks. Then, starting from slightly altered values for 6 parameters, different combinations of algorithm and objective functions are put to the test to try to recover the target parameter values. Each optimization requires many evaluations of an objective function, each of which requires a 3000yr MOPS-2.0 simulation in order to compare the resulting simulated tracers to the "synthetic" observations (taken from the control MOPS-2.0 simulation). A previously used optimization algorithm, CMA-ES, is used as a baseline for the benchmarks. The DFO-LS shows promising performance with one run recovering the 6 parameters in 40 MOPS-2.0 simulations compared to 1200 for CMA-ES.

However, as documented in Table 4, only in 1/6 cases does DFO-LS "recover" all 6 parameters. In other words, a pessimistic take from the study is that DFO-LS "failed" 5

out of 6 times. The authors even state: "DFO-LS had not begun to tune this parameter for one of the experiments (exp_d2) before we terminated it at a maximum of 70 evaluations, although it did find KPHY when initiated from a different starting point (exp_d1). CMA-ES also had difficulty in tuning KPHY and only started optimising this parameter after all the other parameters were recovered at ~1200 evaluations. The maximum number of DFO-LS evaluations was set to 70 as it is a sequential algorithm, therefore it was impractical to allow too many more evaluations. Had it been allowed to run longer the expectation is it would begin tuning KPHY once the other 5 were sufficiently tuned, as was the case with CMA-ES." That is, DFO-LS originally failed 5 times out of 5, until a favourable starting state was used in one experiment that eventually succeeded. Unless I am confused, it appears that the robustness of the proposed DFO-LS algorithm to tackle real-world problems hinges on the expectation that it could be just as efficient as CMA-ES and — rarely — be much better. In my opinion, this leaves a gaping hole in the arguments supporting the claim that DFO-LS is "inexpensive and robust to apply to the calibration of complex, global ocean biogeochemical models". As it stands I would recommend accepting the manuscript only after thoroughly addressing the robustness of the approach by running many more simulations from different starting points (and also after consideration of the other points made below).

- Does the paper address relevant scientific modelling questions within the scope of GMD? Does the paper present a model, advances in modelling science, or a modelling protocol that is suitable for addressing relevant scientific questions within the scope of EGU? Yes.
- Does the paper present novel concepts, ideas, tools, or data? Yes.
- Does the paper represent a sufficiently substantial advance in modelling science? Yes.
- Are the methods and assumptions valid and clearly outlined? Yes.
- Are the results sufficient to support the interpretations and conclusions? No.
- Is the description sufficiently complete and precise to allow their reproduction by fellow scientists (traceability of results)? In the case of model description papers, it should in theory be possible for an independent scientist to construct a model that, while not necessarily numerically identical, will produce scientifically equivalent results. Model development papers should be similarly reproducible. For MIP and benchmarking papers, it should be possible for the protocol to be precisely reproduced for an independent model. Descriptions of numerical advances should be precisely reproducible. Did not try.
- Do the authors give proper credit to related work and clearly indicate their own new/original contribution? Yes (apart from some oversights).
- Does the title clearly reflect the contents of the paper? The model name and number should be included in papers that deal with only one model. Yes.
- Does the abstract provide a concise and complete summary? Yes.
- Is the overall presentation well structured and clear? Yes, apart from the table that should be presented as figures instead.
- Is the language fluent and precise? Yes.
- Are mathematical formulae, symbols, abbreviations, and units correctly defined and used? Yes.
- Should any parts of the paper (text, formulae, figures, tables) be clarified, reduced, combined, or eliminated? Yes, see the points below.
- Are the number and quality of references appropriate? Yes.
- Is the amount and quality of supplementary material appropriate? For model description papers, authors are strongly encouraged to submit supplementary material containing the model code and a user manual. For development, technical, and benchmarking papers, the submission of code to perform calculations described in the text is strongly encouraged. Yes.

Below is a list of line-by-line items, suggestions, and comments.

- l. 1: "performance" of biogeochemical models could be confused with computational cost (the way "performance" is used when talking about DFO-LS). What about "skill"?
- l. 34: I am unsure the Li and Primeau (2008) citation is an application of the Transport Matrix Method. (Maybe clarify or remove it?)
- l. 37: "use finite differences or adjoint" This is an inexact distinction of cases. Derivatives can sometimes be derived symbolically (manually or computationally) and evaluated directly (see, e.g., Dickinson and Gelinas, 1976; doi:10.1016/0021-9991(76)90007-3). Most often symbolic derivations are impractical, so one falls back on numerical techniques, such as finite differences. But there are other more efficient and accurate methods (see, e.g., Griewank and Walther, 2008; doi:10.1137/1.9780898717761).
- l. 37–38: "to calculate derivatives" This is technically incorrect. Gauss-Newton and other derivative-based algorithms *use* derivatives but *do not calculate* them.
- l. 38–39: "They can be both computationally expensive and generally less robust on or even unsuitable for noisy problems" Is there a reference for this? While I am convinced that the authors are correct with regards to the pitfalls of a derivative-based algorithm in the case of noisy problems, I am failing to see a strong argument for computational efficiency. While it is true that evaluating a derivative is not free, it provides information that can drastically improve the convergence rate of the optimization algorithm. This needs to be discussed in more detail in my opinion.
- l. 58: "interpolated" I think one could argue that the authors here mean extrapolated.
- l. 60 "Section 4" and "Section 5" should be spelled out for consistency.
- l. 73: Are the 6 parameters the same as those optimized by Kriest et al. (2017)? If so, this could be made clearer.
- l. 86: "In general" Is there a review to reference here?
- l. 86: "quite" is unnecessary
- l. 87–88: I find the "To determine (...) synthetic observations." sentence hard to parse. Maybe it can be reworded for clarity?
- l. 89–90: Suggestion: "the global minimum [is zero] and optimal parameter values are known"
- l. 90: Suggestion: "We compare the performance (...)"
- l. 95: Why not use a differentiable bijection mapping the range to the real line?
- l. 96: Can "various" be replaced with a more descriptive term? Maybe "randomized"?
- l. 96: It might be useful to briefly describe the covariance matrix there.
- l. 96–98: "It returns only the parameter configurations which provide the best misfits to a multivariate normal distribution of parameters, then in the next iteration it randomly draws several more parameter configurations, and repeats" is unclear. What is "it"? Which "multivariate normal distribution"? How many is "several"? This sentence sounds like the description of a brute-force random search, which paints an unfavourable picture of what CMA-ES actually does.
- l. 98: What is a "population"?
- l. 99: "and therefore aim" reverses the logic. The "aim" is to converge towards the best estimation from the onset. "Therefore", the algorithm employs the strategy to "move" towards lower misfit values.
- l. 102–103: This "In order (...) in practice" sentence could be rearranged. Also, an indication of how many function evaluations would be useful. (And "quite" is unnecessary.) Suggestion: "In practice, CMA-ES requires thousands of function evaluations (...)"
- l. 104: What does "was sourced" mean? Is that the exact code? Is it archived and

publicly accessible?

- l. 109: If $\boldsymbol{x}$ is bounded, then starting with $\boldsymbol{x} \in \mathbf{R}^n$ is misleading. What about: "$\boldsymbol{x} \in D$ a bounded domain of $\mathbf{R}^n$"?
- l. 109–111: Suggestion: "DFO-LS can take into account individual terms of the misfit function and use their structure to improve convergence"
- l. 114: "provably": If there is a convergence proof, then it should be cited. Unless this was supposed to be "probably"?
- l. 117: What is a typical n?
- l. 117: "for a total of $n + 1$ function evaluations". I think this can safely be removed.
- l. 117: What does "nearby" mean?
- l. 118: Suggestion: "only one misfit function evaluation is needed"
- Fig. 1 Caption:
  - How is the information "combined"? Are the squared misfits simply summed over? If that's the case, it should be stated as such. Otherwise, maybe some clarification of what goes on would be useful.
  - How can the misfit function be "evaluated if accepted". It seems this is the other way around.
- l. 119–121: This is only important if the initial sampling constitutes the bulk of the computation. Is that the case in general?
- l. 133: "minima ." (space before dot)
- l. 139: "much" is not needed.
- l. 139: "This is much more computationally expensive than a soft restart" Why is it the case?
- l. 148–149: Are these personally communicated regions available in a public archive? Reproducibility hinges on such availability.
- Eq. (2): Does $V_i^{i \neq j} = 0$ when $i \neq j$? If so, I would suggest just having V_i instead, and summing over only $i \neq j$ (instead of summing, potentially redundantly, over all i)
- l. 158: "eddies" can be large. Maybe "unresolved eddies" is clearer?
- Fig 2. Caption: Which reference defines the 19 regions? Henson et al (2010) or Weber et al (2016)?
- Eq. (30) + many lines: "f_{global}". Usually non-variable subscripts are typeset upright. Also, "global" is misleading, since there are many tracers. On the other hand, for volumes, the authors use "T", for "total", I guess. Maybe no subscript for this "total" f? Or maybe swap the "T" and "global" subscripts throughout?
- l. 171: How is the interpolation done?
- l. 172: Why three noise realizations?
- l. 189 and throughout: I kept going back to read what differentiated each experiment from the other. Maybe the authors can find more expressive names for their experiments? E.g., `exp_d1` could be `D_noise` and `exp_d2` could be `D_smooth`? `exp_drngi` could be `D_randi`? And `exp_d1_sparse` could be `D_smooth_sparse`, and so on.
- l. 183–204: What about an experiment combining sparse and randomized observations?
- Table 2. It is unclear what all the settings do. Also, all the caption experiment names do not match the main text.
- l. 208: the authors say they "plot" values but instead show a table. Indeed, Table 3 looks like it would deserve to be turned into a plot with 6 panels (6x1, one for each parameter) with each experiment on the x-axis, and the optimized value on the y axis. With a color code that conveys the groupings (smooth, sparse, noisy, and so on). The same goes for Table 4, which could be turned into a combination of bar plots (with a broken axis to cater for a large number of evaluations for CMA-ES rather than a misleading logscale).
- Table 4. Some of the experiment names do not match the main text. Is "subsel" the same as "sparse"?
- l. 277: Maybe I read this incorrectly, but it seems the authors report that only 1/6 of the DFO-LS experiments recovered all 6 target parameters (Table 4). This is swept

under the rug here. I think it would be useful to discuss the failures of convergence for some parameters here.

- l. 290–297: This seems like a significant caveat. The justification for capping the number of evaluations for DFO-LS to 70 is unsubstantiated. Combined with the above comment it seems that the authors' (5?) original experiments with DFO-LS *all* failed to recover all 6 parameters, and they then added an extra experiment with a different starting point for K_PHY that is at least twice as close as its target value (Figure 6). In my opinion, this places the robustness of the approach under question, and therefore I would recommend additional experiments with randomized starting values.
- l. 315: While technically feasible, I would consider using "oxygen concentrations" as a constraint rather than the "location of oxygen minimum zones", which is subjectively defined by an arbitrary threshold.
- Figures 7, 9 captions should repeat what the vertical arrows mean (soft restarts)
- One of the main citations of the manuscript (Cartis et al., arXiv, 2018) should be replaced by the more recent and, importantly, peer-reviewed, version (Cartis et al., Optimization, 2021, doi: 10.1080/02331934.2021.1883015).