Reviewer’s comments on Buchanan et al. "*Optimisation of the World Ocean Model of Biogeochemistry and Trophic-dynamics (WOMBAT) using surrogate machine learning methods*"

March 17, 2025

Overview

Research summary

(1) In this paper, the authors introduce a new version of the marine biogeochemical model, WOMBAT-lite. The updated model incorporates several new processes and requires tuning using observational datasets. The authors leverage machine learning methods to optimize the model. First, a machine learning model is trained on outputs from 512 experiments, each with different parameter sets (same parameters but different values), to predict 8 root mean square errors (RMSEs) relative to observational products. This trained machine learning model is used to perform a sensitivity analysis, identifying the parameters that most significantly impact the RMSEs — a total of 11 parameters are identified. Then, a second machine learning model is trained using outputs from 512 new experiments, focusing solely on the 11 selected parameters, to replicate a cost function that measures the deviation from observations. This trained model is utilized for optimization analysis to determine the optimal values and ranges of the 11 parameters. The optimized version of the model is then compared to observational data and a previously unoptimized version of WOMBAT. The new optimization approach proves to be successful, with the optimized model showing improvements across all evaluated variables, albeit to varying degrees.

Overall assessment

(2) This technical paper is clearly written and presents a novel method for optimizing the new version of the biogeochemical model. The authors have made great efforts to describe the machine learning approach, detailing the sequential steps used first to identify key parameters and then to optimize them. I have a few major comments, that summarize as: 1) Additional justifications are needed to ensure that ten years of simulations are long enough, in relation to the model’s intended usage and the influence of the optimization strategies. 2) Further discussion on alternative model tuning approaches would be beneficial. 3) The advantages of this optimization approach over simply selecting the best of the 512 experiments need to be further explained. 4) Further details on the machine learning algorithms may be necessary. I elaborate on these major comments below. Most of these comments can probably be addressed by adding text in the introduction or conclusive section of the manuscript. Additionally, I have a list of minor or specific comments and suggestions. Finally, given the length of the manuscript and the number of figures, any content that could be shortened would be beneficial. I tried to provide some suggestions. Overall, this is a quality piece of work, and I recommend it for publication after some minor revisions.

Major comments

(3) My first concern is the ten-year-long simulations used for conducting the sensitivity/optimization analysis. I understand the constraints associated with running a large number of longer simulations. Nonetheless, I would like to have a clearer picture of the equilibrium achieved in ten years. What are the drifts at various depths? Although ten years may be enough for the surface ocean, what about the deeper ocean that can impact the surface on longer time scale? The authors have noted a drift away from the optimal fit when simulations are extended to 100 years. Could the authors discuss the risk of further drift in these simulations? Additionally, is the drift of the optimized simulation weaker compared to the best initial simulation or any of the 512 simulations? If the drift is indeed weaker, this would be a strong point in favor of the optimization strategy.

(4) Somewhat connected with the previous point, could the authors specify/discuss the intended usage for which the model is optimized? For instance, will the model perform equally well when incorporated into an Earth System model for climate simulations? I am curious whether the optimization based on ten-year simulations and focusing on the surface ocean would remain effective during the longer processes, such as a thousands of years spin-up to achieve preindustrial equilibrium, followed by hundreds of years for simulating historical and future climate conditions.

(5) Another major comments reside in emphasizing the utility of this optimization approach. Considering that the ML approach will identify the optimal parameter set in the space that has been explored, why not using the best performing of the 512 experiment? Could the authors comment a bit around that? I guess the optimization approach also provide some uncertainty and sensitivity quantification but can't we get the same from the 512 experiments? What about comparing the optimal simulations with the best of the 512 initial simulations?

(6) I would appreciate further discussion on model tuning and alternative approaches, particularly those employing ensembles/surrogate models/emulators. For instance, the work of Singh et al. (2025) and Williamson et al. (2017) seem relevant.

(7) Last major comment, while the approach is explained very well, I believe it lacks a detailed description of the machine learning techniques used. This aspect is beyond my area of expertise, so I cannot adequately assess it. However, a thorough explanation might be essential for reproducibility and for the benefit of readers interested in machine learning methods applied to a specific problem.

Minor and specific comments

Introduction

(8) l. 47: Reformulation suggestion for clarification: "This component represents the growth..." -> "This component, in its simplest form, represents the growth..."

(9) l. 96: There is a typo, remove "apt" in "As an apt example, ..."

Methods

(10) l. 163-179: I first counted 7 observation products before understanding that chlorophyll data where actually providing 2 product: depth of the maximum and surface chlorophyll. Can the authors reformulate to clarify?

(11) l. 172: Why optimizing to the CO_2 flux and not pCO_2? The CO_2 flux is itself estimated from pCO_2, thus introducing an additional source of potential errors in the observational target. Can the authors clarify/justify the choice?

(12) l. 225: Can the authors gives some example and eventually some reference (if they exists) to clarify what they have in mind?

(13) l. 239: I find misleading using "Southern Ocean" for all the ocean south of 20°S. Here it includes some of the subtropical gyres. I found this misleading in some other part of the manuscript. Can the authors use an other term?

(14) l.271: For clarifying "... by comparing GPR predictions with WOMBAT-lite RMSE...", the authors can be specific and say that they compare the GPR predictions of RMSE with WOMBAT-lite RMSE.

(15) l. 267-268: Is there one training for the 8 RMSE or 8 training? i.e 1 GPR model predicting 8 RMSE or 8 GPR model? Can the authors clarify?

(16) l. 276: Why not doing the optimization directly with the 24 parameters? Why having first a sensitivity analysis? (ie. why not doing step 10 just after step 5?) Can the authors clarify?

(17) l.285-297: Similar to comment 15, can the authors clarify? Is there one training for the 8 cost functions? Or one for the global cost function? Or 8 training for each cost function? To be entirely clear, the model is trained on a cost function (defined in Eq. 1 or 2), whereas for the sensitivity analysis, it focuses on the RMSE. Is that correct?

Results

(18) l. 349-356: Why not including PI^0 in the most important parameters? On figure 5b it has a value of 0.05 for the depth of chlorophyll maximum, i.e. 5% if I understood correctly the threshold. Can the authors clarify?

(19) l. 405: The depth of the chlorophyll maximum is indeed heavily influenced by α_a and γ_p^0 but only "in interaction" (i.e. panel b of Figure 5). It is the only target (with maybe iron but it is not as marked) that have 2 parameters standing out so much in interaction while no parameters seems to strongly influence it at first order. Can the authors comment one that?

(20) l. 412: It reads a bit strange to mention 2 additional parameters and not have them specified right away. There are mentioned just a bit after, but maybe the authors can slightly reformulate the paragraph so that 2 additional parameters are specified right after being mentioned.

(21) l. 413-418: Any reason why these parameters where not included in the sensitivity analysis? Can the authors clarify?

(22) l. 437: "... due to its complexation with organics..." is not clear to me. What is organics? Organic form of dFe? Can the authors clarify?

(23) l. 506: In the subtropical or subpolar gyres?

(24) l. 504-506: It seems to me there is a very shallow chlorophyll maximum in the subtropical gyres that is not present in the observation (Fig. 8k, l). Any explanation? This is a bit contradictory with the "too deep maxima in the gyres." Can the authors clarify?

(25) l. 486-506: What about the CO2 flux? Can the authors mention that it receive a detailed analysis further in the paper.

(26) l. 483-484: "Our optimised WOMBAT-lite manages to show some improvement..." I don't find Figure 10b to be the most effective way to illustrate this. Instead, Figure 11 demonstrates the improvement in the seasonality of the CO2 flux through better local correlation, whereas the correlation on Figure 10b actually decreases. I agree nonetheless that on figure 10 the amplitude seems to be a bit closer. It also seems to me that comparing a transition zone based on average is rather complex. Could the authors reformulate to provide a more precise explanation/description?

(27) l. 593-599: This paragraph seems to be a repetition of the the one just before. Can the authors merge it and combine figure 10 and 11 (maybe not all panel of figure 10)?

(28) l. 593: The improvement are clear for the equatorial band (10S to 10N) and the high latitude (>50N and <30S). For the rest (10S-30S and 10N-50N) the unoptimized model is already well correlated with the observations and even better (darker red) in a few places. Can the authors reformulate to provide a more precise explanation/description?

(29) l. 593-599: Can the authors mentioned that the observations also have some caveats (Hauck et al., 2023; Gloege et al., 2021)?

(30) l. 602: "the unoptimised version of the model exceeded the observations." The optimized version also exceed the observation on figure 12a. Can the authors clarify their point?

(31) l.601-611: On figure 12 we see that the optimized version has a much better seasonal cycle in the high latitude (in line with what we see on figure 11). This important improvement does not appear so much on the annual average because of compensation between the errors in summer and winter. This should be stated more clearly.

Appendix

(32) l. 1430: There is typo, "is remineralised" is repeated.

Figures

(33) Fig. 1: In the text of step (6) in the figure, consider to keep the same format as the other steps. Maybe something like "Run a global sensitivity..."

(34) Fig. 4: Can the authors add the performance of the optimized model and the 20 optimal experiments after 10 and 100 years.

(35) Fig. 4: Why is the standard deviation of sinking detritus, NPP and chlorophyll so weak for many experiments? Can the authors explain/comment?

(36) Fig. 5: Should the sum of each row in the table be equal to one, including both tables a) and b)? Could the authors clarify the difference between the first-order and higher-order values? It would be helpful to include an explanation of this in the methods section, for example at the end of section 2.4.1.

(37) Fig. 6: The figure is not much referenced in the text, and because the values are normalized, it is difficult to compare them with Table 1 (I assume that the optimal values and ranges are derived from this figure in some way) and it does not provide a finer understanding of the optimal range. Perhaps it could be moved to the supplementary materials.

(38) Fig. 10: As mentioned in comment 13, I find misleading the term Southern Ocean in the caption.

(39) Fig. 11: Can the authors add the latitude on the map.

References

Tarkeshwar Singh, François Counillon, Jerry Tjiputra, and Yiguo Wang. A Novel Ensemble-Based Parameter Estimation for Improving Ocean Biogeochemistry in an Earth System Model. *Journal of Advances in Modeling Earth Systems*, 17(2):e2024MS004237, 2025. ISSN 1942-2466. doi: 10.1029/2024MS004237.

Daniel B. Williamson, Adam T. Blaker, and Bablu Sinha. Tuning without over-tuning: Parametric uncertainty quantification for the NEMO ocean model. *Geoscientific Model Development*, 10(4):1789–1816, April 2017. ISSN 1991-959X. doi: 10.5194/gmd-10-1789-2017.

Judith Hauck, Cara Nissen, Peter Landschützer, Christian Rödenbeck, Seth Bushinsky, and Are Olsen. Sparse observations induce large biases in estimates of the global ocean CO2 sink: An ocean model subsampling experiment. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 381(2249):20220063, May 2023. doi: 10.1098/rsta.2022.0063.

Lucas Gloege, Galen A. McKinley, Peter Landschützer, Amanda R. Fay, Thomas L. Frölicher, John C. Fyfe, Tatiana Ilyina, Steve Jones, Nicole S. Lovenduski, Keith B. Rodgers, Sarah Schlunegger, and Yohei Takano. Quantifying Errors in Observationally Based Estimates of Ocean Carbon Sink Variability. *Global Biogeochemical Cycles*, 35(4):e2020GB006788, February 2021. ISSN 1944-9224. doi: 10.1029/2020GB006788.