Overview

Thanks to the authors for their thorough response to my comments. The additional text in the revised manuscript addresses my concerns. In particular I find the additional figure S6 complementing very well the figure 7 to show how the interest of using the surrogate model to run an extensive optimisation. One remaining question to me: how the optimisation affect the drift of the model? If it could be shown that the optimisation reduce the drift that would be an additional strength of the method. Maybe by comparing the drifts in the best of the sensitivity experiments and the best of the optimal experiments, both run on 100 years?

Because we train the GPR model on the 10th year of output from the full model (with coupled and evolving physics), we do not expect the selection of optimal parameters to counteract any long-term drift in the ocean state. We are therefore unable to show that the method reduces the long-term drift in the ocean nor do we necessarily expect it to, although we acknowledge that it may provide some guardrails against rapid deviations.

We have added the following sentence to clarify this:
Line 261: "Our optimisation therefore does not counteract the potential for long-term drift but does provide some guardrails against rapid deviations from the initial state with respect to dissolved nutrient and carbon fields. 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). However, by assessing the cost function after 100 years and selecting the best performing of the optimal experiments we likely select for a parameter set with minimal long-term drift."

In addition, I have a couple of minor and specific comments, some of which I did not catch during the previous review. Once these minor adjustments are made, the manuscript will be ready for publication.

Minor and specific comments

Abstract

(1) l. 27: I think "earth" in "...of earth system models..." is more commonly written with a capital E ("...of Earth system models...")

Corrected.

Methods

(2) l. 318-333: Is it one surrogate model trained for each cost function J (so 8 surrogate models in total) or one for the global cost function (equation 2)? Can you clarify? I had a similar question for the sensitivity analysis during the first round of review. The authors clarified very well, notably by using plural forms or emphasizing singular terms when appropriate.

We thank the reviewer for their thoroughness. We have clarified this in the text on line 346 by extending the sentence:

"Therefore, we aim to select parameter sets that optimise overall model performance"

To:

"Therefore, we select parameter sets that optimise overall model performance by summing the prediction errors from all eight surrogate GPR models."

Results

(3) l. 463-464: "The 512 experiments were used to calibrate the global cost function synthetically using the machine learning model." I do not think this sentence really state what was done. What I understood is that: the 512 experiment are used to trained the machine learning model to reproduce the global cost function. The

trained model is then used to look for the optimal values of the parameters, i.e. the ones minimizing the global cost function. It do not think we can say that the global cost function is calibrated. Can the authors rephrase?

Agreed. We have corrected this by including your interpretation within the text.

(4) l. 478-479: "We also note that the model predicted optimal values that often aligned well with ecological theory." Is the initial range of the parameters' values much wider than the theory? Otherwise it seems expected that the optimal 31 *values are aligned with the ecological theory and thus the argument is not very strong and could be removed.*

Here, we refer to and immediately discuss the fact that the optimal values of two parameters were in a ratio that reflected theory and/or observational evidence. So, while we chose prior ranges that are legitimate, there are ratios of two parameters that are not, and it was encouraging to see that the optimal parameter set included parameter ratios that were in line with observations/theory.

(5) l. 531: "unoptimised biogeochemical model". I assume the authors mean the former version of the model that even if not optimised was considered good enough to conduct analysis and numerical experiment. Correct? This model was still tuned. Can the author specify this? Otherwise, it sounds like the new optimised version is compared to a version that should even not be used for any study.

Agreed. When referring to this prior version of the model, we have now introduced it as a prior model "that did not undergo the same optimisation process but was nonetheless tuned manually". This has been clarified in the text.

(6) Sec. 3.4.3: Is the bloom phenology better than with the unoptimised model version?

The previous version did not have prognostic chlorophyll concentrations, so a direct comparison was not possible.

Summary

(7) l. 666-668: The authors should mention here that the surrogate model can provide the large number of samples at a low computational cost.

We have followed the reviewer's suggestion. The sentence now reads:

Line 712:
"To do so, we used a surrogate machine learning model trained on a limited sample of real model output that provided many synthetic estimates of model performance (i.e., global NRMSE and the global cost function) at low computation cost."

(8) l. 677: Shouldn't it be "statistically" instead of "statically"?

Statically is correct here. We refer to the idea of having time-evolving, or even space-evolving, parameter values, which could potentially even better performance.

We have nonetheless altered the sentence to improve clarity:

Line 725:
"While these approaches can provide optimal parameter values that evolve in time and/or space, the surrogate approach employed herein represents a simple yet valid approach that provides globally optimised parameter values that are fixed in time and space, but importantly without large computation overhead."

Figures

(9) Fig. 5: I do not think emphasizing "at most" and "at least" with bold font is necessary.

Agreed. Corrected.

General comments

The paper by Buchanan et al., provides an innovative and elegant case study for parameter sensitivity analysis and optimization of a ocean biogeochemical model using surrogate models and Bayesian methods. In the study the authors use Gaussian Processes (GPs) to predict model error (specifically, root mean square error - RMSE) for an ensemble of process-based model runs (from the WOMBAT-lite model). These surrogate models are then used to (1) test sensitivity analysis of the WOMBAT-lite model and (2) optimize key WOMBAT-lite parameter values. Using the method the authors show the importance of key parameters, and illustrate significant model improvements, including for nonoptimized metrics such as bloom phenology.

The authors have made significant improvements to the manuscript since the initial submission. However, due to 1) the intended audience of process-based biogeochemical modelers - which do no necessarily have statistical backgrounds and 2) the general novelty of the methods I would like to see a bit more discussion and a few more clarifications which are suggested in detail below.

High-level description of method

The surrogate modeling is novel and should be explained conceptually at a higher level. For instance, it is not initially clear that the surrogate model is used to predict model error - which can be confusing to researchers more familiar with surrogate models that predict the full model fields. While this detail is provided in the methods, stating this clearly in the overview (section 2.1) of the methods section, as well as in the conclusion, would aid in interpretation.

The reviewer is right to ask for greater clarity. Following their request, we have included additional remarks within sections 2.1 and the conclusion that remind the reader that the surrogate GPR model is predicting the univariate, global statistics of model error.

Section 2.1; Lines 133 – 135:
"This computationally inexpensive surrogate was trained on hundreds of real simulations with WOMBAT-lite, predicted global univariate statistics of model error and was able to produce large samples of synthetic results that enabled sensitivity analysis and optimisation (Fig. 1)."

Summary; Lines 712 – 713:
"To do so, we used a surrogate machine learning model trained on a limited sample of real model output that provided many synthetic estimates of model performance (i.e., global NRMSE and the global cost function) at low computation cost."

Definition of priors

The authors use uniform priors which are scaled to range between 0-1. However: 1) the scaling is not mentioned in the text and should mentioned, justified and explained; 2) how the priors were chosen is not clear - ideally references for each value should be provided in table 1; 3) the uniform priors seem to work nicely, but the downside and potential implications of a uniform prior on the posterior estimates should be discussed - in particular since the parameters are likely Gamma or Normally distributed in reality.

We acknowledge the reviewers concerns and we believe we have addressed them completely:

1) *We do normalize the prior distributions to vary between 0 and 1 via a Max-Min scaling, since we know the lower and upper bounds of each parameter. The reason for normalising the prior distribution is to ensure more reliable and efficient exploration of the high-dimensional parameter space during MCMC.*
2) *We have added the appropriate references that guide the a priori range choices in Table 1.*
3) *We want the optimisation to determine the posterior distribution from as generic a distribution as possible. However, we also note that the burn-in phase of the MCMC is designed to forfeit the uniform distribution and its effect on the posterior, so although we do assume a uniform as input to the MCMC procedure, the burn-in substantially reduces its effect on the posterior.*

These concerns have been addressed in what is now written on lines 361 – 379:

"Bayesian optimisation enables iterative learning of optimal model parameters using observational data. Here, we apply uniform priors under the assumption that there is equal probability of the optimal values falling anywhere between what are known lower and upper bounds (Table 1). The priors are normalised from 0 to 1 via max-min scaling, and the normalized cost function value predicted by the GPR model serves as the likelihood function (Reddy et al., 2024a). Since computing marginal likelihood directly is often complex, Markov Chain Monte Carlo (MCMC) sampling is employed, which estimates the posterior distribution without explicit calculation of this constant (Issan et al., 2023). Normalisation of priors between 0 and 1 ensures more efficient exploration of the high-dimensional parameter space (i.e., better "mixing") and therefore more reliable convergence, while normalisation of the cost functions ensures equal weighting of each target field to the solution. Among MCMC methods, Affine invariant ensemble sampling, implemented using the "emcee" Python package (Foreman-Mackey et al., 2013), is selected for its efficient convergence properties. This method uses an ensemble of chains to simplify sampling from anisotropic distributions. Fifty walkers and a stretch move of two are applied, with the first 10,000 steps used as a burn-in phase to ensure convergence, followed by 90,000 additional steps to achieve stable posterior distribution estimates (Foreman-Mackey et al., 2013; Goodman and Weare, 2010). This substantial burn-in phase converts our uniform priors into distributions that look more like the posterior (Foreman-Mackey et al., 2013). In total, 90,000 steps for each of 50 walkers means a total of 4,500,000 random walks, and with a mean autocorrelation time of 690 steps, we achieve 6500 independent samples of the posterior. This number of samples, made possible by the efficiency of the surrogate GPR model, is more than sufficient for convergence. Convergence was also evident by a mean acceptance fraction of 0.22 (Foreman-Mackey et al., 2013), a Gelman-Rubin statistic of between 1.005 and 1.009 for each parameter and visual assessment of the chains (Fig. S6)"

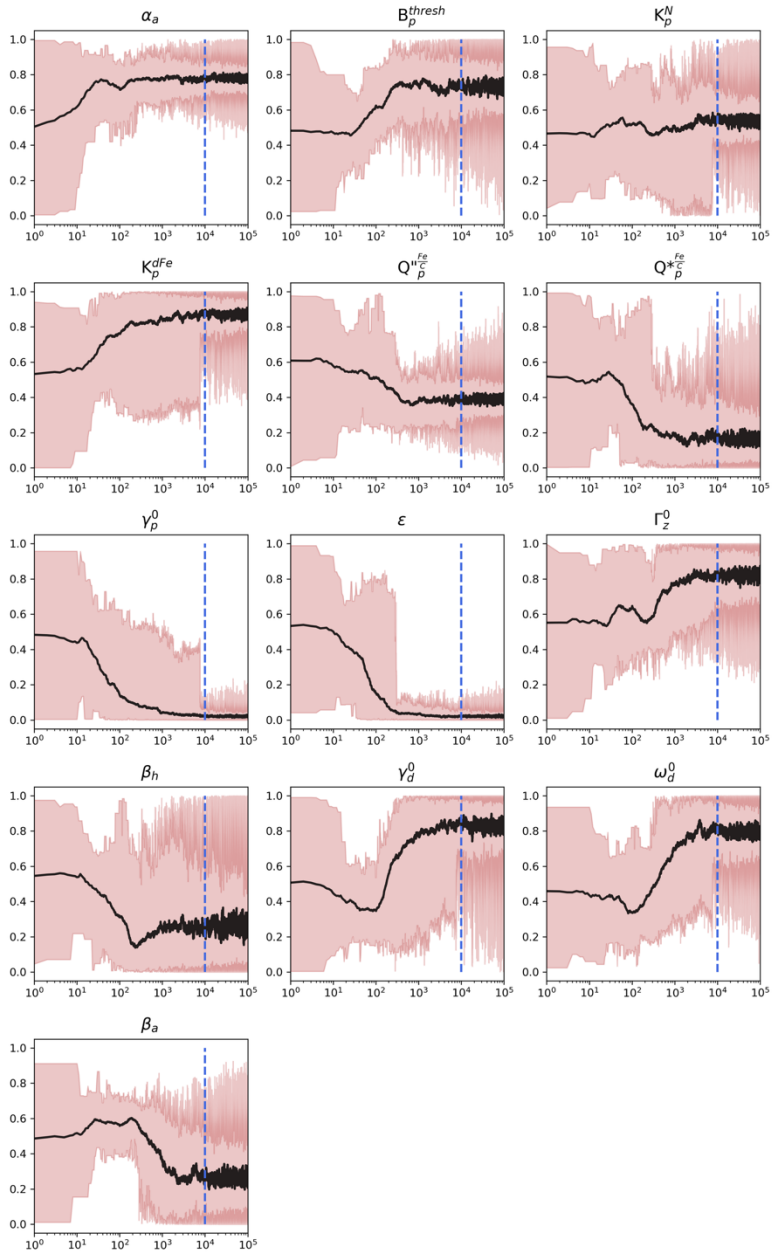

Figure S6. Mean (black line) and range (pink shading) of sampling by all 50 walkers during the Markov Chain Monte Carlo (MCMC) optimisation of the 13 parameters. *The blue dashed line demarcates the transition from the burn-in phase to the actual sampling used to construct the posterior. The x-axis is on a log10-scale, and the y axis is the same for all parameters because we normalised their ranges to 0-1 based on Max-Min scaling. A total of 100,000 steps is taken by the 50 workers.*

As well as the new Table 1 with a new column "Reference".

Table 1. Key ecosystem parameters for WOMBAT-lite and their predefined ranges for ocean-only experiments. Parameter values for other configurations, include for the Earth System Model (ACCESS-ESM1.6), are available at https://github.com/ACCESS-NRI.

Component	Parameter	*a priori* range	Default (optimal range)	Description	Units	Reference
Phytoplankton	α_a	0.25 - 1.25	1.0 (0.89 – 1.16)	Scaler control on phytoplankton maximum growth rates	day^{-1}	Anderson et al. (2021a)
	[(i)]β_a	1.040 - 1.080	1.050 (1.041 – 1.063)	Base for temperature-dependent autotrophy	-	Anderson et al. (2021a)

Component	Parameter	*a priori* range	Default (optimal range)	Description	Units	Reference
	PI^0	1.5 - 3.0	2.25	Initial slope of the photosynthesis-irradiance curve	$(W\ m^{-2})^{-1}$ (mg Chl / mg C)$^{-1}$	MacIntyre et al. (2002)
	[ii]B_p^{thresh}	0.01 - 1.0	0.6 (0.48 – 0.94)	Biomass threshold of phytoplankton for implicit allometric scaling	mmol C m^{-3}	N/A
	$K_p^{N_0}$	0.01 - 3.0	2.0 (1.04 – 2.30)	Half-saturation coefficient for nitrogen uptake	mmol N m^{-3}	Litchman (2007)
	$K_p^{dFe_0}$	0.01 - 3.0	2.5 (2.07 – 2.97)	Half-saturation coefficient for dissolved iron uptake	µmol Fe m^{-3}	Shaked et al. (2020)
	$Q_p^{"\frac{Fe}{C}}$	20 - 100	50 (39 – 64)	Maximum Fe:C quota of the cell	µmol Fe (mol C)$^{-1}$	Twining et al. (2021)
	$Q_p^{*\frac{Fe}{C}}$	4 - 15	10	Optimal Fe:C quota of the cell	µmol Fe (mol C)$^{-1}$	Twining et al. (2021)
	$Q_p^{'\frac{Chl}{C}}$	0.001 - 0.01	0.004	Minimum Chl:C quota of the cell	mg Chl (mg C)$^{-1}$	Geider (1987)
	$Q^{*\frac{Chl}{C}}$	0.02 - 0.06	0.036 (0.020 – 0.038)	Optimal Chl:C quota of the cell	mg Chl (mg C)$^{-1}$	Geider (1987)
	γ_p^0	0.01 - 0.10	0.01 (0.010 – 0.016)	Linear mortality rate of phytoplankton	day^{-1}	Baker and Geider (2021)
	Γ_p^0	0.01 - 0.10	0.05	Quadratic mortality rate of phytoplankton	(mmol C m^{-3})$^{-1}$ day^{-1}	Suttle (1994)
Zooplankton	g_h	2.0 - 4.0	3.0	Scaler control on maximum zooplankton grazing rate	day^{-1}	Rohr et al. (2022)
	ε	0.05 - 1.5	0.05 (0.05 – 0.15)	Zooplankton prey capture rate coefficient	(mmol C m^{-3})$^{-2}$ day^{-1}	Rohr et al. (2022)
	[iii]ϕ_z^p	1.0	1.0	Preference of zooplankton for phytoplankton	-	N/A
	ϕ_z^d	0.01 - 0.50	0.25	Preference of zooplankton for detritus	-	N/A
	[iii]λ	0.6	0.6	Zooplankton assimilation efficiency	-	Anderson et al. (2021b)
	γ_z^0	0.01 - 0.10	0.05	Linear mortality of zooplankton (respiration)	day^{-1}	Anderson et al. (2021b)
	K_z^γ	0.01 – 0.5	0.25	Half-saturation coefficient of zooplankton mortality	mmol C m^{-3}	N/A
	Γ_z^0	0.1 - 1.0	0.9 (0.61 – 0.99)	Quadratic mortality rate of zooplankton (predation)	(mmol m^{-3})$^{-1}$ day^{-1}	N/A
	β_h	1.060 - 1.080	1.065 (1.060 – 1.075)	Base for temperature-dependent heterotrophy	-	Chen et al. (2012a)
Detritus	ω_d^0	5 - 20	18 (12.7 – 19.9)	Scaler to sinking speed of detritus	m day^{-1}	De La Rocha and Passow (2007)
	ω_d^{max}	20 - 50	35	Maximum sinking speed of detritus	m day^{-1}	De La Rocha and Passow (2007)
	γ_d^0	0.025 - 0.1	0.09 (0.064 – 0.099)	Linear rate of (implicit) bacterial remineralisation	day^{-1}	del Giorgio and Cole (1998)
	$R_{\frac{CaCO_3}{detritus}}$	0.01 – 0.15	0.050	$CaCO_3$ to organic detrital ratio	mol C / mol C	Lehmann and Bach (2025)
	$\omega_{CaCO_3}^0$	3 - 10	6.0	Scaler to sinking speed of $CaCO_3$	m day^{-1}	Pantorno et al. (2013)
	$\gamma_{CaCO_3}^0$	0.0005 - 0.01	0.01	Scaler control on (implicit) $CaCO_3$ dissolution rate	day^{-1}	Kwon et al. (2024)
Iron cycling	[iii]Lig	0.7	0.7	Concentration of Fe-binding organic ligand	µmol m^{-3}	Johnson et al. (1997)
	[iii]K_{nanop}^{Fe}	0.01	0.01	Precipitation of Fe' as nanoparticles (in excess of solubility)	day^{-1}	Tagliabue et al. (2023)
	[iii]K_{scav}^{Fe}	0.00005	0.00005	Scavenging of Fe' onto biogenic particles	(mmol C m^{-3})$^{-1}$ day^{-1}	Tagliabue et al. (2023)
	[iii]K_{coag}^{Fe}	0.0001	0.0001	Coagulation of dissolved Fe into colloidal Fe	(mmol C m^{-3})$^{-1}$ day^{-1}	Tagliabue et al. (2023)

[i]Parameter variations not included in initial sensitivity experiments for sensitivity analysis (steps 3-6 in Fig. 1). Only in optimization (steps 7-10 in Fig. 1).
[ii]Parameter range was set equal to 0.01 to 0.1 in the initial sensitivity experiments for sensitivity analysis (step 3-6 in Fig. 1).
[iii]Parameter space not explored.

Choice of surrogate model objective

The authors train the surrogate models to predict RMSE of each parameter configuration. This is an elegant and cost effective approach, but the benefits of this approach should be highlighted and it's use justified in text. It would also be useful to highlight any downsides of this approach compared to a spatially-resolving emulator. For instance, it is conceivable that sensitivity of each parameter is not spatially uniform (e.g. parameters influencing predictive performance in the Southern Ocean could be quite different from equatorial upwelling regions).

Lines 312 – 313:
"NRMSE was chosen as a metric of performance because of our diversity of target fields each with different units, which required normalization to ensure that each contributed equally to an assessment of global performance."

A spatially resolving emulator was avoided in this first instance because we wanted to determine how the simplest possible approach performed. A more sophisticated emulator could be developed later.

Specific comments

• L294: include the short name of the package (UQ-PyL)
Addressed.

• L296: consider defining this as rRMSE (relative RMSE) or NRMSE (as used further down in the text) and then using NRMRSE/rRMSE where appropriate for clarity
Addressed.

• L298: why was a K value of 8 chosen?
Addressed.

Line 313: "K-fold cross-validation is used to evaluate the GPR model's accuracy and we chose 8 folds (K=8) to cleanly divide the 512 experiments and provide a balance between sufficient training (448) and testing data (64) (Reddy et al., 2024a)."

• L329: which hyperparameters were used for these kernels? Was any hyperparameter optimization conducted to find the best values?
Parameter and kernal choices were guided by previous work.
Line 348: "Using a composite kernel—constant, Matern, and white noise kernels with hyperparameters guided by previous work (Reddy et al., 2024a)—"

• L342: how was convergence assessed? A plot illustrating chain convergence should be included, and if any test (e.g. Rhat) was used, these values should be reported.
See answer above referencing Fig S6. We also note here that we do assess convergence and have now reported the Rhat and other statistics in the text.

• L538-545: these estimates should be compared to the literature (e.g. field et al., 1998, falkowski et al., 1998, johnson et al., 2021 - or others as relevant)
We make this comparison later in the Summary section on Line 739.