

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 non-optimized 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 not 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.

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 be 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.

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 its 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).

Specific comments

- L294: include the short name of the package (UQ-PyL)
- L296: consider defining this as rRMSE (relative RMSE) or NRMSE (as used further down in the text) and then using NRMSE/rRMSE where appropriate for clarity
- L298: why was a K value of 8 chosen?
- L329: which hyperparameters were used for these kernels? Was any hyperparameter optimization conducted to find the best values?
- 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.
- 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)