RC4: 'Comment on essd-2023-67'

Anonymous Referee #4

Summary

This study uses a machine learning approach to reconstruct global Ba concentrations in the ocean, and uses the model output to calculate Ba* and barite saturation state in the global ocean. In general this is solid study that provides model output that will be useful to other researchers, and the methodology is sound, with one exception that I detail below. I think that with minor revisions the study should be acceptable for publication.

We're pleased to read that the reviewer felt that this was a solid study and we are grateful for their comments, which have improved the contribution.

Specific Comments

- Line 89: I disagree that mechanistic modeling should be called the "gold standard". A model is useful if one can learn something from it, period. Some mechanistic models are useful, some statistical models are useful.

Fair enough. We took out this language. The new sentence reads:

"In mechanistic or process-based modeling, model outputs are derived from sets of underlying equations that are based on fundamental theory. As such, mechanistic model outputs can be interrogated to obtain understanding of processes and their sensitivities."

- Line 104: The entire process and methodology of this study seems to owe a large intellectual debt to ML-based trace metal modeling studies of Roshan et al. These pioneering studies should be acknowledged here, e.g. Roshan et al. (2018), Roshan et al. (2020)

We are happy to acknowledge these earlier studies. The new sentence reads:

"Machine learning is increasingly being used to solve problems in Earth and environmental sciences, including simulating the dissolved distribution of tracers in the sea (e.g., for cadmium, Roshan & DeVries, 2021; copper, Roshan et al., 2020; iodine, Sherwen et al. 2019; nitrogen isotopes of nitrate, Rafter et al., 2019; and zinc, Roshan et al., 2018)."

- Line 196: Explain what you mean by "non-parameteric" and "kernel-based"

Excellent suggestion. We added a sentence to clarify this:

"This particular ML algorithm is non-parametric, kernel-based, and probabilistic, which means that it does not make strong assumptions about the mapping function, can handle nonlinearities, and takes into account the effect of random occurrences when making predictions."

We also added the following to make clear why we used GPR:

"Gaussian Process Regression algorithms are widely used in geostatistics, where it is often referred to as 'kriging' (e.g., Cressie, 1993; Rasmussen & Williams, 2006; Glover et al., 2011). This type of algorithm is ideal when working with continuous data that also contains a certain level of noise, such as from measurement uncertainty or oceanographic variation."

- Line 196: What is the specific MATLAB function, and what options did you specify

An excellent idea. We now name the function in the main text:

"The MATLAB function, `fitrgp`, was used for model training."

We also note the following:

"A full list of the parameter selections used in `fitrgp` is provided in Table S1."

We then provide a table in the Supplement (Table S1) that explains all the function options, what they do, the value we selected, and why we chose that value. This table is reprinted below:

Table S1. Function parameters specified for the function used to train ML models. The MATLAB function `fitrgp` was used to perform model training (MathWorks, 2023). Each option, its purpose, the value assigned, and a justification for the value chosen are shown.

Option	Description of option	Value selected	Description of the value selected
Fit Method	Method to estimate parameters of the GPR model	`'sd'`	Subset of data points approximation (i.e., selects a smaller subset of training data points and computes the inverse of the covariance matrix only for that subset, while the remaining data points are used to estimate the hyperparameters of the model.)
Basis Function	Explicit basis in the GPR model	`'constant'`	H=1 (n-by-1 vector of 1s, where n is the number of observations, i.e., sets the mean of the GPR model to be a constant value, which is equal to the mean of the training output data and is applied to all observations

			in the training data
Beta	Initial value of the coefficients		Inferred from the data, thus changes with each run.
Sigma	Initial value for the noise standard deviation of the Gaussian process model	`std(y)/sqrt(2)`	Depends on the response data, thus changes with each run.
Constant Sigma	Constant value of Sigma for the noise standard deviation of the Gaussian process model	`false`	allows the noise standard deviation to vary across different input points
Sigma Lower Bound	Lower bound on the noise standard deviation	`1e-2*std(y)`	Depends on the response data, thus changes with each run.
Categorical Predictors	Categorical predictors list	logical vector of length p where each element is false and p is the number of predictors	None of our predictors are categorical.
Standardize	Specify whether or not the data should be standardized using mean and standard deviation	`true`	When true, each predictor is centered and scaled to have a mean of zero and a standard deviation of unity.
Kernel Function	Form of the covariance function	`'exponential'`	sets an exponential kernel function (i.e., a type of radial basis function that computes the similarity or covariance between two input vectors based on their distance or proximity in the input space) to be used to model the covariance between the input variables.
Distance Method	Method for computing inter-point distances	`'fast'`	e.g., $(x-y)^2$ is computed as $x^2 + y^2 - 2*x*y$ when the distance method is fast.
Active Set	When specified, the active set indicates the observations to be used in model training. If the active set is predetermined, ActiveSetSize and ActiveSetMethod are not used.	[]	We do not assign a predetermined active set and let the model chose a random active set

Active Set Method	selection method for the Active Set	'random'	random selection of active set
Random Search Size	Random search set size	59	MATLAB default value
Tolerance Active Set	Relative tolerance for terminating active set selection	1e-6	Controls the convergence tolerance level for the active set algorithm used in the "subset of data points" fitting method.
Predict Method	Method used to make predictions	'exact'	Specifies that the exact method should be used to make predictions with the trained GPR model
Optimizer	Optimizer to use for parameter estimation	'quasinewton'	Sets a quasi-Newton method (i.e., a gradient-based optimization algorithm) to estimate the hyperparameters or other parameters of the GPR model.
Initial Step Size	Initial step size	[]	Empty. Initial step size is not used to determine the initial Hessian approximation.
Holdout	A cross-validation method where a fraction of the data is used for validation.	0.2	Use 20% of training data for validation and 80% for training.

- Line 199: Explain the meaning of "basis" and "kernel-function" parameters

These are now all described in Table S1 (above). We think that this change makes the main text simpler to follow and our methods easier to replicate.

- Line 310: The p-values seem to be meaningless. Not sure they add any value here.

This is a good point and we have updated this analysis. Rather than exploring the probability that a feature changes the model we now explore how different features affect the model for the training, testing, and 'good' models. This change is detailed in response to a comment made by Reviewer #1.

The main changes are a new figure (Fig. 3, see response to Reviewer #3) and a new table (Table 3; shown in a response to Reviewer #1).

- Figure 8: Are these values volume-normalized? If not, they would skew toward surface values where grid boxes are smaller.

Yes, these are all volume weighted. This is now noted in the caption:

"Figure 9. Stacked, volume-weighted histograms showing the relative frequency distribution of dissolved [Ba] (A, B) and Ωbarite (C, D) in the global ocean."

- Section 5.1: It makes sense to remove models with lat and lon as predictors. After that, I disagree with all of the choices presented in this section, which ultimately lead to the choice of 1 model out of a possible 1,687 — talk about overfitting!

We agree with parts of this comment and disagree with others. We provide our reasoning in response to the next point.

- Eliminating models with Chl-a and MLD predictors: I will accept eliminating Chl-a, since including it degraded the median model. But just because including MLD only improved the average model by 3% is not a good reason to remove it as a predictor. You have a small sample size in the validation set, and MLD may encode key information for particular environments that are under-represented in the validation set. If it improves the model on average, it is reasonable to keep it.

This last point – *If it improves the model on average, it is reasonable to keep it* – got us thinking about the best way to approach the feature significance analysis, which is summarized in the new Table 3. We performed this analysis for the training data (random holdout cross folding), the testing data (regional cross validation), and for the 1,687 'good' models (also regional cross validation). We consider this last group the most relevant because:

"... these 1,687 models … are superior to existing methods for estimating [Ba] in seawater."

which we now state in Section 5.1. Looking at it this way, we see that six features improved the model on average ($[PO_4]$, $[NO_3]$, T, $[O_2]$, z, [Si]), five degraded it (bathy., Chl. a, MLD, lat., and long.) , and one (S) had no effect. Since there is only one model that contains $[PO_4]$, $[NO_3]$, T, $[O_2]$, z, and [Si], model #3112, we started with that. However, when we plotted the output from model #3112 it became clear that this model, while excellent (statistically speaking), was missing an important aspect of Ba geochemistry: input from rivers. We included some plots illustrating this comparison in the Supplement (see response to Reviewer #1) and note in Section 5.1.:

"Though volumetrically minor, riverine inputs are a geochemically important component of the marine Ba cycle, and the existence of nearshore Ba plumes underpins a major proxy application of Ba. Near-shore riverine influence is easily discerned by low S; we thus explored output from

model #3080, which is identical to model #3112, but includes S as a seventh feature during training. Models #3080 and #3112 exhibit identical statistical performance for the testing data (MAE = 4.3 nmol kg–1; Fig. S1) and make similar predictions for mean marine [Ba] and Ωbarite (89 nmol kg–1 and 0.82, respectively; see Supplement)."

- Eliminating models with Si eliminates the strongest predictor, which seems foolish. There is no reason to eliminate Si just because it appears in the definition of Ba*, which is not even in the target data. If you want the model to predict Ba* in addition to Ba, you could add that to the target when you train the models, but that is still no reason to remove Si from the predictor data (if it were, Si wouldn't even be in the list of features that you consider for this model).

Excellent point. We now retain Si as a feature when winnowing the list of good models.

As a result of this comment, we added [Si] to the model. The new model, #3080, is equivalent to model #3336+[Si], and the performance of the model is improved by about 3 %.

- The reason given for eliminating models with <=4 features is not valid. The analysis shows that *on average* the models with 5-8 predictors performed best (Figure 3). But that doesn't mean that there are not models with <5 predictors that could perform just as well and be just as probable (in fact there clearly are, as shown in Figure 3). It is arbitrary to eliminate these models.

This is a fair point and, given the changes made in response to an earlier comment ("*[i]f it improves the model on average, it is reasonable to keep it*") it no longer applies.

- In general, there is simply no good reason to choose 1 model as the "optimal" model. In fact the great benefit of the model testing that the authors have done is that it affords an ensemble of models from which to choose, many of them being equally or approximately equally probable. It one wants to "weight" the models one could do so be defining a probability function (MAD or something similar would do) and assigning a probability to each of the models. This would be better than simply choosing one single model (equivalent to assigning that model a probability of 1 and all the other models a probability of 0).

This is an interesting point. However, we did not implement this suggestion for two reasons. First, the analysis we performed in response to the reviewer's earlier comment showed that only six features consistently improved ML model performance—[PO_4], [NO_3], T, [O_2], z, [Si]. We thus decided to start with the only model that contained these six features (#3112). Adding S to this model (#3080) had no effect on its MAE or MAPE, but it did mean that the model got rivers right. Second, while an ensemble of good models might be interesting, we believe that end users of this data product may find it easier to simply use our 'best estimate' of marine [Ba]. As more data become available, such as from the next GEOTRACES IDP release, a new best model could well emerge and we can update the model output incorporating those data.

- Line 414: Figure 3 doesn't show sea surface Ba.

Good catch.

This particular cross reference has been cut.

- Line 426: Or maybe the model is just wrong in those regions. Do any other of the possible models (e.g., not model #3336) show elevated Ba at those locations?

Great point.

We now have a section in the Supplement showing a comparison of ML model outputs close to the mouths of major rivers. It appears that including *S* is important if the models are to recognize that there is elevated [Ba] close to shore.

- Line 430: Sure, it's reasonable. It's just unreasonable to say that there are no other possibilities.

We've added another possibility to this section (underlined text):

"The reasons for the lack of elevated [Ba] near the outflow of these two rivers is less clear. It is possible that the model is simply inaccurate in these regions, though we have no particular reason to believe that this is the case. Alternatively, it may reflect seasonal variations in Ba release that are not captured by our mean annual model (e.g., Joung & Shiller, 2014). It could also indicate that these particular rivers are not major net sources of Ba to the surface ocean, which might be the case if dissolved Ba is being retained in the catchment (e.g., Charbonnier et al., 2020) or estuary (e.g., Coffey et al., 1997)."

- Line 551: It would be better to base such uncertainties on an ensemble of most-probable models (rather than a single model)

Based on the revised feature significance analysis suggested by the reviewer, we restricted our analysis to a single model (#3080) and base our uncertainties on the generalization error.

No changes made.

References

- Roshan, S., DeVries, T., Wu, J., & Chen, G. (2018). The internal cycling of zinc in the ocean. Global biogeochemical cycles, 32(12), 1833-1849.

- Roshan, S., DeVries, T., & Wu, J. (2020). Constraining the global ocean Cu cycle with a data‐assimilated diagnostic model. Global Biogeochemical Cycles, 34(11), e2020GB006741.

Citations added.