Supplement of Earth Syst. Sci. Data, 15, 4023–4045, 2023
https://doi.org/10.5194/essd-15-4023-2023-supplement

Supplement of

Barium in seawater: dissolved distribution, relationship to silicon, and barite saturation state determined using machine learning

Öykü Z. Mete et al.

Correspondence to: Öykü Z. Mete (omete@fas.harvard.edu) and Tristan J. Horner (tristan.horner@whoi.edu)

Model initialization and validation

Specific parameters used in model training

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.

Effect of salinity

The feature significance analysis described in the main text indicates that *S* is not, on average, a strong predictor of [Ba]. However, models lacking *S* tend not to reproduce the elevated [Ba] in nearshore environments associated with riverine discharge. Though volumetrically minor, riverine discharge is a geochemically important aspect of the marine Ba cycle, and the existence of nearshore Ba plumes underpins a major proxy application of Ba.

To explore the importance of *S* in predicting [Ba], we compared the output from two models with similar performance whereby the only difference was whether *S* was included as a feature during training. This analysis helps to isolate the effects of *S* on the accuracy of [Ba] predictions. We focused on comparing models #3112 (*z*, *T*, [O_2], [PO_4], [NO_3], and [Si]) and #3080 (*z*, *T*, *S*, [O_2], [PO_4], [NO_3], and [Si]), noting that *S* is the only difference between the two models. An overview of model performance is shown in Fig. S3.

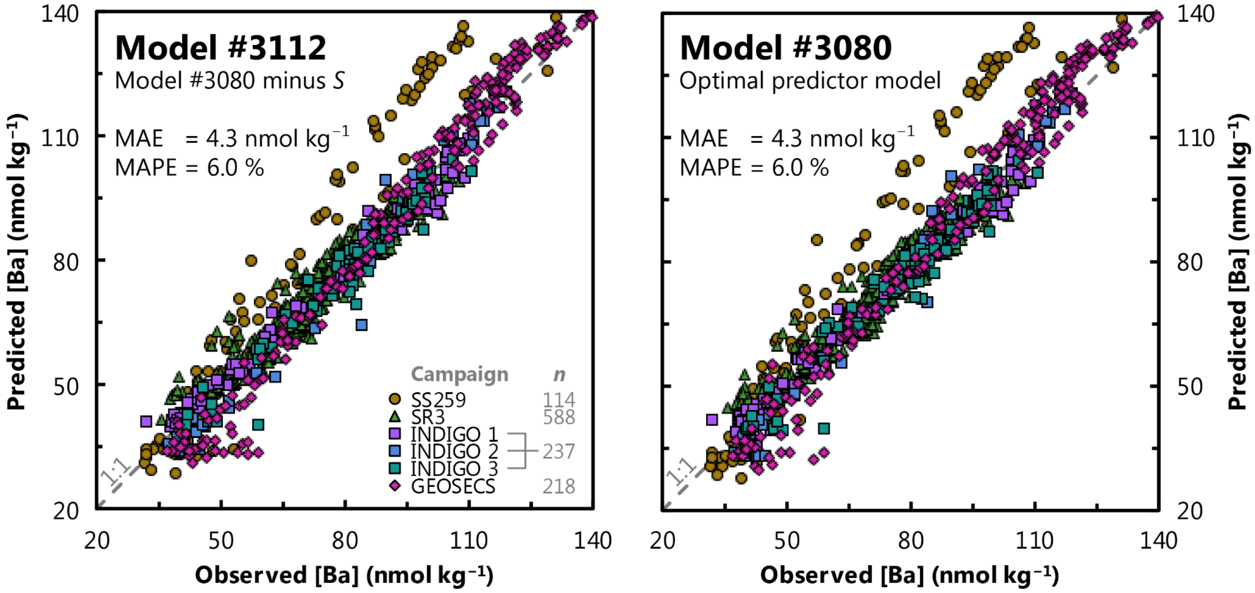

Figure S1. Comparison of model #3112 and model #3080 outputs for the Indian Ocean testing data. Model #3112 was trained using six features: *z*, *T*, [O_2], [PO_4], [NO_3], and [Si]. These are the same features as in model #3080, minus *S.* The statistical performance of the two models is highly similar, though model #3112 misses important geochemical features, discussed in the text.

Statistical comparison of output from models #3112 and #3080 reveals that both are highly adept at predicting [Ba] in the Indian Ocean testing data (Fig. S1). Both models exhibit essentially identical performance and yield similar estimates of mean ocean [Ba] (89 nmol kg^{-1}) and Ω_{barite} (0.82). However, visual inspection of the output reveals stark differences in the model performance that are easily missed by statistical methods. Results of the visual comparison are shown in Figs. S2–S8, below. Within each figure, [Ba] data from model #3080 and #3112 are plotted using the same color scale; however, the scales differ when comparing different geographic regions. Mean annual sea-surface salinity from the WOA 2018 (Zweng et al., 2018) is also shown to assist the reader in identifying regions that are influenced by riverine discharge; note that the salinity scale bars also differ between regions.

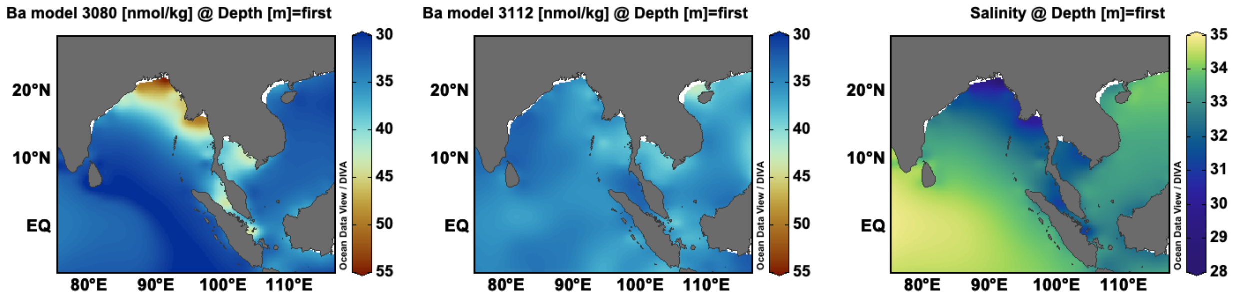
Ba model 3080 [nmol/kg] @ Depth [m]=first
Ba model 3112 [nmol/kg] @ Depth [m]=first
Salinity @ Depth [m]=first
20°N
10°N
EQ
80°E
90°E
100°E
110°E
30
35
40
45
50
55
35
34
33
32
31
30
29
28
Ocean Data View / DIVA

Figure S2. Seawater chemistry in the Bay of Bengal and Andaman Sea. Left and center panels show [Ba] at the sea surface from model #3080 and #3112, respectively. Right panel shows sea-surface salinity.

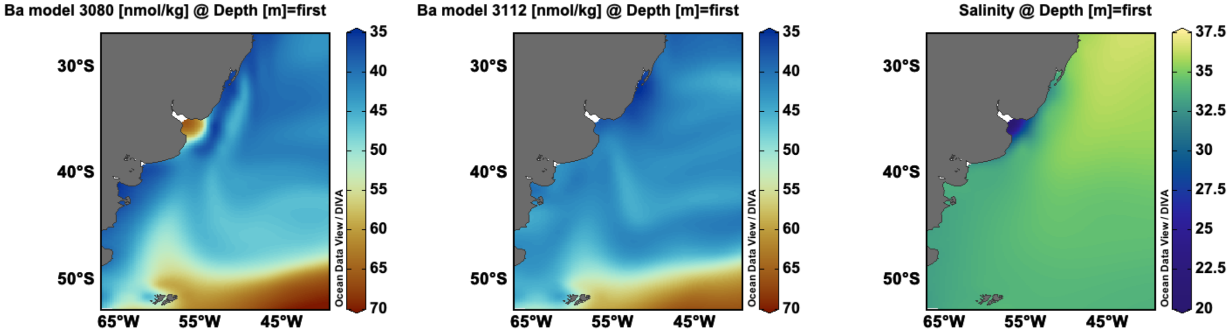
Ba model 3080 [nmol/kg] @ Depth [m]=first
Ba model 3112 [nmol/kg] @ Depth [m]=first
Salinity @ Depth [m]=first
30°S
40°S
50°S
65°W
55°W
45°W
35
40
45
50
55
60
65
70
37.5
35
32.5
30
27.5
25
22.5
20
Ocean Data View / DIVA

Figure S3. Seawater chemistry in the Southwest Atlantic close to the Río de la Plata outflow. Left and center panels show [Ba] at the sea surface from model #3080 and #3112, respectively. Right panel shows sea-surface salinity.

Model #3080 correctly identifies elevated sea-surface [Ba] in the Bay of Bengal and Andaman Seas (Fig. S4), in the Southwest Atlantic close to the Río de la Plata outflow (Fig. S5), and in the East China Sea close to the Yangtze outflow (Fig. S6). Likewise, model #3080 predicts higher-than-background [Ba] in the Northwest Atlantic associated with the St. Lawrence River (Fig. S7) and in the Gulf of Guinea (Fig. S8), though we are not aware of any corroborating data for these latter two regions. In contrast, outputs from model #3112 do not show any nearshore increases in [Ba] associated with river outflow (Figs. S4–10). Since model #3112 did not encounter any information regarding *S* during training, we infer that *S* must be an important feature for predicting near-shore elevated [Ba] associated with river discharge.

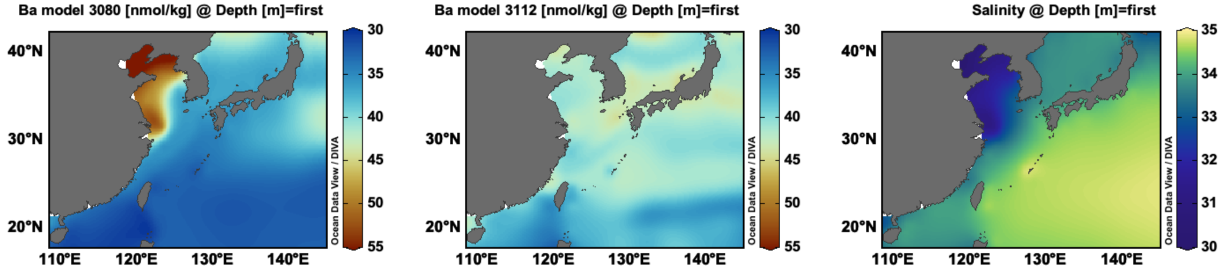
Ba model 3080 [nmol/kg] @ Depth [m]=first
Ba model 3112 [nmol/kg] @ Depth [m]=first
Salinity @ Depth [m]=first
40°N
30°N
20°N
110°E
120°E
130°E
140°E
30
35
40
45
50
55
35
34
33
32
31
30
Ocean Data View / DIVA

Figure S4. Seawater chemistry in the East China Sea. Left and center panels show [Ba] at the sea surface from model #3080 and #3112, respectively. Right panel shows sea-surface salinity.

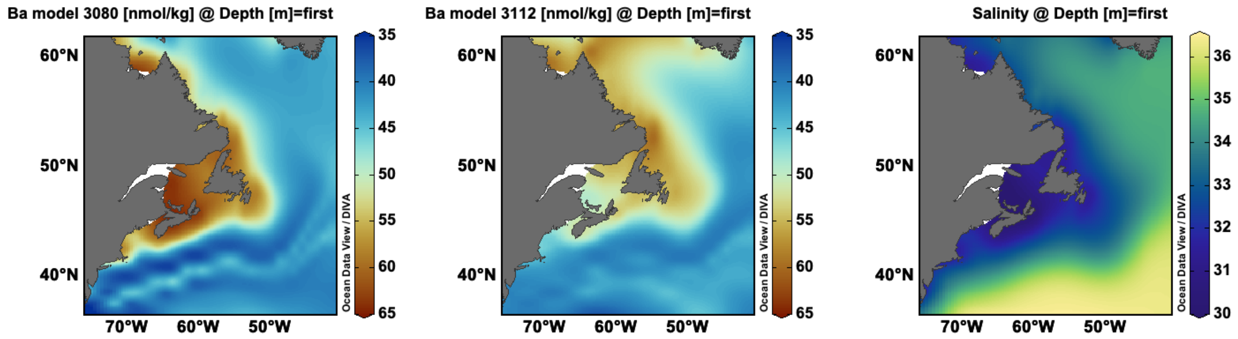
Ba model 3080 [nmol/kg] @ Depth [m]=first
Ba model 3112 [nmol/kg] @ Depth [m]=first
Salinity @ Depth [m]=first
60°N
50°N
40°N
70°W
60°W
50°W
35
40
45
50
55
60
65
36
35
34
33
32
31
30
Ocean Data View / DIVA

Figure S5. Seawater chemistry in the Northwest Atlantic. Left and center panels show [Ba] at the sea surface from model #3080 and #3112, respectively. Right panel shows sea-surface salinity.

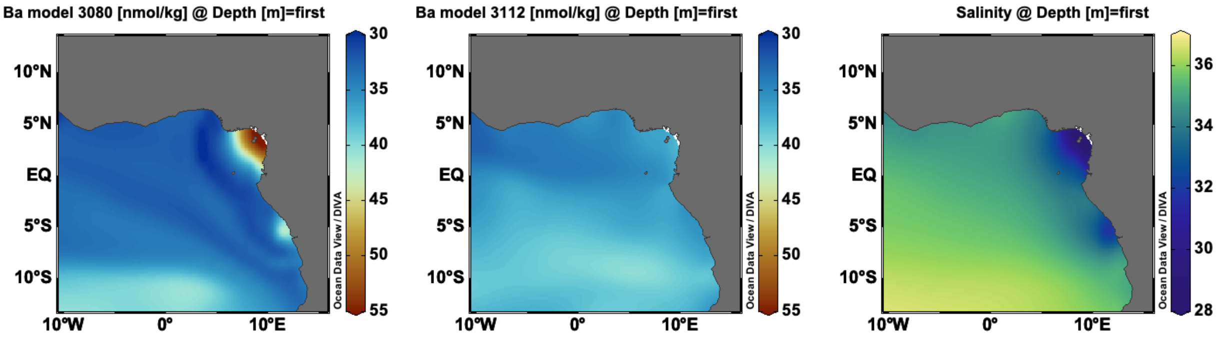
Ba model 3080 [nmol/kg] @ Depth [m]=first
Ba model 3112 [nmol/kg] @ Depth [m]=first
Salinity @ Depth [m]=first
10°N
5°N
EQ
5°S
10°S
10°W
0°
10°E
30
35
40
45
50
55
36
34
32
30
28
Ocean Data View / DIVA

Figure S6. Seawater chemistry in the Gulf of Guinea. Left and center panels show [Ba] at the sea surface from model #3080 and #3112, respectively. Right panel shows sea-surface salinity.

Interestingly, model #3080 does not predict elevated sea-surface [Ba] at the mouths of all major rivers. For example, there is no obvious [Ba] feature associated with the outflows of either the Amazon (Fig. S8) or Mississippi Rivers (Fig. S9). The reason for the lack of a near-shore [Ba] feature in these regions is unclear, and we speculate on some geochemical possibilities in the main text. It is also possible that the lack of a surface [Ba] feature relates to the overall higher salinity in these regions (>33), whereas the largest [Ba] anomalies manifest only when mean annual sea-surface salinity ≤32 (cf. Figs. S4–S8, Figs. S9–10). As with model #3080, model #3112 does not predict surface [Ba] plumes at either the Amazon or Mississippi outflows.

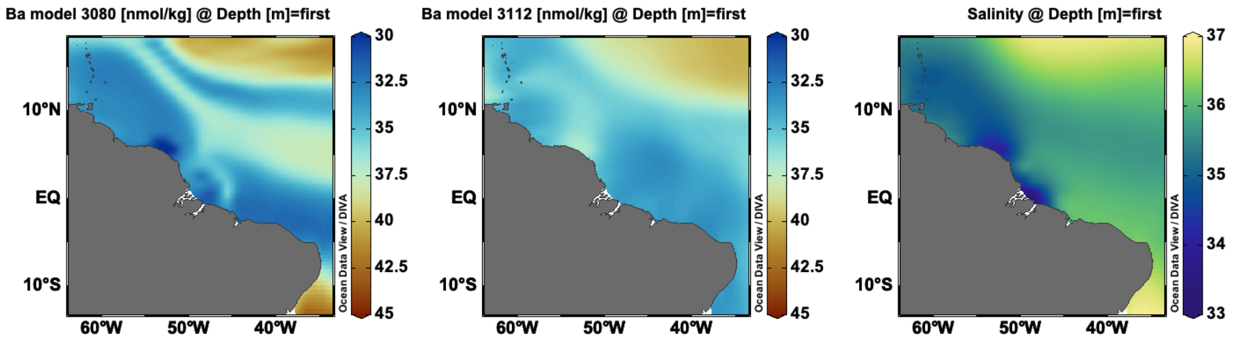
Ba model 3080 [nmol/kg] @ Depth [m]=first
Ba model 3112 [nmol/kg] @ Depth [m]=first
Salinity @ Depth [m]=first
10°N
EQ
10°S
60°W
50°W
40°W
30
32.5
35
37.5
40
42.5
45
37
36
35
34
33
Ocean Data View / DIVA

Figure S7. Seawater chemistry in the Western Tropical Atlantic. Left and center panels show [Ba] at the sea surface from model #3080 and #3112, respectively. Right panel shows sea-surface salinity.

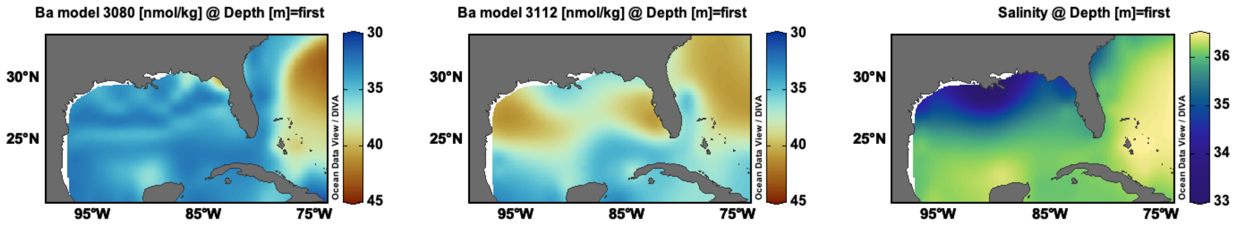
Ba model 3080 [nmol/kg] @ Depth [m]=first
Ba model 3112 [nmol/kg] @ Depth [m]=first
Salinity @ Depth [m]=first
30°N
25°N
95°W
85°W
75°W
30
35
40
45
36
35
34
33
Ocean Data View / DIVA

Figure S8. Seawater chemistry in the Gulf of Mexico. Left and center panels show [Ba] at the sea surface from model #3080 and #3112, respectively. Right panel shows sea-surface salinity.

Overall, this analysis highlights that *S* is an important predictor of [Ba] in certain coastal environments. The importance of *S* is only revealed by visual inspection of model output, which is an important component of data analysis (see e.g., Anscombe, 1973). Since the statistical performance of model #3080 is identical, within uncertainty, to model #3112 (Fig. S1), and model #3080 reproduces known riverine [Ba] features, we select model #3080 as our preferred model for making global predictions of [Ba].

Reducing the range of the training data

Gaussian Process Regression (GPR) models are highly adept at making accurate geospatial predictions of a target variable, particularly when the training data contain a certain level of noise. However, GPR models are oftentimes less accurate than other methods when making predictions beyond the ranges encountered during training (Cressie, 1993). To investigate whether our preferred predictor model was subject to similar biases, we analyzed the performance of model #3080 when provided with a narrower range of training data. This meant restricting the range of

[Ba] values seen during model training and comparing these outputs against those generated by model #3080 when trained on the full training dataset.

To achieve this, we identified the bottom (51.4) and top (98.9 nmol kg^{-1}) sextile in the Indian Ocean testing data and removed all [Ba] observations from the training data that were outside of this range (i.e., only samples with [Ba] between 51.4–98.9 nmol kg^{-1} were included). This reduced the number of [Ba] observations in the training data from 4,345 to 2,295. We then retrained model #3080 (z, T, S, [O_2], [PO_4], [NO_3], and [Si]) on these 2,295 data and used this retrained model (hereafter model #3080N) to predict [Ba] for the Indian Ocean and on a global basis. A comparison of [Ba] predictions made using model #3080N and from model #3080 are shown below in Figs. S9 and S10.

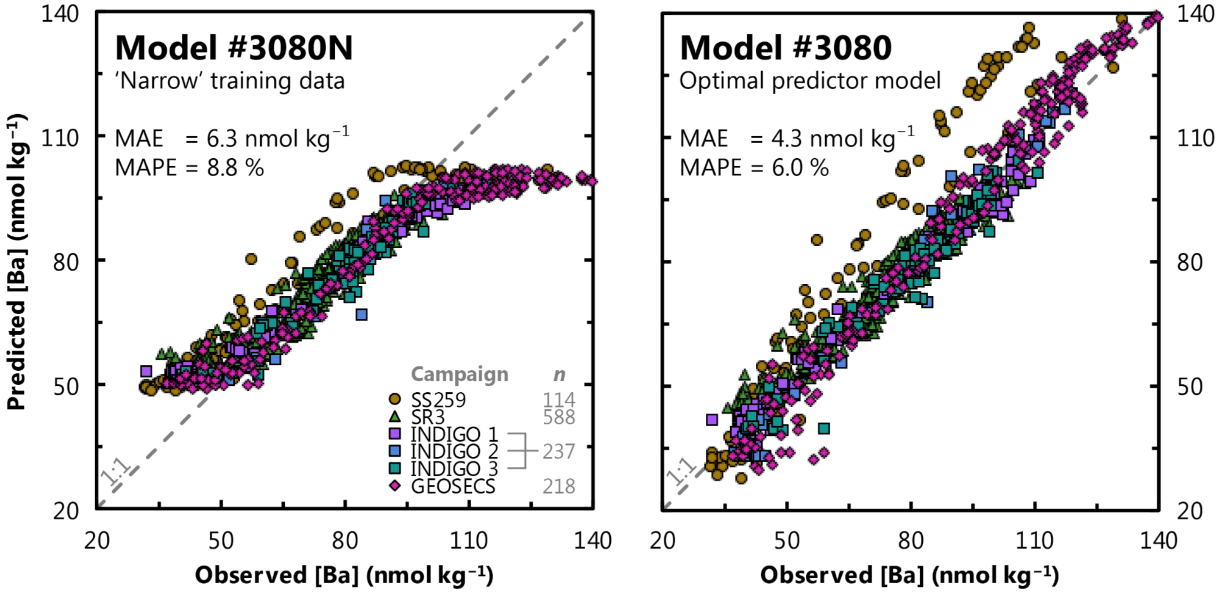

Figure S9. Comparison of model #3080N and model #3080 outputs for the Indian Ocean testing data. Model #3080 was trained using a narrowed version of the training data compared to model #3080, which saw the full training database.

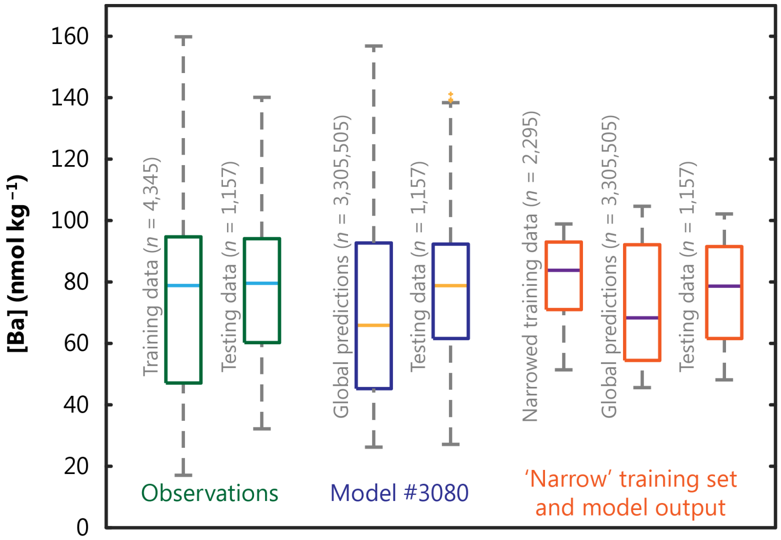

Figure S10. Boxplot of [Ba] values for training, testing, and ML model predictions. Each box shows a five-number summary for the relevant dataset: median (horizontal line), the 75th and 25th percentiles (top and bottom of box, respectively), and the maximum and minimum non-outlier values (upper and lower whiskers, respectively). Statistical outliers are indicated by '+'. Boxes labeled 'observations' summarize the *in situ* data used in model training and testing, respectively. The next two boxes show model #3080 predictions for the global ocean and the testing data. The final three boxes show the distribution of [Ba] values in the 'narrow' training data and the resultant spread of [Ba] predicted by model #3080N, which was trained using only these data.

This analysis shows that model #3080N reproduces the median and interquartile range of the Indian Ocean testing data (Figs. S9) as well as for the global predictions (Fig. S10). Likewise, model #3080N can predict values of [Ba] outside of the ranges encountered in model training, but only by between 5–10 %. As such, model #3080N underestimates the true range of [Ba] values seen in the ocean and achieves a lower overall accuracy of [Ba] predictions compared to model #3080 (#3080N MAPE = 8.8 % vs 6.0 %; Fig. S9). We conclude that the output from model #3080, and likely other models, is most accurate when it falls within the range of [Ba] encountered during training. Since model #3080 was trained on [Ba] data spanning 17.1–159.8 nmol kg^{-1}, the entire range of model #3080 predictions (26.2–156.8 nmol kg^{-1}; Fig. S10) falls within the range seen during training. Thus, we conclude that the results from model #3080 are generally robust as the model did not extrapolate beyond the range of [Ba] encountered during training.

Supplementary References

Anscombe, F. J. (1973). Graphs in Statistical Analysis. *The American Statistician*, 27(1), 17-21. https://doi.org/10.1080/00031305.1973.10478966.

Cressie, N.A.C. (1993). Spatial Prediction and Kriging. In *Statistics for Spatial Data*, N.A.C. Cressie (Ed.). https://doi.org/10.1002/9781119115151.ch3

MathWorks (2023). `Fitrgp` Documentation. *Natick, Massachusetts: The MathWorks Inc.* https://www.mathworks.com/help/stats/fitrgp.html

Zweng, M.M, J.R. Reagan, D. Seidov, T.P. Boyer, R.A. Locarnini, H.E. Garcia, A.V. Mishonov, O.K. Baranova, K.W. Weathers, C.R. Paver, and I.V. Smolyar (2018). World Ocean Atlas 2018, Volume 2: Salinity. A. Mishonov, Technical Editor, NOAA Atlas NESDIS 82, 50pp. http://www.nodc.noaa.gov/OC5/indprod.html