A spatially and vertically resolved global grid of dissolved barium concentrations in seawater determined using Gaussian Process Regression machine learning

Website: https://www.bco-dmo.org/dataset/885506
Data Type: model results
Version: 2
Version Date: 2023-07-11

Project
» The Speed, Signature, and Significance of Barium Transformations in Seawater (The Three S's)

Contributors	Affiliation	Role
Horner, Tristan J.	Woods Hole Oceanographic Institution (WHOI)	Principal Investigator
Mete, Oyku Z.	Woods Hole Oceanographic Institution (WHOI)	Student
Rauch, Shannon	Woods Hole Oceanographic Institution (WHOI BCO-DMO)	BCO-DMO Data Manager

Abstract
We present a spatially and vertically resolved global grid of dissolved barium concentrations ([Ba]) in seawater determined using Gaussian Process Regression machine learning. This model was trained using 4,345 quality-controlled GEOTRACES data from the Arctic, Atlantic, Pacific, and Southern Oceans. Model output was validated by assessing the accuracy of [Ba] simulations in the Indian Ocean, noting that none of the Indian Ocean data were seen by the model during training. We identify a model that can accurate predict [Ba] in the Indian Ocean using seven features: depth, temperature, salinity, as well as dissolved dioxygen, phosphate, nitrate, and silicate concentrations. This model achieves a mean absolute percentage error of 6.0 %, which we assume represents the generalization error. This model was used to simulate [Ba] on a global basis using predictor data from the World Ocean Atlas 2018. The global model of [Ba] is on a 1°x 1° grid with 102 depth levels from 0 to 5,500 m. The dissolved [Ba] output was then used to simulate dissolved Ba* (barium-star), which is the difference between 'observed' and [Ba] predicted from co-located [Si]. Lastly, [Ba] data were combined with temperature, salinity, and pressure data from the World Ocean Atlas to calculate the saturation state of seawater with respect to barite. The model reveals that the volume-weighted mean oceanic [Ba] and and saturation state are 89 nmol/kg and 0.82, respectively. These results imply that the total marine Ba inventory is 122(±7) $\times 10^{12}$ mol and that the ocean below 1,000 m is at barite equilibrium.

Table of Contents

Coverage

Spatial Extent: **N**:89.5 **E**:179.5 **S**:-77.5 **W**:-179.5
Temporal Extent: 2007 - 2018

Methods & Sampling

The data are output from a machine learning model that was trained using GEOTRACES dissolved Barium ([Ba]) data. Full protocols for sample collection and analysis are provided in the GEOTRACES Cookbook and 2021

Intermediate Data Product (see References), respectively.

Full methods are provided in a companion study, which is in revision for E*arth System Science Data* (Mete et al., 2023). A summary of methods is provided below.

The features used to predict [Ba] and their associated data sources are summarized in Table 1 of Mete et al. (2023). The first three features (latitude, longitude, depth) record geospatial information that defines the location of an observation in three-dimensional space. Features 4–9 encode physical (temperature, salinity) and chemical (oxygen, nutrients) information that is routinely measured alongside [Ba]. These data were generally available for the same bottle as the [Ba] measurements; however, when that was not the case, nutrient data were taken from the corresponding location during a separate cast, or, in the case of oxygen, from linearly interpolated sensor data. Features 10-12 are independent of depth, meaning that all samples within a given vertical profile exhibit the same value for mixed-layer depth, sea-surface chlorophyll a, and bathymetry.

Table 2 of Mete et al. (2023) identifies all dataset sources of d[Ba] ingested into the master record. The data ingestion process resulted in a master record containing 5,502 observations of [Ba] that also contained a corresponding value for all 12 of the features of interest described above. The record was then split into a Pareto partition: the first partition was used for ML model training (4,345 observations, 79 % of data) and the second for model testing (1,157 data; 21 %).

We opted for supervised ML using a Gaussian Process Regression learner, implemented in MATLAB. The training partition of the master record was used to train 4,095 different machine learning models with the goal of finding a model that could accurately simulate the global distribution of [Ba]. Each model uses a unique combination of the 12 features and our testing followed a factorial design whereby each feature was either enabled or disabled. In the second stage of cross validation, trained models were used to predict [Ba] for the withheld data from the Indian Ocean. The accuracy of the models was assessed by comparing ML model predictions against observed [Ba]. We then winnowed the list of models from 4,095 to a single, highly accurate model (#3080), which we used to simulate Ba* and the saturation state of seawater with respect to barite on a global basis.

Refer to Mete et al. (2023) for complete methodology, results, and discussion.

The data provided here include the resulting global grid of dissolved [Ba], Ba*, and barite saturation state as well as Supplemental Files used in testing and training of the model.

The code used in running the model is also provided here in the Supplemental File "Model_3080_code.zip". "predictBa.m" is a code that allows users to predict [Ba] in seawater based on input data for seven predictors: depth, temperature, salinity, dioxygen, phosphate, nitrate, and silicate. Predictions of [Ba] are made using "trainedModel_Exp3080.mat", which is a Gaussian Process Regression Machine Learning Model that was trained to simulate [Ba] based on these seven inputs. Instructions on how to use the model are provided in the comments to predictBa.m and example input data are provided in "exampleData.xlsx". The code was written in MATLAB, and should work on all versions beyond 2018a. All settings, configurations, and the training process are described in a companion study by Mete et al. (2023).

Data Processing Description

Data Processing:
Model training and testing were performed in MATLAB.

BCO-DMO Processing
Version 1:
- in .csv file "horner_and_mete_global_ba_grid.csv", replaced the unprintable omega symbol with the text "omega"; removed the empty date/time column; removed Cruise and Type columns (unnecessary for data re-use); and renamed columns to comply with BCO-DMO naming conventions for text files.
- in ODV file "Horner_and_Mete_Ba_grid_ODV_BCO.txt", replaced the unprintable omega symbol with the text "omega".

Version 2:
- extracted original data file "ML_model_3080.txt" from .zip file;
- replaced tab separators with commas in the "ML_model_3080.txt" file and imported into the BCO-DMO data system;
- removed the empty date/time column and the Cruise and Type columns (unnecessary for data re-use);

- renamed columns to comply with BCO-DMO naming conventions for text files;
- named the final data file "885506_v2_global_ba_grid.csv";
- replaced version 1 of the following Supplemental Files with new/updated versions: "Experiment_list.csv", "Global_and_basin_averages.csv".
- added the "Model_3080_code.zip" Supplemental File.

[table of contents | back to top]

Data Files

File
885506_v2_global_ba_grid.csv (Comma Separated Values (.csv), 175.47 MB) MD5:bdda9a9964dd664d2802a60a9bbfda05 Primary data file for dataset ID 885506, version 2. See the "Parameters" section of the metadata for column descriptions and units. Missing values are reported as NaN.

[table of contents | back to top]

Supplemental Files

File
ML_model_3080_global_and_basin_averages filename: Global_and_basin_averages.csv (Comma Separated Values (.csv), 34.65 KB) MD5:199c056747157ab66edde53f3d19d06e Average depth profiles of [Ba] and barite saturation state. (Supplemental file for dataset ID 885506, version 2.) Depth profiles of mean, median, and the standard deviation of [Ba] (nmol/kg) and barite saturation state (unitless) for the whole ocean and the major ocean basins (Arctic, Atlantic, Indian, Pacific, and Southern Oceans). All profiles are provided on the World Ocean Atlas 2018 depth spacing and the number of profiles in each bin is shown.
ML_model_3080_global_grid_ODV filename: ML_model_3080.txt (Plain Text, 285.35 MB) MD5:18c9963f020f27dbaff51c0b429c475b An alternative version of the primary data file for dataset ID 885506, version 2. Provided to facilitate use with ODV. This file was exported from ODV and contains metadata at the top of the file before the data begins. The data themselves are identical to file "885506_v2_global_ba_grid.csv" with the exception of a few additional columns which have been removed from the .csv as they are not needed (Cruise, Type, and date/time). Missing values are reported as NaN.
ML_model_experiment_list filename: Experiment_list.csv (Comma Separated Values (.csv), 208.65 KB) MD5:4bcde4a4169cc1517950f50d33d70750 List of trained Gaussian Process Regression models and their respective skill metrics. (Supplemental file for dataset ID 885506, version 2.) List showing the 4,095 Machine Learning models trained and tested in this study. Each model uses a unique combination of the 12 features tested—longitude, latitude, bathymetry, depth, temperature, salinity, oxygen, phosphate, nitrate, silicate, mixed-layer depth, and chlorophyll a. If a feature was used in model testing it is denoted by a ‘1’, else it is ‘0.’ Column 1 lists the model number; model 3080 (the model used for global simulations) is shown first, model 3112 second, and the others are listed in a random order. Columns 2-13 show the features included in that model. The final four columns show the model skill in terms of Mean Absolute Error (MAE; units nmol/kg) and Mean Absolute Percentage Error (%) for the training data and then for the testing data.

File

ML_model_testing_data

filename: Testing_data.csv

(Comma Separated Values (.csv), 152.11 KB)
MD5:7fff49de2d2f8af0e40db6f78e430d1a

Data used to test GPR ML models. (Supplemental file for dataset ID 885506, version 1 and version 2.)

File containing the Indian Ocean data used to test the ML models. The file contains 16 columns. Columns 1–4 indicate the Cruise ID, the station, and coordinates from which the in situ barium data were obtained. Column 5 is our best-estimate of the bathymetry. Column 6 is the depth in the water column from which the same was collected. Columns 7–12 contain interpolated World Ocean Atlas data for physical (temperature, salinity) and biogeochemical parameters (nutrients, oxygen). Columns 13 and 14 show the interpolated mixed-layer depth and sea-surface chlorophyll a used in model test. Column 15 contains the observed in situ dissolved barium from the respective data source, listed in column 16.

ML_model_testing_results

filename: Testing_results.csv

(Comma Separated Values (.csv), 32.18 MB)
MD5:768e112fb6b64ef30c08475552e46321

Model testing results. (Supplemental file for dataset ID 885506, version 1 and version 2.)

File containing the results from ML model testing in the Indian Ocean. Columns 1–5 provide a unique identifier that corresponds to the samples listed in 'Testing_data.csv.' Column 6 shows observed [Ba]. Columns 7–4,102 show model-predicted [Ba] (in nmol/kg) for each trained model shown in 'Experiment_list.csv.'

ML_model_training_data

filename: Training_data.csv

(Comma Separated Values (.csv), 641.82 KB)
MD5:ca8d9a144f39b564884cd83474a59ded

Data used to train GPR ML models. (Supplemental file for dataset ID 885506, version 1 and version 2.)

File containing the data used to train the ML models. The file contains 16 columns. Columns 1–4 indicate the GEOTRACES cruise ID, the station, and coordinates from which the in situ data were obtained. Column 5 is our best-estimate of the bathymetry. Column 6 is the depth in the water column from which the same was collected. Columns 7–12 contain in situ data for physical (temperature, salinity) and biogeochemical parameters (nutrients, oxygen). Columns 13 and 14 show the interpolated mixed-layer depth and sea-surface chlorophyll a used in model training. Column 15 contains the observed in situ dissolved barium. Column 16 shows a citation to the data source and/or originator (if unpublished).

Model_3080_code.zip

(ZIP Archive (ZIP), 140.06 MB)
MD5:237647d1bbd10c4c04d29b9c679bd654

This folder contains the following 3 files:

(1) predictBa.m;

(2) trainedModel_Exp3080.mat;

(3) exampleData.xlsx.

"predictBa.m" is a code that allows users to predict [Ba] in seawater based on input data for seven predictors: depth, temperature, salinity, dioxygen, phosphate, nitrate, and silicate. Predictions of [Ba] are made using "trainedModel_Exp3080.mat", which is a Gaussian Process Regression Machine Learning Model that was trained to simulate [Ba] based on these seven inputs. Instructions on how to use the model are provided in the comments to predictBa.m and example input data are provided in "exampleData.xlsx".

The code was written in MATLAB, and should work on all versions beyond 2018a. All settings, configurations, and the training process are described in a companion study by Mete et al. (2023).

[table of contents | back to top]

Related Publications

Cutter, Gregory, Casciotti, Karen, Croot, Peter, Geibert, Walter, Heimbürger, Lars-Eric, Lohan, Maeve, Planquette, Hélène, van de Flierdt, Tina (2017) Sampling and Sample-handling Protocols for GEOTRACES Cruises. Version 3, August 2017. Toulouse, France, GEOTRACES International Project Office, 139pp. & Appendices. DOI: http://dx.doi.org/10.25607/OBP-2
Methods

GEOTRACES Intermediate Data Product Group. (2021). The GEOTRACES Intermediate Data Product 2021 (IDP2021). (Version 1) [Data set]. NERC EDS British Oceanographic Data Centre NOC. https://doi.org/10.5285/CF2D9BA9-D51D-3B7C-E053-8486ABC0F5FD
References

Mete, Ö. Z., Subhas, A. V., Kim, H. H., Dunlea, A. G., Whitmore, L. M., Shiller, A. M., Gilbert, M., Leavitt, W. D., & Horner, T. J. (2023). Barium in seawater: dissolved distribution, relationship to silicon, and barite saturation state determined using machine learning. Earth System Science Data, 15(9), 4023–4045. https://doi.org/10.5194/essd-15-4023-2023
Results

[table of contents | back to top]

Parameters

Parameter	Description	Units
Station	World Ocean Atlas station number	unitless
Longitude_degreesE	Longitude	degrees East
Latitude_degreesN	Latitude	degrees North
Depth_m	Depth	meters (m)
dBa_nmol_kg	Dissolved barium concentration	nanomoles per kilogram (nmol/kg)
omega_Ba	Saturation state with respect to barite	unitless
Ba_star_nmol_kg	Barium-star, which is the difference between 'observed' and expected dissolved barium concentrations	nanomoles per kilogram (nmol/kg)

[table of contents | back to top]

Project Information

The Speed, Signature, and Significance of Barium Transformations in Seawater (The Three S's)

Coverage: Eastern Tropical Pacific, Subtropical South Pacific, Sub-Antarctic Pacific, and Southern Oceans

NSF Award Abstract:
The biological cycling of carbon in the oceans entrains many other elements, some directly (like nutrients that are essential for life) and some indirectly, as they become chemically involved in the processes that are affecting carbon. One such element is barium (Ba). Particles of the mineral barite (barium sulfate) have been found to form in association with microbial consumption of organic material in the ocean's "twilight zone." These particles settle to the ocean floor, and their presence in sediments has been used to infer changes in the conditions in the ocean back in time. Both the amount of barite in sediments and the isotope composition of Ba in barite are potentially sensitive to processes occurring in the twilight zone. However, several long-standing questions remain about Ba cycling in the oceans, which complicates the interpretation of barium-based proxy records. Examples of remaining questions include how much barium enters the oceans at mid-ocean ridge

hydrothermal sites, and what controls the precipitation and dissolution of barite in the water column. This project seeks to tackle these questions using new approaches, on three scheduled research expeditions in the Pacific and Southern Oceans. In doing so, this project will support the education, training, and career development of a graduate student, postdoctoral researcher, and junior investigator. Undergraduate students from underrepresented groups will be recruited to conduct complementary shore-based experiments.

This proposal seeks to answer four questions central to the utility of barium-based proxies in oceanography: What are the major inputs of new Ba to the ocean? What are their isotopic compositions? What controls the amount of pelagic barite precipitated during the remineralization of organic matter? What influences its isotopic composition? These questions will be addressed using a field-centric approach combining: in situ and shipboard tracer-incubation experiments, AUV-led adaptive sampling of Ba cycling 'hotpots', and section-based surveying of the surrounding oceanographic features. This multi-pronged approach will be used to investigate: the flux and isotopic composition of Ba released from the largest hydrothermal fields in the ocean, the Southern East Pacific Rise, with a focus on low-temperature venting; rates and signatures of pelagic barite precipitation associated with different phytoplankton assemblages in the Southern Ocean; and, the importance of environmental conditions, such as low ambient oxygen concentrations, in setting the efficiency of barite precipitation in the Eastern Tropical Pacific. The significance of each transformation will be assessed, which may lead to ruling out the importance of certain processes, or identifying new dependencies that could form the basis of new proxies.

This award reflects NSF's statutory mission and has been deemed worthy of support through evaluation using the Foundation's intellectual merit and broader impacts review criteria.

[table of contents | back to top]

Funding

Funding Source	Award
NSF Division of Ocean Sciences (NSF OCE)	OCE-2023456
NSF Division of Ocean Sciences (NSF OCE)	OCE-2048604

[table of contents | back to top]