RC3: 'Comment on essd-2023-67’

Frank Pavia

I really enjoyed reading this paper by Mete et al. The manuscript was well-written, well-organized, extremely clear, and the work generates a product that should be used by chemical oceanographers and paleoceanographers alike. I have a few comments, and only the first is substantial.

We are delighted to read such an encouraging and positive response from Frank Pavia. We found these comments very helpful.

The decision to only use Indian Ocean data as the validation dataset is definitely curious, and I think not justified well-enough in the text. The authors cite Rafter et al. 2019 as their source for doing location-based separation of training and test data to avoid overfitting. Rafter et al., however, don't isolate a single basin for this - they use whole ship transects as their witheld data, and these transects span multiple basins, hemispheres, and latitudes. Testing a globally-trained dataset on a regionally-confined subset of data doesn't, at least to a reader not well-versed in these sorts of choices, inspire the maximum amount of confidence in the results of the global output of the model. Perhaps the authors could more completely explain this choice to bulwark against this criticism.

This is a great point and we’ve taken the opportunity to clarify our reasoning. Before addressing it, however, we want to note a related point raised by Reviewer #1, who wanted us to include the randomly assigned training–testing separation results for comparison. We now do so in Figure 3 (below), which shows that the model training process vastly underestimates the error in [Ba] predictions. The main reason for this, stated on p. 108 of Rasmussen & Williams (2006) is:

> *“[T]he training error is usually a poor proxy for the generalization error, since the model may fit the noise in the training set (over-fit), leading to low training error but poor generalization performance.”*

This means that if we partition the data randomly we will get small, but unrealistic errors. If we partition geographically we get larger, but more-realistic errors. We now note this in Section 3.2.:

“A significant problem in supervised ML, and particularly Gaussian Process Regression learning, is overfitting: models may fit the noise in the training data, leading to poor generalization performance (Rasmussen & Williams, 2006). Since our goal was to develop a global model of [Ba] using regional training data, we deemed it especially important to identify generalizable models. Generalizable models were identified through a testing process involving regional cross-validation; each trained model was used to predict [Ba] for the 1,157 samples from the Indian Ocean and model predictions were again compared against observations. Importantly, no [Ba] data from the Indian Ocean were seen by any of the models during training. This process helped to identify models that may have been overfit to the training data and can further be used to calculate generalization errors (Sect. 4.1).”

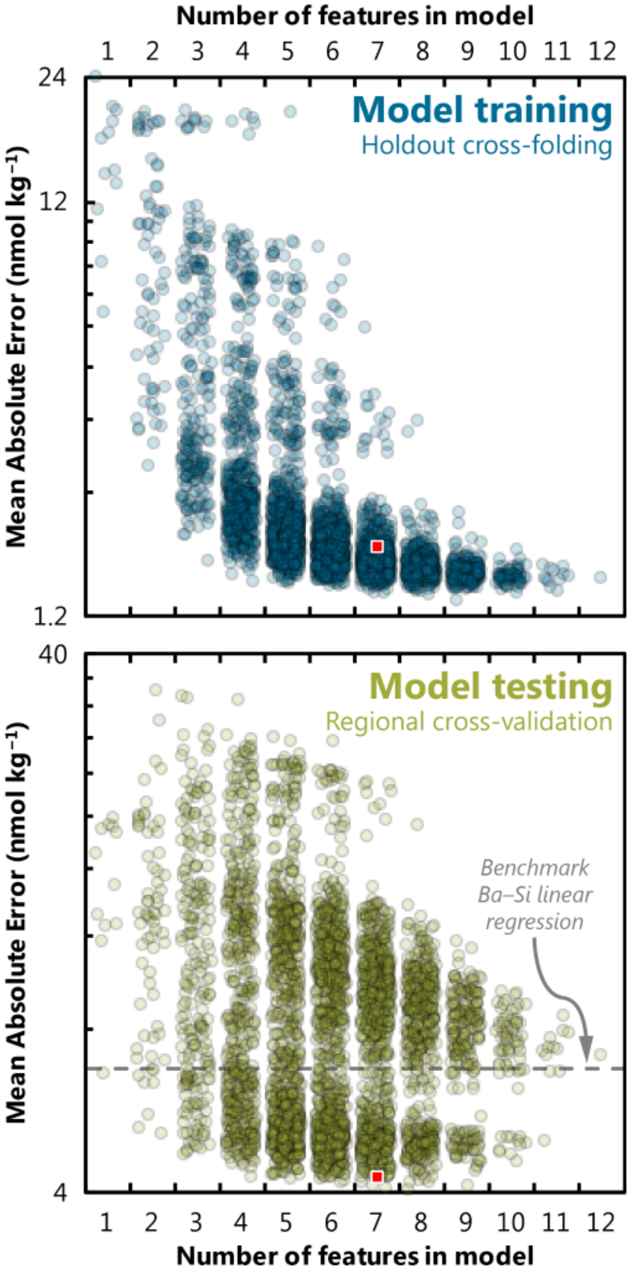

Figure 3. Effect of feature addition on ML model accuracy. Accuracy was quantified for each of the 4,095 trained models and quantified here using MAE (note log scale, which differs between panels). The accuracy of trained models is shown for random holdout cross-validation during training (top) and for regional cross-validation during testing (bottom). Square indicates the performance of our favored predictor model, #3080 (see Fig. 4, Sect. 5.1). The accuracy of the Ba–Si linear regression benchmark is shown as a dashed line in the lower panel (MAE = 6.8 nmol kg^{-1}). To illustrate data density, points have been randomly positioned within their respective bin and plotted with 80 % transparency.

To further address Reviewer #3's point, we added some discussion to Section 5.1. explaining why we opted for regional cross-validation. We think that these arguments are now much stronger than in the original submission and we appreciate the reviewer's insights.

"Choosing a single, optimal model configuration is challenging given the sheer number of skillful ML models. Below we winnow the list from 4,095 to a single model (#3080). We base our winnowing primarily on the results of the regional cross-validation performed in the Indian Ocean, rather than from the errors determined from random holdout cross folding of the training data. We believe that there are three strong reasons for winnowing in this way. First, Gaussian Process Regression Learners tend to fit the noise in the training data, meaning that the training error is significantly lower than the generalization error (Rasmussen & Williams, 2006). Indeed, trained models showed overall lower performance during testing compared to training, which we believe is evidence of overfitting (Fig. 3, Table 3). Second, a generalizable global model should be able to make predictions in regions where it has not already learned anything about the target variable. Our regional cross-validation approach satisfies this consideration since no Indian Ocean data were seen during model training. Third, the Indian Ocean is an ideal basin for testing as it exhibits the full diversity of features expected to influence [Ba] (riverine inputs, oxygen-minimum zones, coastal upwelling, etc.) and constitutes ≈20 % of the global ocean volume. Likewise, the Indian Ocean captures most of the range in [Ba] seen elsewhere in the ocean (Fig. 9); this likely reflects the input of Atlantic waters through the Aughulas leakage, transport of old Pacific waters via the Indonesian Throughflow, and northward spreading of mode and intermediate waters from the Southern Ocean. We thus assume that the Indian Ocean testing errors are a good approximation of the generalization error, which we now use to winnow the list of models."

Figures 4-7: The 0.75 value for barite saturation is labeled as 1.75 in all these figures.

Good catch (thanks also to Reviewer #2!).

Changed.

Line 342: Change Look to Looking

Done. Now reads:

"Looking at the ocean as a whole, the probability density function of [Ba] roughly resembles a uniform distribution, with a mean ocean [Ba] of 89 nmol kg–1 (Fig. 9A)."

Figure 8: I think it would be helpful to have a key/legend for the basins in A and C, similar to the key for depths in B and D. It is not easy to match the text color of the basins to the colors of the histograms in panels A and C.

Good suggestion. These have been added to the updated histograms for model #3080:

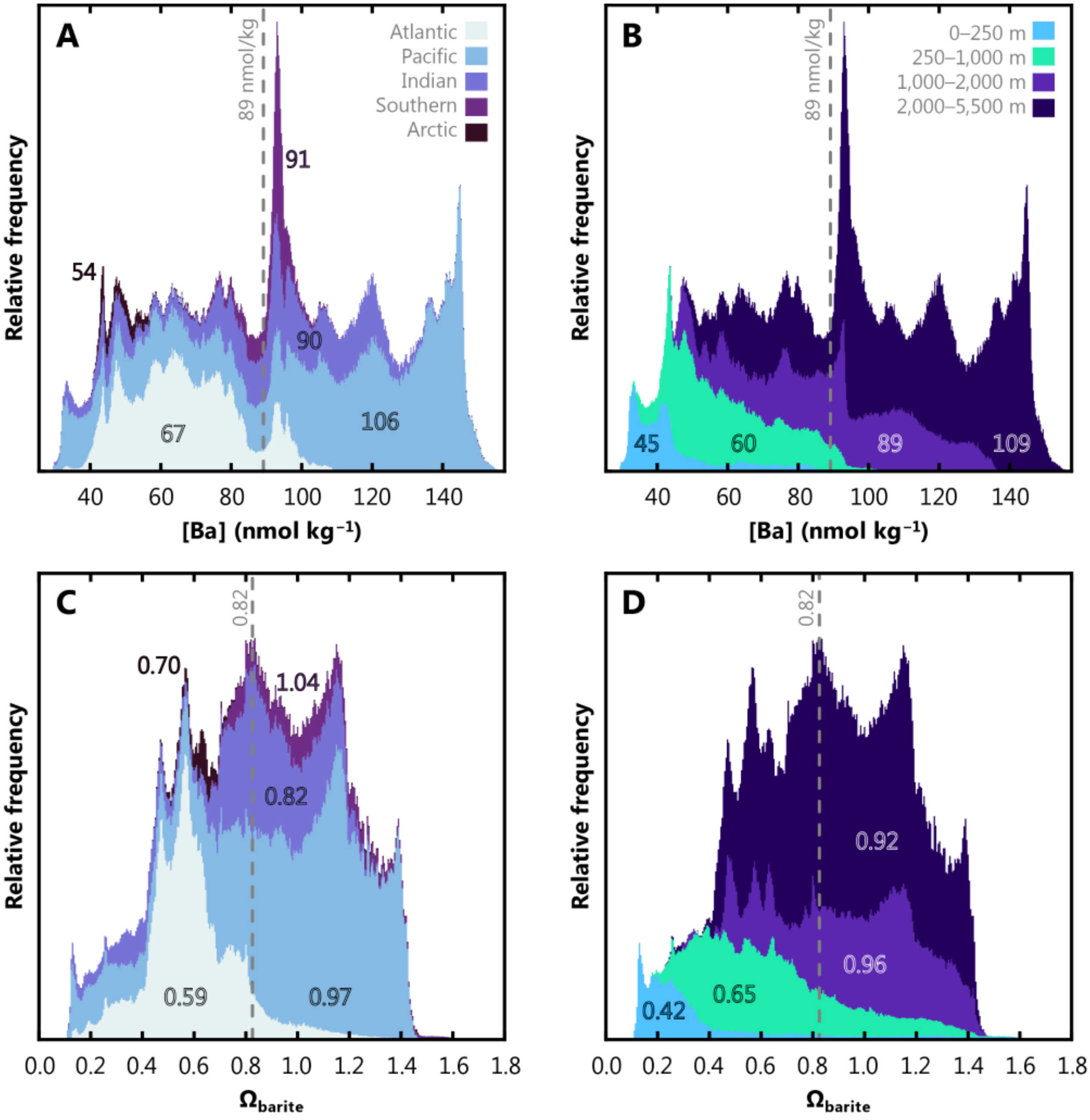

Figure 9. Stacked, volume-weighted histograms showing the relative frequency distribution of dissolved [Ba] (A, B) and Ω_{barite} (C, D) in the global ocean. The left column shows data grouped by basin and the right column shows data grouped by a prescribed depth bin. Numbers in each panel display the mean property value for that bin. Dashed line shows the global mean.

Lines 507-513: Did all or many of the best models tend to produce the same systematic mismatch between predicted and measured Ba for the Singh et al. 2013 data? It would be helpful to know to make sure it isn't a quick [quirk?] of this specific model.

Reviewer #1 had a similar question and we added a few sentences in Section 5.2.3. to clarify this point (reprinted in response to Reviewer #1's comment).

Overall, we don't think it's a specific quirk of this model because several other good models reproduced this offset; indeed, our new 'optimal' predictor model shows similar behavior as model #3336 (the optimal model in the original submission). However, we did not investigate whether every model predicted this mismatch because our focus was on making a globally (rather than regionally) accurate model of [Ba]. It's worth noting that model #3080 does well on the shallow samples from Singh et al. (2013), implying that it's something specific to the deep Bay of Bengal, namely: changes in the deep dissolved Ba inventory since GEOSECS, or inaccuracies with the measurements. Since we cannot distinguish between these two possibilities without more recent samples from the region, we have left it open ended.

Section 5.3. I really enjoyed reading this section and am looking forward to the community's use of this data product.

Excellent!

No changes made.