Supplementary Information for

Mapping the global distribution of lead and its isotopes in seawater with explainable machine learning

Arianna Olivelli[1,2,*], Rossella Arcucci[1], Mark Rehkämper[1], Tina van de Flierdt[1]

[1] Grantham Institute for Climate Change and the Environment, Imperial College London, South Kensington Campus, London SW7 2AZ, United Kingdom

[2] Department of Earth Science and Engineering, Imperial College London, South Kensington Campus, London SW7 2AZ, United Kingdom

[*]Corresponding author: a.olivelli21@imperial.ac.uk

Keywords:

Lead, lead isotopes, marine geochemistry, machine learning, explainable AI

Note 1: Exclusion of geospatial coordinates as features for Pb concentration, ^{206}Pb/^{207}Pb and ^{208}Pb/^{207}Pb models

As briefly mentioned in the main text, geospatial coordinates were initially included as features on which to train the Pb concentration, ^{206}Pb/^{207}Pb, and ^{208}Pb/^{207}Pb models. To ensure continuity of the data, coordinates were transformed using *n*-vector transformations of latitude (λ) and longitude (μ), such that:

$$\begin{aligned} Coordinate\ A &= \sin(\lambda) \\ Coordinate\ B &= \sin(\mu)\cdot\cos(\lambda) \\ Coordinate\ C &= -\cos(\mu)\cdot\cos(\lambda) \end{aligned}$$

While the models trained with coordinates returned performances comparable to those of models trained without them (Pb concentration: MAPE = 23.6 %, RMSE = 3.81 pmol/kg, R^2 = 0.91; ^{206}Pb/^{207}Pb: MAPE = 0.2 %, RMSE = 0.005, R^2 = 0.82; ^{208}Pb/^{207}Pb: MAPE = 0.1 %, RMSE = 0.006, R^2 = 0.72), global reconstructions of Pb concentrations and isotope compositions showed spatial artefacts biased by the inclusion of coordinates. These artefacts included sharp transitions from low to high values at adjacent locations, as well as strong impacts of coordinate values on the predicted Pb concentrations and isotope compositions. These model artefacts could be explained by the very different frequency distributions of coordinate values between the models' training set and prediction dataset, due to the scarcity of data and clustering of sampling efforts in areas such as the Atlantic Ocean and North Pacific (Fig. 1, main text).

For these reasons, we opted to exclude geospatial coordinates from the list of features on which the models were trained.

Note 2: Performance of the Pb concentration, ^{206}Pb/^{207}Pb and ^{208}Pb/^{207}Pb models build using the Random Forest algorithm

The same procedure used to develop the Pb concentration, ^{206}Pb/^{207}Pb and ^{208}Pb/^{207}Pb models with the XGBoost algorithm (Sect. 2.2 of the article) was also followed to build models based on the Random Forest (RF) algorithm, in order to identify the best performing architecture. In contrast to XGBoost, the trees in the RF are built all at the same time and only use a random subset of training features during the creation of each individual tree. The final prediction of the RF algorithm for a regression task is the average of the predictions made by all decision trees in the ensemble.

A set of hyperparamters were tuned for all three models, including the number of trees in the ensemble (‘*n_estimators*’), the maximum depth of each tree (‘*max_depth*’), the minimum number of observations in a node for it to split (‘*min_samples_split*’), the minimum number of samples required to be at a leaf node (‘*min_samples_leaf*’), whether bootstrapping was used when building trees (‘*bootstrap*’), and the maximum number of features considered by each tree (‘*max_features*’). A detailed overview of the hyperparamter space explored and the best values for each hyperparameter is provided in Table S1 below.

Compared to XGboost, RF performed overall worse for the Pb concentration and ^{208}Pb/^{207}Pb models both on the random test set and the geographic one. Indeed, the best RF Pb concentration model performance returned R^2 = 0.86, RMSE = 4.97 pmol/kg, and MAPE = 21.3% on the random test set and R^2 = 0.80, RMSE = 5.29 pmol/kg, and MAPE = 19.7% on the geographic test set. Similarly, the best RF ^{208}Pb/^{207}Pb model performance was R^2 = 0.72, RMSE = 0.006, and MAPE = 0.1% on the random test set and R^2 = 0.33, RMSE = 0.008, and MAPE = 0.002 on the geographic test set. Contrarily, the RF ^{206}Pb/^{207}Pb had an only slightly better performance than its XGBoost counterpart. In detail, it had a R^2 of 0.81, RMSE of 0.005, and MAPE of 0.3% on the random test set, and R^2 of 0.77, RMSE of 0.005, and MAPE of 0.3% on the geographic test set.

Overall, given the much better performance of the XGBoost algorithm for the Pb concentration and ^{208}Pb/^{207}Pb models, and comparable performance for the ^{206}Pb/^{207}Pb model, we decided to use XGBoost for the development of all final models.

Table S1. Hyperparameter space explored for the Random Forest regression models. Bold values identify the combination of hyperparameters that returned the best model performance.

Hyperparameter	***Pb concentration***					**$^{206}Pb/^{207}Pb$**					**$^{208}Pb/^{207}Pb$**				
bootstrap	**True**	False				**True**	False				**True**	False			
n_estimators	400	600	800	**1000**	1200	**400**	600	800	1000	1200	**400**	600	800	1000	1200
max_depth	**None**	10	20			**None**	10	20			**None**	10	20		
min_samples_split	**2**	5	10			**2**	5	10			**2**	5	10		
min_samples_leaf	1	**2**	4			1	**2**	4			1	**2**	4		
max_features	**None**	sqrt	log2			**None**	sqrt	log2			**None**	sqrt	log2		

Figures

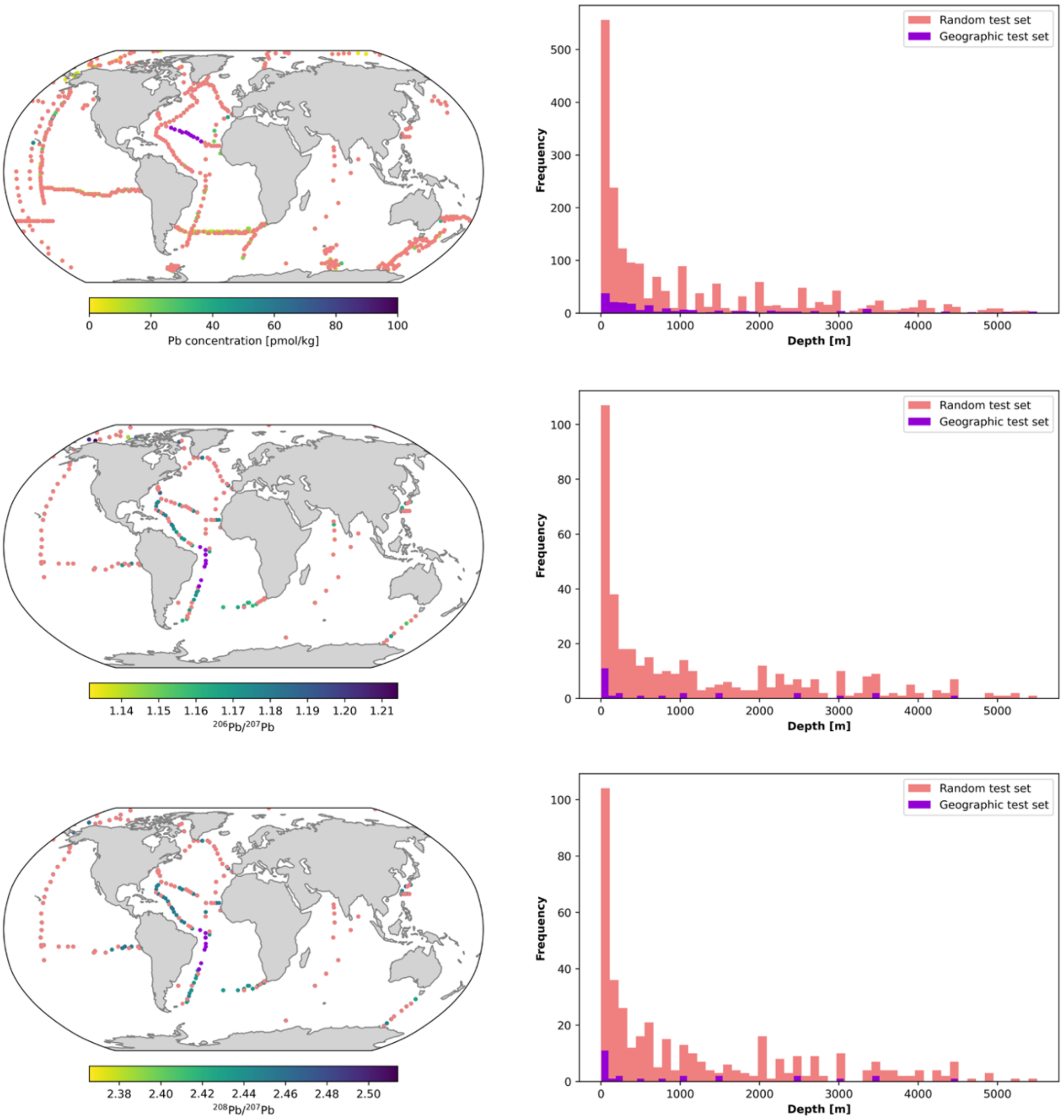

Figure S1. Location (left) and depth frequency distribution (right) of samples that make up the random and geographic test sets for the Pb concentration (top), ^{206}Pb/^{207}Pb (middle), and ^{208}Pb/^{207}Pb (bottom) models.

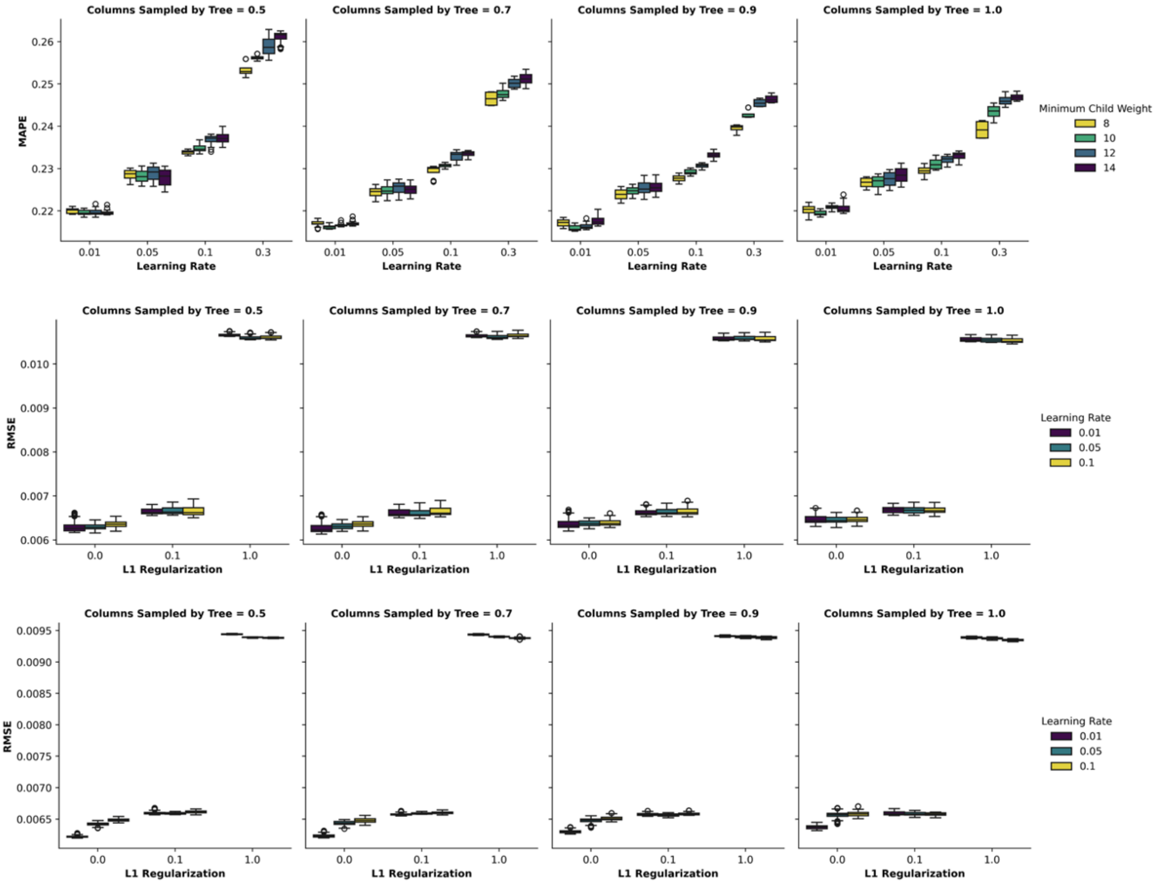

Figure S2. Distribution of MAPE and RMSE values for different hyperparameter combinations for the Pb concentration (top), ^{206}Pb/^{207}Pb (middle), and ^{208}Pb/^{207}Pb (bottom) models.

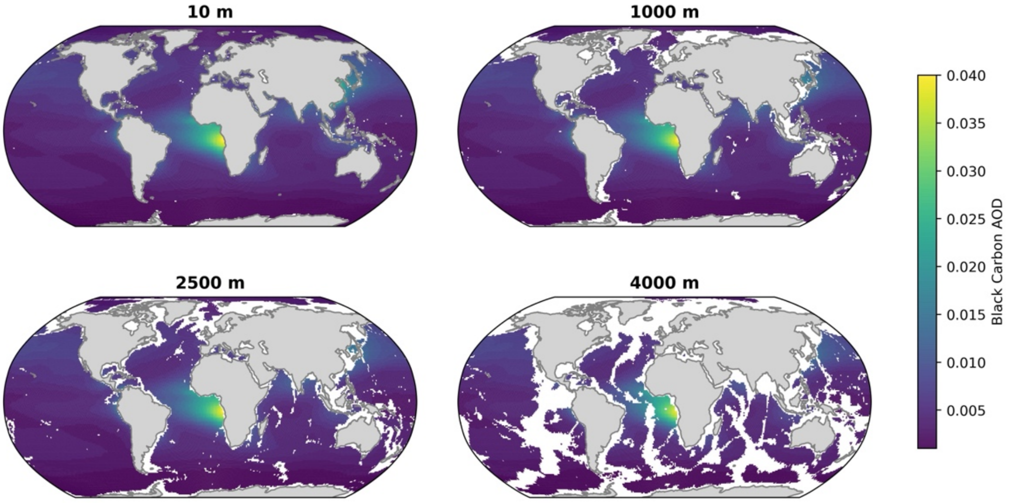

Figure S3. Global distribution of Black Carbon AOD at 10 m, 1000 m, 2500 m, and 4000 m. Data from CAMS global reanalysis (ECMWF Atmospheric Composition Reanalysis 4; EAC4). All depth levels in each 1x1 cell column were assigned the same Black Carbon AOD values.

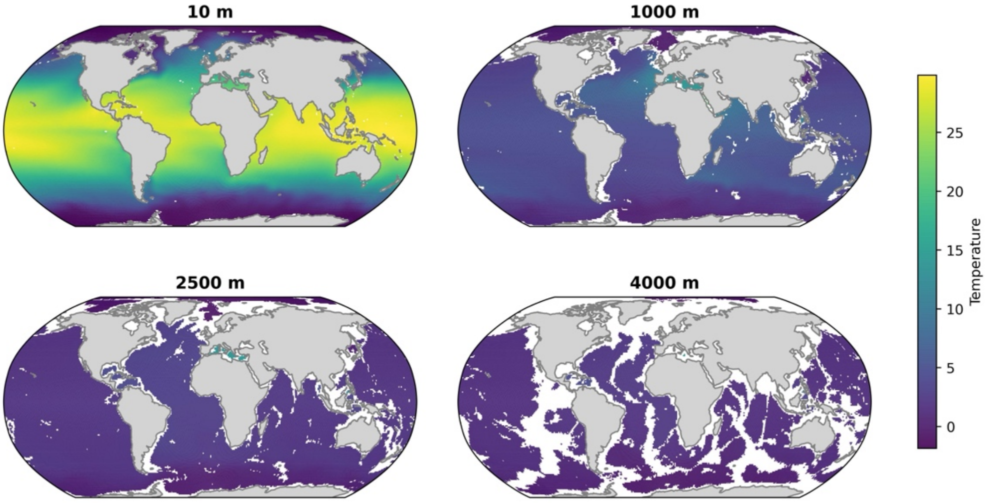

Figure S4. Global distribution of seawater temperature [°C] at 10 m, 1000 m, 2500 m, and 4000 m. Data from the World Ocean Atlas 2018.

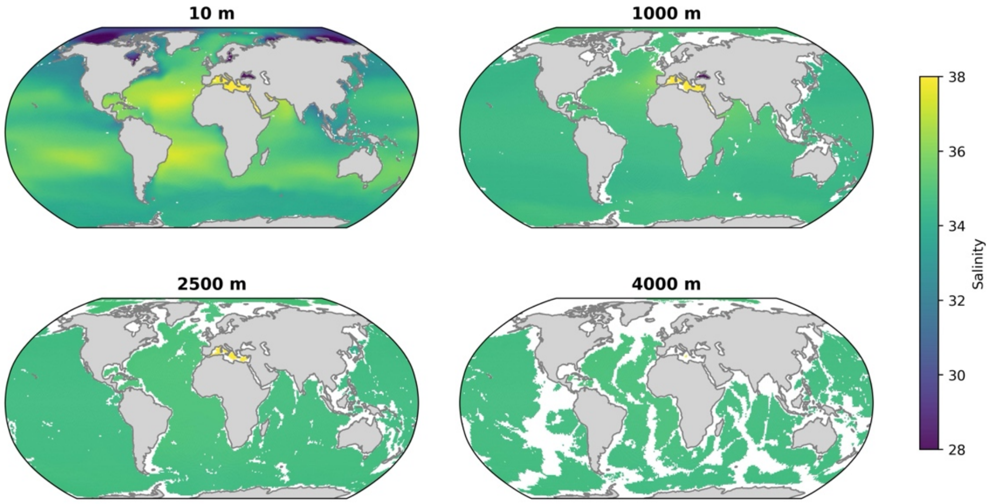

Figure S5. Global distribution of salinity at 10 m, 1000 m, 2500 m, and 4000 m. Data from the World Ocean Atlas 2018.

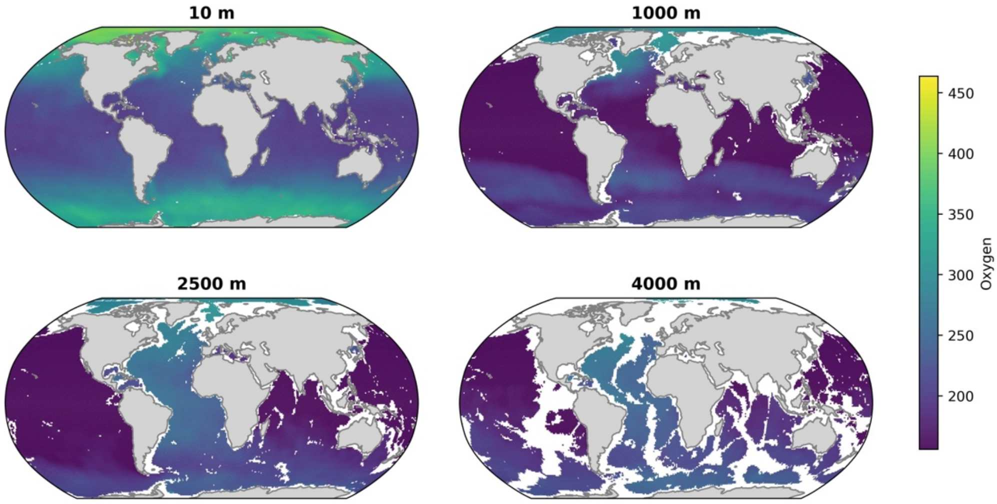

Figure S6. Global distribution of dissolved oxygen concentration [µmol/kg] at 10 m, 1000 m, 2500 m, and 4000 m. Data from the World Ocean Atlas 2018.

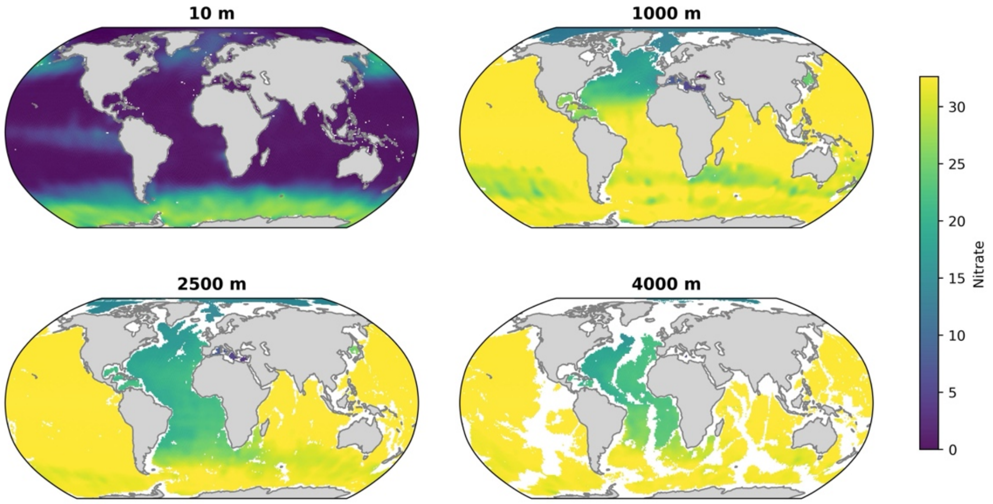

Figure S7. Global distribution of dissolved nitrate concentration [µmol/kg] at 10 m, 1000 m, 2500 m, and 4000 m. Data from the World Ocean Atlas 2018.

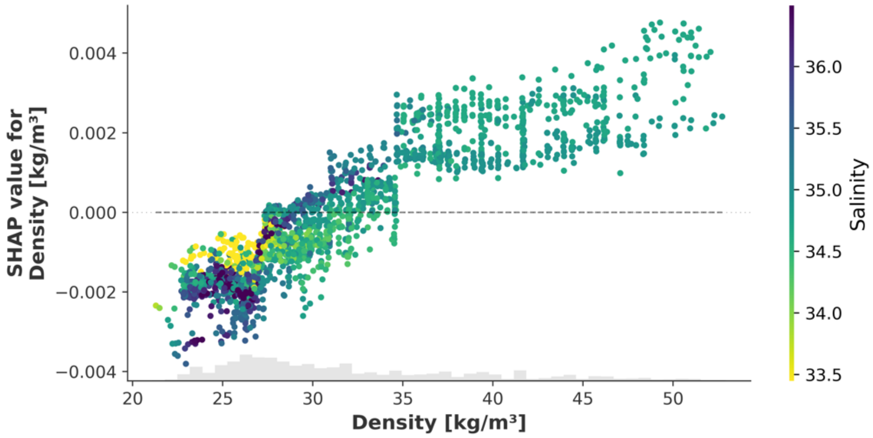

Figure S8. SHAP values for density for the ^{206}Pb/^{207}Pb model.

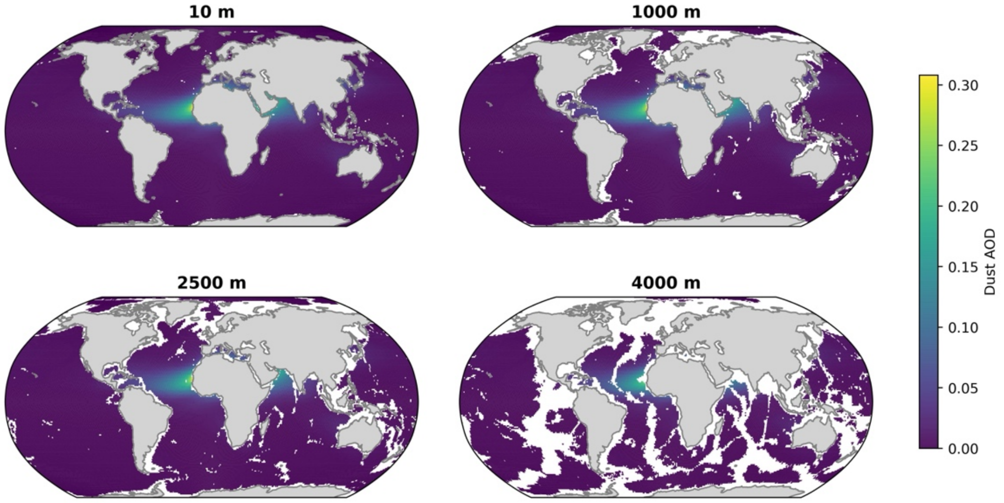

Figure S9. Global distribution of Dust AOD at 10 m, 1000 m, 2500 m, and 4000 m. Data from CAMS global reanalysis (ECMWF Atmospheric Composition Reanalysis 4; EAC4). All depth levels in each 1x1 cell column were assigned the same Dust AOD values.

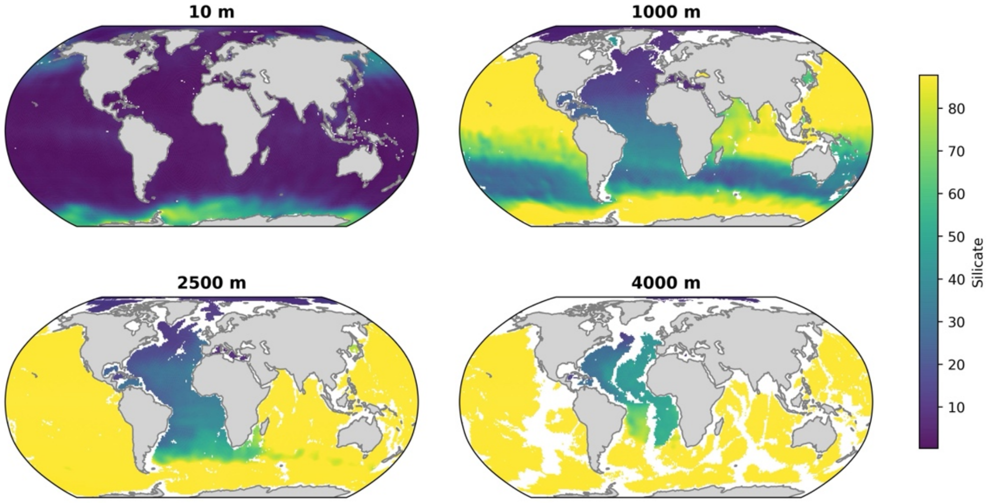

Figure S10. Global distribution of dissolved silicate concentration [µmol/kg] at 10 m, 1000 m, 2500 m, and 4000 m. Data from the World Ocean Atlas 2018.

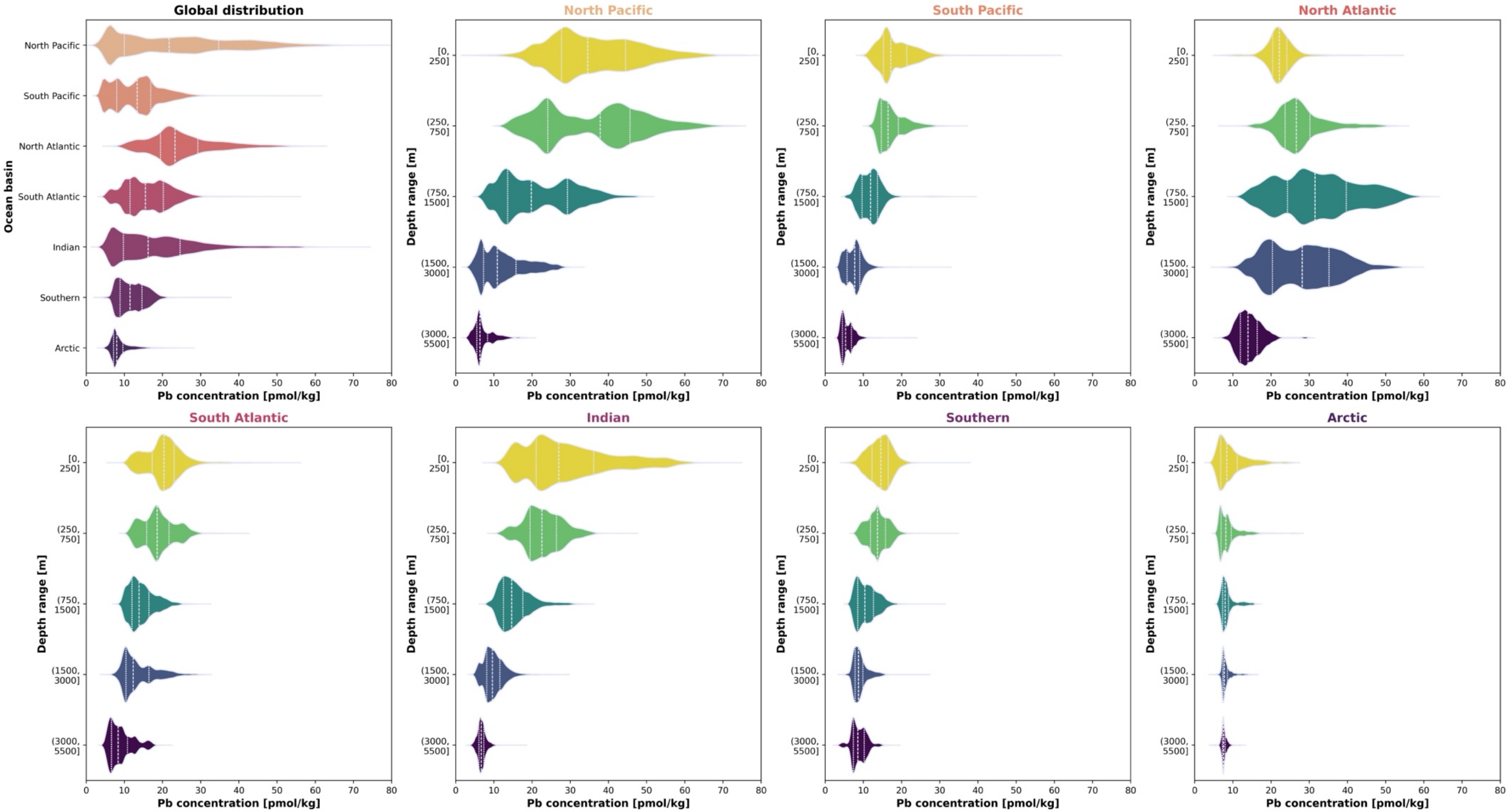

Figure S11. Violin plots of Pb concentration distributions in the different ocean basin (top left panel) and within each basin at different depth ranges. The white dashed line in each violin represents the median value, while the dotted lines represent the lower and upper quartiles (Q1 and Q3, respectively).

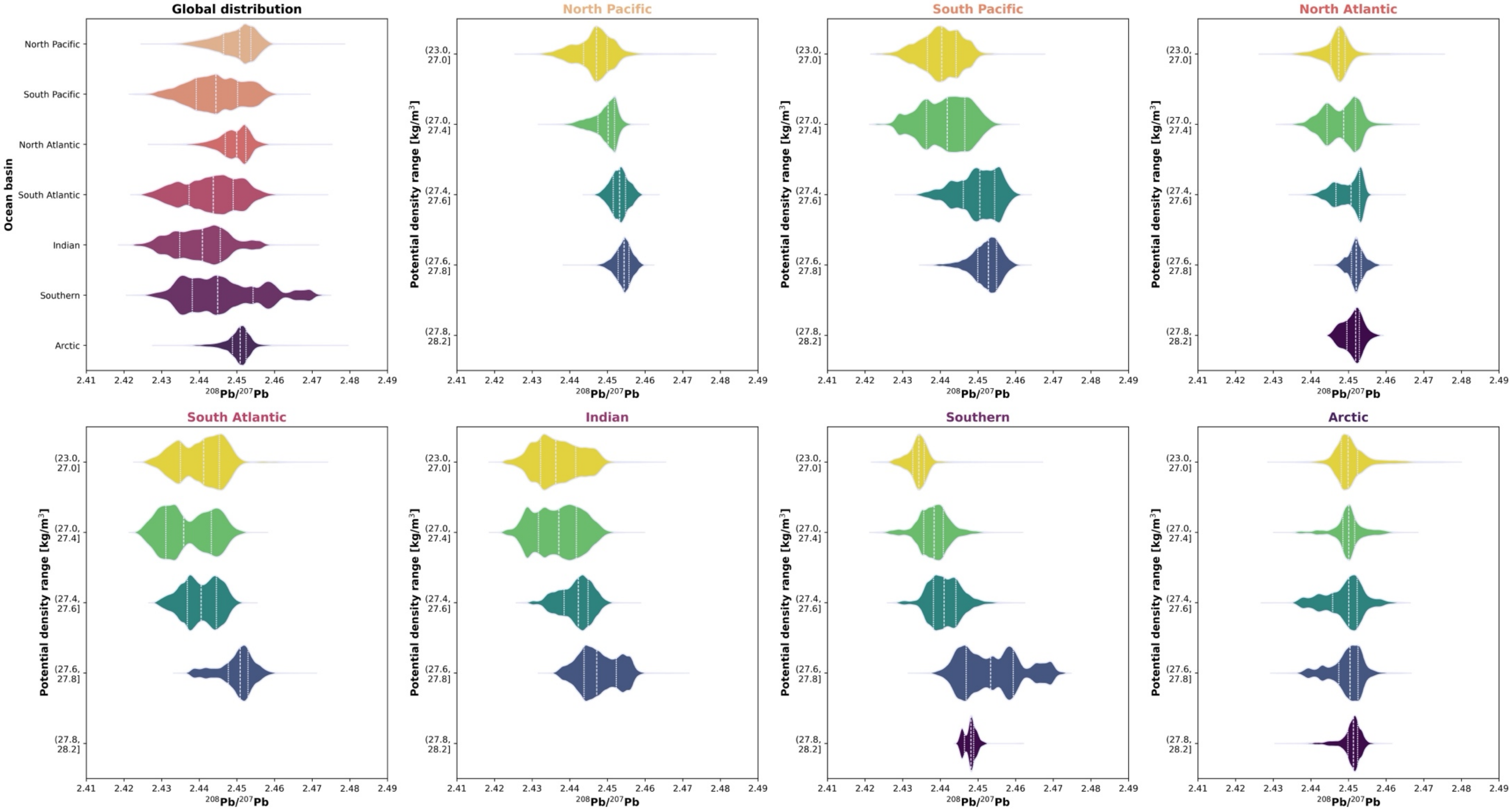

Figure S12. Violin plots of ^{208}Pb/^{207}Pb distributions in the different ocean basin (top left panel) and within each basin at different potential density ranges. The white dashed line in each violin represents the median value, while the dotted lines represent the lower and upper quartiles (Q1 and Q3, respectively).