Predicting Coccolithophore Abundances Using Statistical Learning Methods

Thomas W. Duck
10 December 2025
STAT 4950
Statistics Honours Thesis

Introduction

Coccolithophores are a group of phytoplankton noted for their external plates of calcium carbonate called coccoliths (Winter & Siesser, 1994). The most common coccolithophore species is *Emiliania huxleyi*, which can be found from subpolar regions to the equator (Taylor et al., 2017). Coccolithophores are important for two main reasons. First, as primary producers (Holligan et al., 1993), they use photosynthesis to produce energy and therefore support the food webs in the oceans. Second, they are very important to the carbon cycle of the ocean. During photosynthesis, coccolithophores remove carbon dioxide (CO_2) from the ocean, by extension removing CO_2 from the atmosphere. However, during the calcification process that makes coccoliths, CO_2 is created as a by-product, therefore putting some CO_2 back into the ocean. Coccolithophore biogeography is also important, as changes in the distribution and abundance of these organisms can be indicators of temperature and nutrient changes in the ocean.

Coccolithophore abundance data are obtained by taking samples of seawater at different depths, locations (latitude and longitude), and times of the year. Microscopic identification of species is then done for each cell, and the cells are counted to get abundances for each species. Counting is done using the Utermöhl method, which assumes a Poisson distribution of cells in the counting chamber, and are identified and counted under a microscope after waiting many hours for particles to settle out of the water column to the bottom of the sampling chamber. Sampling is done through research cruises, which collect seawater samples at different points around the ocean, and dedicated stations which take samples at the same coordinates at different times of the year. This can provide a good time-series, particularly at the dedicated observation points, but also cruises recording observations at the same points either months or years later.

There are several key environmental factors that influence coccolithophore abundance. As with everything in the ocean, water temperature is key, as abundance will be lower if the water is too warm or too cold (Charalampopoulou et al., 2016). Light exposure is also a major influence (Charalampopoulou et al., 2016), since coccolithophores use photosynthesis to get their energy. Levels of nutrients in the water, and the ratios of these nutrients, impact the calcification process. Key nutrients include nitrate, phosphate, and silicate (Charalampopoulou et al., 2016). For most of these environmental data we have collections of both compiled and analysed observations to provide global, monthly climatologies. Grazing of phytoplankton by zooplankton is undoubtedly a key factor regulating their abundance, but data on grazing is very sparse and unlikely to be a reliable predictor.

The main database used to train the models is called coccobase (Naud et al., 2025). Coccobase is a compilation of coccolithophore observations taken between 1905 and 2015, with most data being from 1980 onwards. This database was compiled partly from two previously made compiled databases, namely the 2013 MARine Ecosystem DATa (MAREDAT) project compiling global coccolithophore abundance observations, and the 2020 PhytoBase project compiling mainly observations of coccolithophore presence only. Coccobase was expanded from those two compilations by systematically scraping Google Scholar and PANGAEA for relevant datasets, and added accordingly. In total, coccobase has 213,671 individual coccolithophore observations, which contain 76,769 presences recorded and 42,873 cell density estimates. 63.1% of observations are recorded in the northern hemisphere, very disproportionate since the northern hemisphere has only 43% of the oceans. There are 190 species recorded in coccobase, but 42 species account for 75% of observations, with 24 species having more than 2,000 observations or

500 presences. 21 species account for 75% of cell density observations. *E. huxleyi* has the most at over 8,500 observations, with five other species having over 950 cell density observations.

Given the sampling methods of coccolithophore abundance, there are several challenges impacting turning this database into an observational map. First, with cruise data, these can be seasonal data depending on the viability of sailing to various parts of the ocean. This causes cruises to give good spatial coverage with the data, but it is not as thorough temporally, particularly in the northern and southern winters. Secondly, with both cruises and dedicated observation stations, the sampling is not random. The locations for coccolithophore abundance observations are pre-determined, and as a result, much of the ocean has no data. This is a particular problem in the middle of the Indian Ocean, and the central and south Pacific Ocean. Conversely, there is no shortage of data off the coasts of Peru, Japan, and western and central Europe, causing these regions to be far easier to predict by comparison.

The goal of this project is to turn the obtained database of coccolithophore abundance data into a predictive map of coccolithophore abundance in the oceans for each month. This will be done with both total coccolithophore abundance, and species-specific maps. The four species that will be used for species-specific maps are *Coccolithus pelagicus*, *Discosphaera tubifera*, *Emiliania huxleyi*, and *Ophiaster hydroideus*. The general method of this project will be to round observation locations to the nearest degree, then attach these abundance measurements to a dataset of various oceanographic and environmental factors compiled from multiple sources. Using this dataset, five models will be trained, and then predictions will take place across the oceans for each month based on the model. These will include linear regression, tree-based models, and full machine learning models.

Materials and Methods

Data

This project uses two datasets. The first is a compiled dataset of coccolithophore observations, called coccobase. This dataset contains observations taken between 1903 and 2015, and has been filtered to exclude observations more than 50 metres deep. There are 244 species of coccolithophores in this dataset (Naud et al., 2025). Coccobase records the coordinates, date, depth, species, abundance, and whether coccolithophores were present or absent at the coordinates, among other things which are not of interest to this project. The second dataset used is a compiled dataset called environment data ocean, referred to as EDO. EDO was compiled from several sources. These are the 2023 edition of the World Ocean Atlas (WOA) (Reagan et al., 2024), dissolved iron measurements (Tagliabue et al., 2016), the SeaWiFS satellite, the NEMO project, the Glory and de Boyer Montégut (de Boyer Montégut et al., 2004) mixed layer depth models, and some variables that were mutated from those that were compiled from the sources. EDO has a total of 534,349 observations with 29 variables, corresponding to nearly complete coverage of the ocean at 1° spatial and monthly temporal resolution. EDO has no coccolithophore abundance measurements. Because of this, coccolithophore abundance data coupled with climatological and environmental conditions will be used to train the models, and the models will predict coccolithophore abundance on EDO where we do not have biomass observations. This coupled dataset has 34 predictor variables (Table 1).

The response variable in this project is called log biomass. Log biomass is the base 10 logarithm transformation of the biomass per litre variable. This conversion is done to account for different sizes in the coccolithophore species. Obtaining biomass per litre starts with abundance counts of cells per litre (Naud et al., 2025). These abundance counts are converted to carbon per

cell, which is then multiplied by the cell count to get the total biomass per litre. Thus, the biomass per litre accounts for picograms of organic carbon per litre. All biomass per litre values equal to zero are disregarded, and the models are just trained on those observations for which there are abundance values. This was done because biomass per litre values equal to zero become undefined when the logarithm is taken, which would needlessly complicate the model and have an unreliable scale for predictions. Log biomass is the response variable for the total coccolithophore abundance model, as well as the aforementioned species-specific models.

The inconsistencies in locations of coccolithophore abundance observations make the model potentially very strong in some places and very weak in others. For example, the North Atlantic Ocean and the North Sea have a large number of observations (Figure 1). Conversely, there are many undersampled regions, such as the South Pacific and central Indian Oceans. Thus, the data will be very strong in achieving the goals of this project in the regions with lots of observations, while the lack of data in other regions must be dealt with to get statistically sound abundance predictions. Concurrent with this, there are many fewer observations in November and December compared to the other months (Table 2). This is due to the northern hemisphere winter, so there are fewer samples being taken in that half of the world, where the vast majority of observations are.

Missing data

EDO has a lot of missing data. After removing 913 observations which do not have a value for month, there are 463,960 missing data points of the 15,469,644 total data points, accounting for just under 3% of the data. Much of this data is in the Arctic or Antarctic during the winter. Rather than removing all observations with any missing values, the missing data is

imputed. This is done using the gstat R package (Pebesma & Graeler, 2025), which uses inverse distance weighted k-nearest neighbour (k-NN) imputation. The gstat package is designed specifically for doing k-NN imputations using spatial data. Accordingly, it assigns a weight to each of the nearest neighbours based on real-world distance, rather than using a Euclidean distance calculated from all predictor variables. The weight assigned to each neighbour is $\frac{1}{d^{idp}}$, where d is the real-world distance, and idp is a pre-defined weighting constant. This gives closer points more weight among the nearest neighbours, and that weight increases exponentially based on the choice of idp. In this project, imputation was done using $k = 5$, and $idp = 2$. These values were chosen due to strong performance when imputing variables on a testing set in both shallow (<500 metres, >0.95 R^2 for all variables) and deep water (≥500 metres, >0.98 R^2 for all variables). Since EDO has 12 observations at each point, one per month, the missing data were imputed based only upon other observations from the same month. This means that the imputations will not be impacted by 5 observations at the same location for different months; rather the 5 nearest neighbours will be at different locations than the observation being imputed.

Statistical model development

In developing the models, the response variable is log biomass, which is the base 10 logarithm of the biomass of the whole coccolithophore community (pg C L^{-3}). For observations that have zero biomass, the log biomass is undefined. Those observations are ignored in the model. Five models will be run; neural network, random forest, recursive partitioning and regression trees (rpart), extreme gradient boosted trees (XGBoost), and generalised linear model with elastic net regularisation (GLM). All of the predictor variables were used in training the model so that variable importance could be determined, with the exception of month, latitude,

longitude, and area of each 1° by 1° cell. Those four variables were excluded so that abundance predictions could be made only using environmental and oceanographic variables. These models will be compared by their R^2 when predicting on a test set to determine the best model, which will be analysed further. Each of these models will be run on the total coccolithophore abundance dataset, as well as the four coccolithophore species datasets.

All models in this project were developed within the Tidymodels framework (Kuhn & Silge, 2022), using the workflow set and grid search capabilities to run many reproducible models at once. The first model is a normalised feed-forward neural network implemented using the nnet package (Venables & Ripley, 2002). The nnet neural network consists of an input layer of predictor variables, a single hidden layer with nonlinear activation functions, and an output layer producing predictions of log_biomass. Before training, all predictor variables were normalised to the range [0, 1], as noted to be necessary in the documentation. Model training used gradient descent to minimise the sum of squared errors between observed and predicted values, with an L2 weight decay term penalising the magnitude of the connection weights to reduce overfitting. The hyperparameters for number of hidden units, penalty strength, and number of training epochs, were tuned using resampling procedures provided by the tune package. A neural network was chosen because of its flexibility and ability to capture nonlinear relationships, as may exist in data like this.

The second model is a random forest implemented using the ranger package (Wright & Ziegler, 2017). This package is a fast implementation of random forests, originally introduced by Breiman (2001). Random forests operate as ensembles of decision trees, each trained on a bootstrap sample of the data. 1,000 trees were used in the model in this project. During tree construction, a random subset of predictors of size mtry is considered at each node. This

introduces randomness and reduces correlation between trees. Each tree grows until reaching a minimum node size (min_n), and final predictions are obtained by averaging across all trees. The hyperparameters mtry and min_n were optimised using the tune package. A random forest was selected for its robustness to noise and ability to handle large datasets.

The third model is a single decision tree implemented using the rpart package (Therneau & Atkinson, 1997). This package recursively partitions the predictor space by minimising the within-node sum of squares. Model complexity was governed by two hyperparameters; first being the maximum tree depth (tree_depth), which restricts the number of sequential splits, and second being the cost-complexity parameter (cost_complexity), which prunes back overly complex trees to prevent overfitting. Both were tuned using the tune package. A model using rpart was chosen to use another tree based model, but not one with a large number of trees per iteration, like the random forest. This makes a model using rpart much less computationally expensive.

The fourth model uses extreme gradient boosted trees implemented using the XGBoost package (Chen & Guestrin, 2016). This package builds trees sequentially, with each new tree attempting to correct the residuals of the previously constructed trees. During optimisation, a loss function is minimised using gradient descent, while regularisation terms constrain model complexity. Key hyperparameters included the learning rate (learn_rate), which controls how much each tree contributes to the model; the maximum depth of trees (tree_depth); the minimum loss reduction required for further partitioning (loss_reduction); the minimum number of observations per terminal node (min_n); and the subsampling rate (sample_size). The total number of trees (trees) determined the ensemble's overall size. All hyperparameters were tuned using the tune package. Extreme gradient boosted trees were chosen as a model due to its

relatively high predictive accuracy compared to other models, and its relative computational efficiency by using a greedy algorithm as opposed to a global optimum.

The fifth model is a generalised linear model with elastic net regularisation, implemented using the glmnet package (Friedman et al., 2010). This method estimates regression coefficients by minimising a weighted least squares objective with an added penalty term that combines ridge (L2) and lasso (L1) regularisation. The penalty parameter controls the overall degree of shrinkage applied to the coefficients, while the mixture parameter (α) determines the balance between ridge and lasso regularisation: $\alpha = 1$ is pure lasso, $\alpha = 0$ is pure ridge, and intermediate values are elastic net regularisation. These penalties reduce overfitting by shrinking coefficients, with lasso performing variable selection and ridge stabilising correlated predictors. Hyperparameters were tuned using the tune package. A generalised linear model with elastic net regularisation was chosen for its computational efficiency, interpretability, and strong performance when relationships between the predictor variables and the response variable are approximately linear.

Results

Model performance varied substantially across the five statistical learning approaches tested (Table 3). XGBoost achieved the strongest predictive accuracy for total coccolithophore abundance predictions, with an R^2 of 0.59 on the test data and the lowest root-mean-squared error among all models. This indicates that the boosted tree model captured a substantial proportion of the variance in log biomass and was best able to learn the nonlinear and interacting effects present in the environmental predictors. Random forests performed nearly as well, with an R^2 of 0.58, demonstrating that ensemble tree methods in general were effective at modelling the

structure of coccolithophore biomass. The neural network achieved moderate predictive performance, with an R^2 of 0.54, suggesting that while it can model nonlinear relationships, it was somewhat more sensitive to the sparsity and uneven sampling characterising the training data. The rpart decision tree produced an R^2 of 0.44, reflecting its inability to capture complex interactions with only a single tree. The elastic-net GLM performed worst, with an R^2 of 0.22, showing that the dominant ecological relationships in the data are not well approximated by a primarily linear model.

These results were also seen in the species-specific coccolithophore abundance models (Table 3), with the exception of the *Ophiaster hydroideus* model. *O. hydroideus* had no model with a prediction R^2 above 0.35, which was the random forest model, but both the neural network and the generalised linear model performed better than the XGBoost model. The tree-based models performed the best for each of the other three species, with neural networks not far behind, and both generalised linear model and recursively partitioned trees performed the worst. This is consistent with the prediction trends for total biomass. Models for both *Coccolithus pelagicus* and *Emiliania huxleyi* had better prediction R^2 than the total biomass predictions (Table 3).

Residual diagnostics showed that the XGBoost model produced spatially and temporally unbiased predictions. The spatial residual map (Figure 2) revealed no regionally concentrated over- or under-prediction, even in areas historically undersampled, such as the South Pacific and central Indian Ocean. This suggests strong generalisation ability and indicates that the environmental drivers encoded in the model are sufficiently informative to extrapolate to remote regions. The temporal residual boxplots (Figure 3) likewise revealed no systematic patterns across months. Despite the strong seasonality of coccolithophore populations and the imbalance

of observations across the annual cycle, residual variability remained consistent, implying that the model does not disproportionately misrepresent either hemispheric winter or summer months.

Predicted global biomass showed broad seasonal variability in the mean, reflecting the underlying seasonal structure of coccolithophore populations. Standard errors (Figure 4) varied little across the annual cycle, despite combining data from two hemispheres with opposite seasons. The consistency of these standard errors suggests that model uncertainty is dominated by local environmental variability rather than systematic seasonal effects. Importantly, no month exhibited unusually high uncertainty, supporting the model's temporal robustness even in months with sparse empirical coverage.

Variable importance values, derived from permutation-based metrics (Greenwell & Boehmke, 2020), revealed clear patterns in the ecological drivers of coccolithophore biomass (Figure 5). Two derived stoichiometry variables, si_star (silicate minus nitrate) and n_star (nitrate minus 16 times phosphate), were consistently identified as among the most influential predictors. These variables encapsulate nutrient balance rather than raw nutrient concentration, suggesting that the competitive and physiological constraints imposed by nutrient ratios are more predictive of community biomass than absolute concentrations alone. This is consistent with observational work showing that post-mixing nutrient stoichiometry helps determine the relative dominance of calcifiers and organisms requiring silicate (Leblanc et al., 2009). Other high-importance variables included day length, iron availability, mixed layer depth, PAR, chlorophyll, and several physical variables including temperature and diffuse light attenuation. Although chlorophyll and nitrate are typically linked to general phytoplankton biomass, their high importance here indicates that broader ecosystem productivity and nutrient cycling are key determinants of coccolithophore population structure (O'Brien et al., 2013).

The correlation heatmap (Figure 6) and hierarchical clustering dendrogram (Figure 7) provided additional insight into the structure of the predictor variables. The nutrient variables; nitrate, phosphate, and silicate from both WOA and NEMO, formed predictable high-correlation clusters. However, most of the remaining predictors exhibited weak or moderate correlations, forming small, largely independent clusters. This suggests that the predictor space is rich in non-redundant information, allowing models such as XGBoost to exploit numerous independent environmental gradients. The dendrogram confirmed this structure by grouping variables into tight clusters. This independence likely contributed to the strong performance of tree-based models, which can leverage many distinct and interacting environmental pathways.

The generated global prediction maps provide a foundation for comparison with satellite-derived particulate inorganic carbon. Because particulate inorganic carbon is influenced partly by coccolithophore calcification, spatial concordance between predicted biomass and particulate inorganic carbon should offer an external validation pathway. These comparisons will help assess whether high-biomass regions coincide with elevated particulate inorganic carbon concentrations, particularly in known coccolithophore hotspots such as the North Atlantic bloom region and the Southern Ocean margins. These global prediction maps of total coccolithophore biomass based on the XGBoost model capture the seasonal variation well. Between October (Figure 8) and March (Figure 9), the southern hemisphere has more coccolithophores and the northern hemisphere has fewer coccolithophores, while these trends flip between April (Figure 10) and September (Figure 11). The high abundances in the sub-polar regions are well captured, while around the equator has many fewer coccolithophores than the tropical regions.

Discussion

The modelling results demonstrate that coccolithophore biomass can be predicted reasonably from large-scale environmental conditions, with XGBoost providing the strongest performance across all approaches tested. The relatively high R^2 of 0.59 (Table 3) for the boosted tree model predicting total biomass indicates that the majority of meaningful variability in coccolithophore abundance is captured by a combination of nonlinear interactions among light, nutrient balance, iron availability, mixed layer structure, and temperature. This aligns with ecological understanding of coccolithophore physiology and supports the idea that these populations respond to multiple limiting factors rather than any single environmental driver (Nissen et al., 2018).

Tree-based models clearly outperformed the neural network and linear models (Table 3). The random forest's R^2 of 0.58, only marginally lower than XGBoost, suggests that ensemble decision trees offer a particularly suitable architecture for this type of data. This is likely because the predictor space is ecologically heterogeneous and non-Gaussian, with many interactions and region-specific parameter spaces. Tree-based models, especially boosted trees, are able to partition complex relationships in a way that mirrors ecological processes. The neural network performed reasonably well but likely struggled with the uneven sampling density, wide range of value scales, and noise inherent in abundance data compiled from many independent studies. Conversely, the GLM's poor performance reinforces that coccolithophore biomass responds to nonlinear environmental gradients that cannot be approximated by additive or near-additive linear effects, even with shrinkage penalties.

Residual analyses strengthen the conclusion that the XGBoost model generalises well beyond the training data. Neither spatial (Figure 2) nor temporal (Figure 3) residuals displayed

systematic structure, suggesting that the model does not disproportionately rely on heavily sampled regions such as the North Atlantic and European coastal waters. Instead, the model appears to base its predictions on environmental gradients that hold across basins, which is critical given the sparse coverage in much of the Pacific and Indian Oceans. The absence of strong temporal structure in the errors is also notable, given that sampling is strongly seasonal. The stability in monthly standard errors (Figure 4), despite opposite hemispheric seasons, indicates that the model is integrating signals from both hemispheres without overfitting to one.

The strong importance of nutrient stoichiometry (si_star and n_star) suggests that coccolithophore biomass is controlled not merely by nutrient supply but by the relative balance between nutrients (Figure 5). Coccolithophores typically thrive in low-nutrient, stratified systems where diatoms are disadvantaged, and the stoichiometric variables reflect the potential for competitive exclusion by high-silicate or high-phosphate conditions. This is consistent with the idea that coccolithophores thrive when there is less silicate, leading to less competition from diatoms (Leblanc et al., 2009). The importance of iron also reflects known ecological controls, as coccolithophores frequently co-occur with other small phytoplankton in regions where iron limitation suppresses diatom competitors (Leblanc et al., 2009). Combined, these patterns support a view that coccolithophore distributions emerge from multi-nutrient limitation dynamics embedded within broader ecosystem structure, emphasising a need for models that capture many complex environmental relationships.

The relatively low correlation among most predictors suggests that the model benefits from a set of largely independent environmental dimensions (Figure 6). This helps prevent multicollinearity issues that might hamper GLM performance and enables models like XGBoost to exploit multiple weak signals that together form a strong predictive framework. The

dendrogram structure (Figure 7), showing only a few tight clusters, implies that simplification of the predictor space within these clusters may not risk a loss of meaningful ecological information or a meaningful reduction of predictive power, while still capturing the many complex relationships between predictors.

Finally, the predicted global biomass maps (Figure 8, Figure 9, Figure 10, Figure 11) provide an opportunity for external validation through comparison with satellite-derived particulate inorganic carbon. Particulate inorganic carbon is influenced primarily by coccolith mass and production rates, but its spatial patterns are known to align with regions of high coccolithophore abundance (Nissen et al., 2018). Comparing predicted biomass to particulate inorganic carbon layers will help determine whether the model successfully identifies known coccolithophore bloom regions, western boundary current systems, and subpolar transition zones. Consistency between modeled biomass and particulate inorganic carbon would provide strong independent support for the ecological validity of the predictions, whereas discrepancies may reveal regions where calcification does not track biomass linearly, or where species differences in calcification require species-specific modelling. Monthly prediction maps show a strong correlation away from the coasts with monthly climatology satellite maps of particulate inorganic carbon from Moderate Resolution Imaging Spectroradiometer (MODIS) on the Aqua satellite. This shows the model in this project largely captures patterns well in deep water. However, it does not correlate very well at all near the coasts. This can be attributed to shallow water near the shoreline having much more variation than deeper water, as this shallow water has a much higher short-term variability and strong local drivers that may not be captured well on a global scale (Leblanc et al., 2009).

Conclusion

Tree-based methods showed a solid predictive performance in predicting coccolithophore biomass. The ability of these models to capture complex, non-linear relationships make it an easy choice over linear models in trying to predict these ecological data. While the models captured relationships well, and compared well against satellite data, more work is needed to better predict abundances along the coasts. Additionally, species-specific predictions showed consistency with the total biomass in terms of which models performed best. More data is needed to improve the predictions for some species, but the model performed well for *Coccolithus pelagicus* and *Emiliania huxleyi*.

Appendix

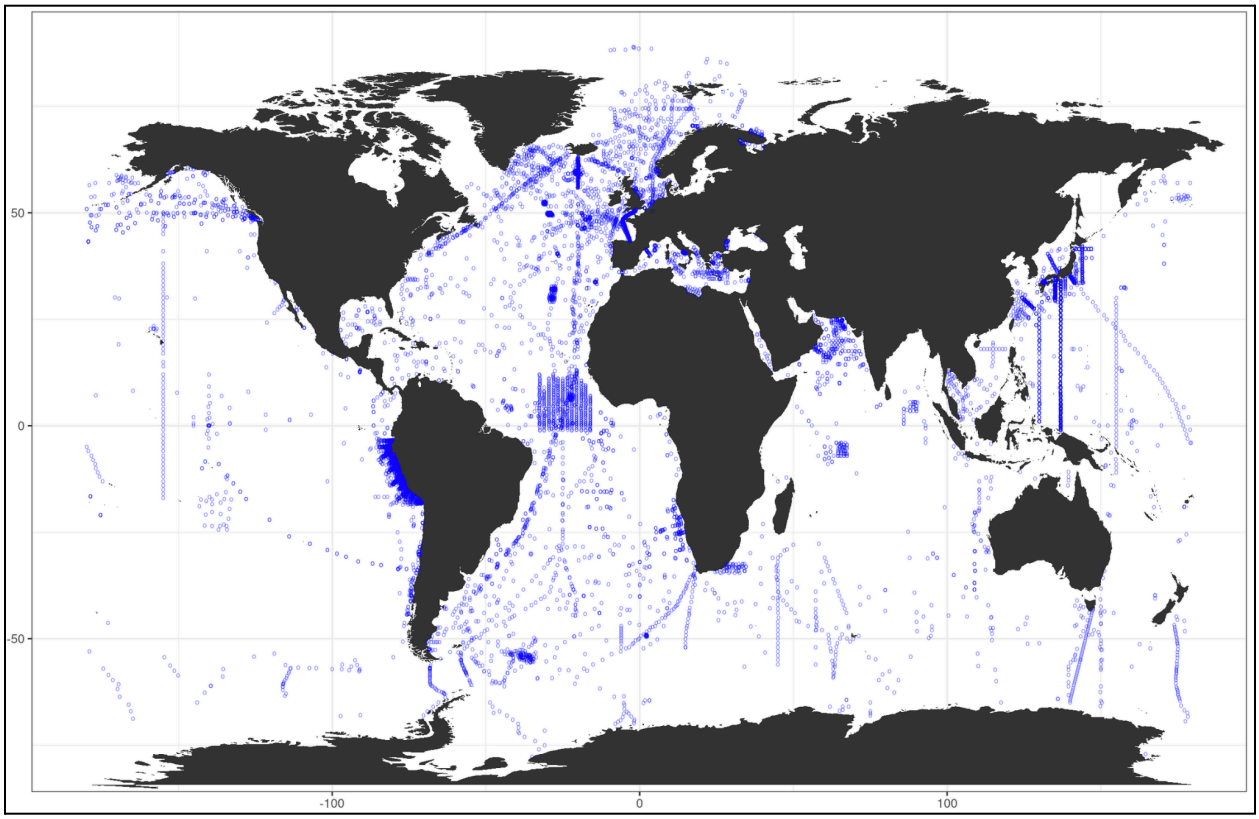

Figure 1. World map of coccolithophore abundance observations.
Distinct locations of where a coccolithophore abundance observation was obtained

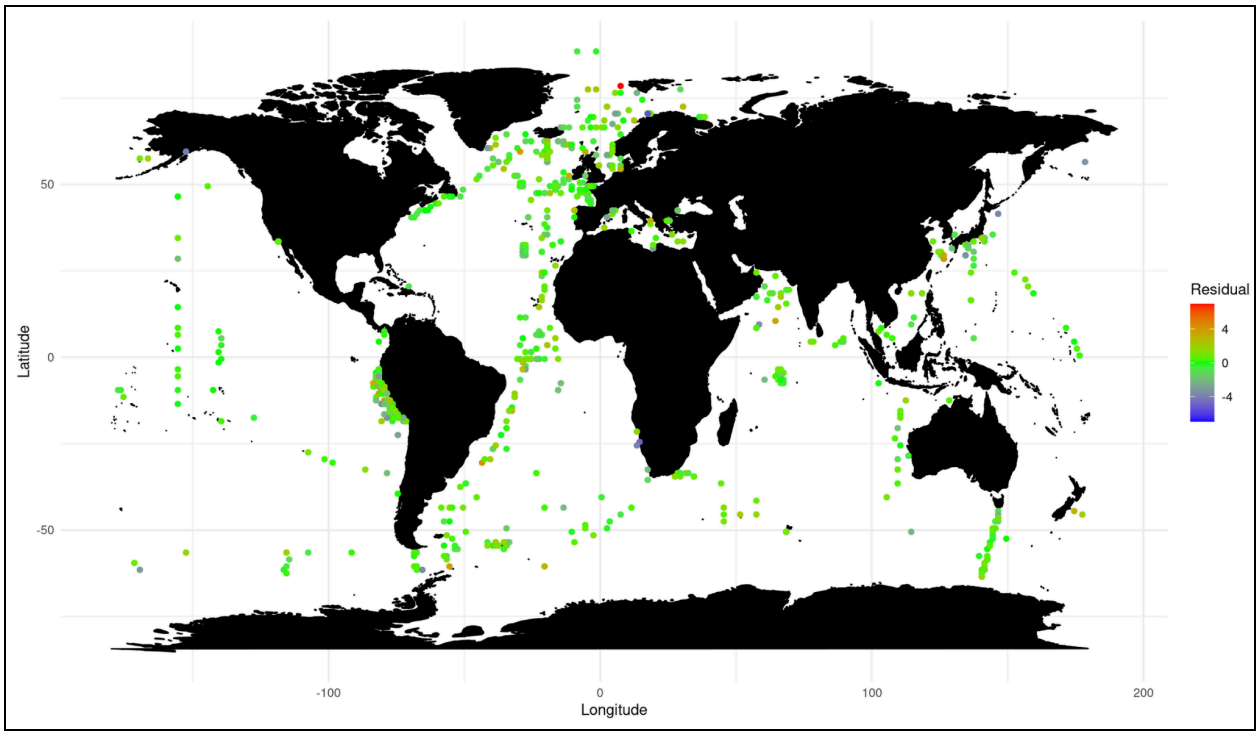

Figure 2. Spatial residual plot for the XGBoost model.
Residuals of the test dataset on the XGBoost model trained by the training dataset, plotted spatially. These data points have known abundance values, are predicted using the trained model, and then compared to find the prediction error. There is no evident spatial pattern in the residuals. The lack of structure indicates that the errors are random noise as opposed to systemic biases associated with any particular location.

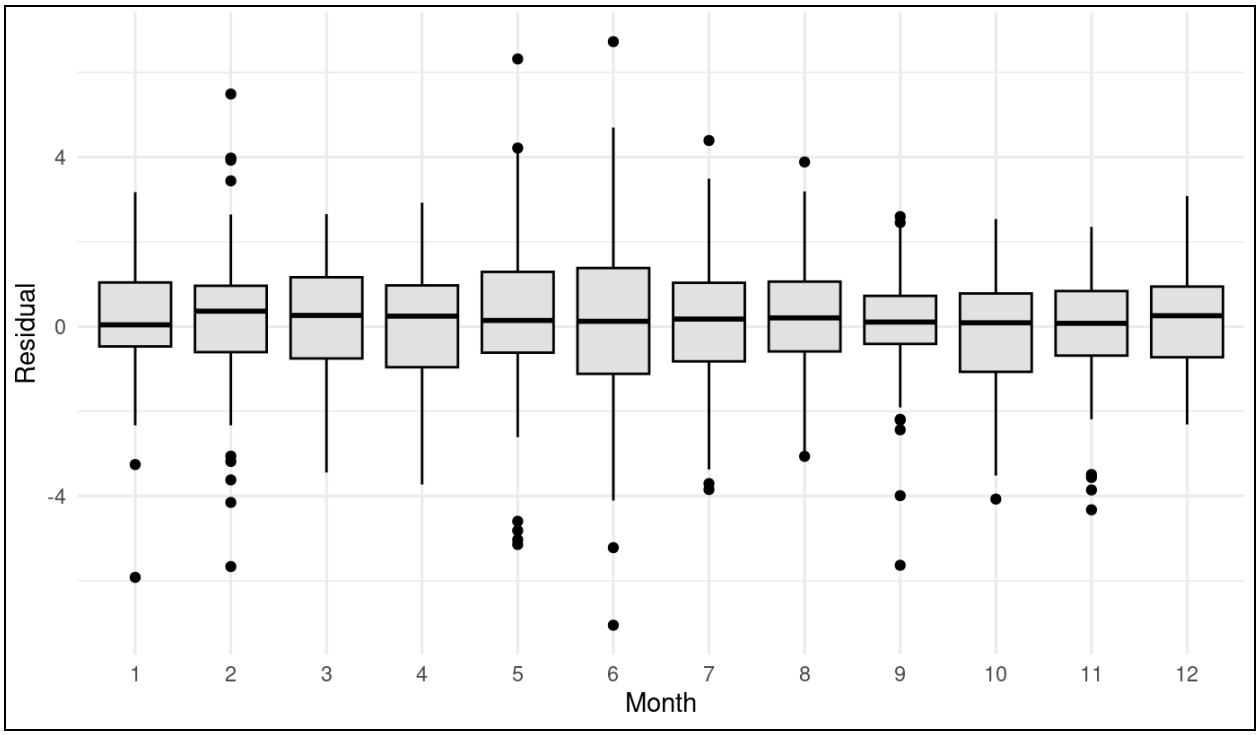

Figure 3. Temporal residual boxplots.
Residuals of the test dataset on the XGBoost model trained by the training dataset, plotted temporally. These data points have known abundance values, are predicted based on the trained model, and then compared to find the prediction error. There is no temporal pattern in the residuals. The lack of structure indicates that the errors are random noise as opposed to systemic biases associated with any particular location.

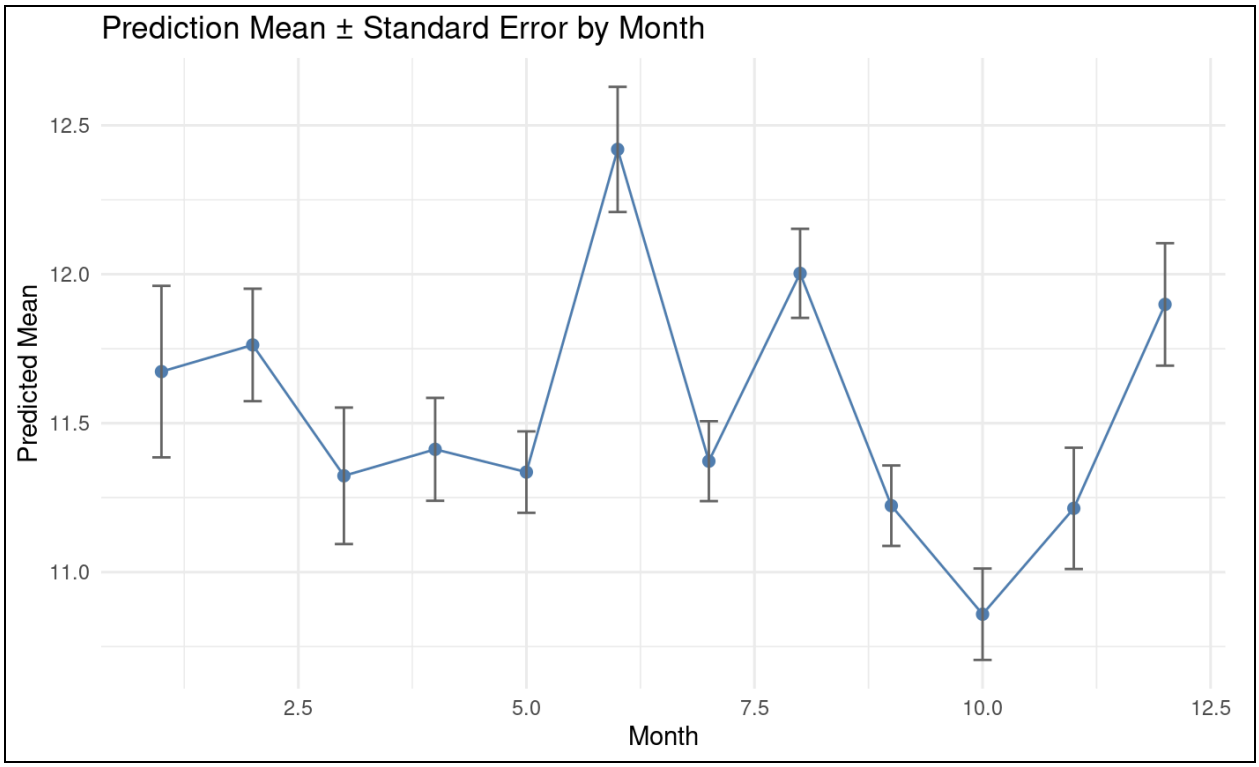

Figure 4. Standard error plot.
Coccolithophore abundance prediction mean and standard error by month. The large variation in this plot is explained by observations in each month coming from both hemispheres, so the seasons are opposite. The standard error has some variation month to month, but not enough to indicate that any month has a range large or small enough to reduce the accuracy of the overall model.

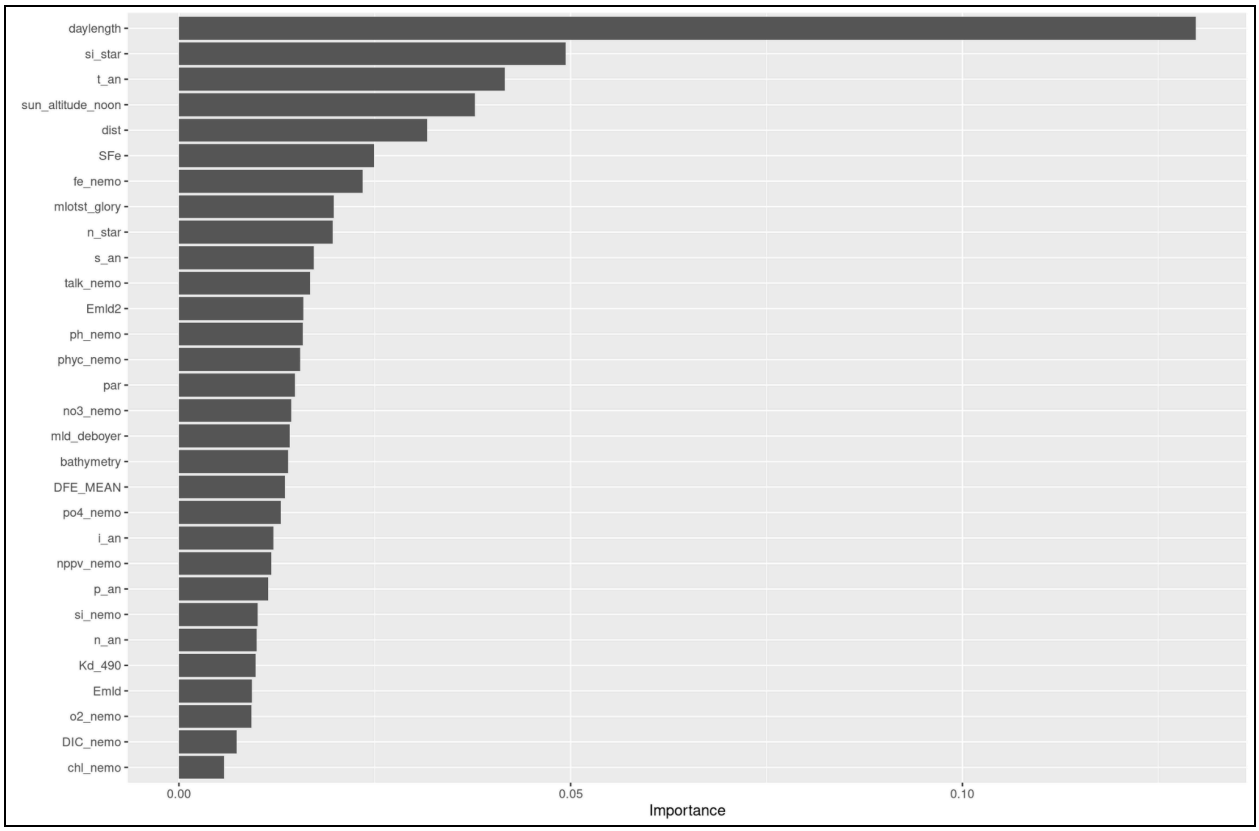

Figure 5. Variable importance in the XGBoost model.
Variable importance for the predictor variables used in the training model, calculated using a permutation based method, where each variable had its values permuted across all observations. This keeps the distribution of that predictor, but breaks its relationship with the response. A recomputation of the performance metric is then done; in the case of this project, that metric is R^2. The new value is compared to the previous value, and if the new model is significantly worse, then the variable importance is greater. In other words, the difference in R^2 between the permuted model and original model is proportional to variable importance.

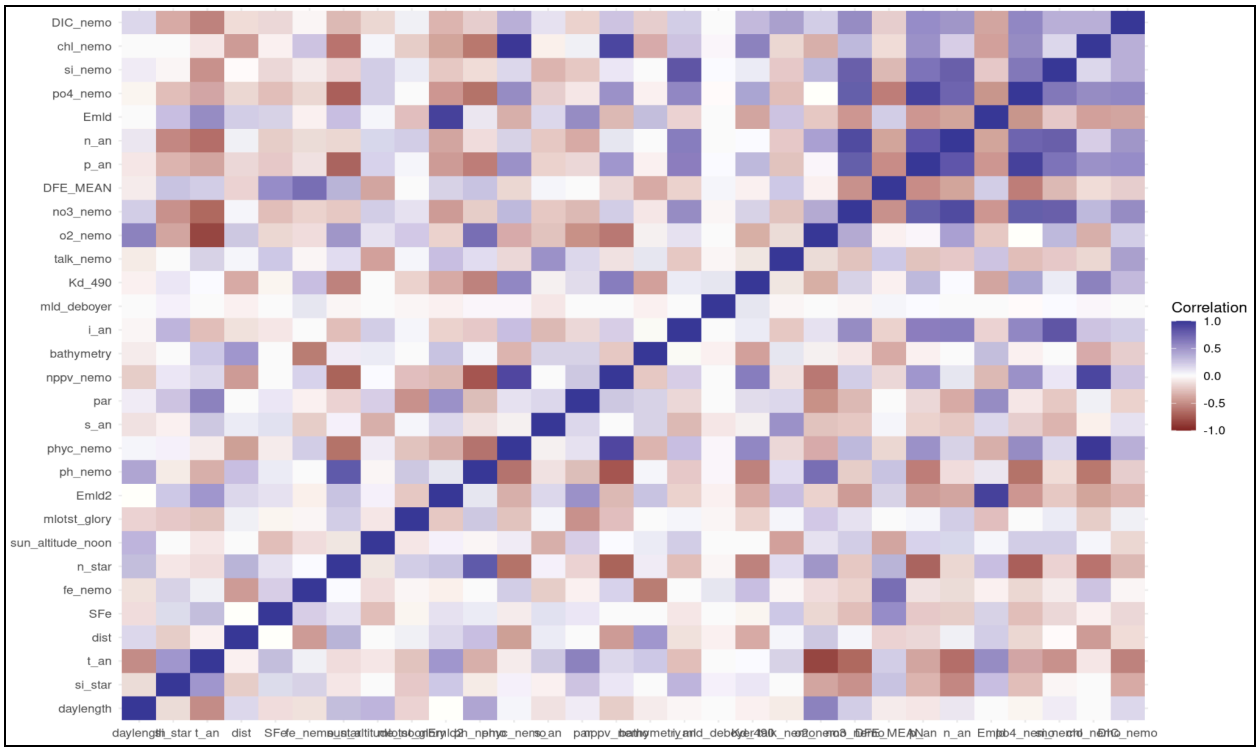

Figure 6. Correlation Heatmap.

The correlation between each predictor variable, and is ranked in order of importance based on the permutation-based importance calculation. Nitrate, phosphate, and silicate are all highly correlated with each other across both the WOA and NEMO model, while environmental factors such as day length, the altitude of the Sun at noon, and temperature are not particularly correlated with other variables, positively or negatively. Most of the most important variables are not highly correlated with anything else, indicating that the number of predictor variables can be decreased without losing much information in the model.

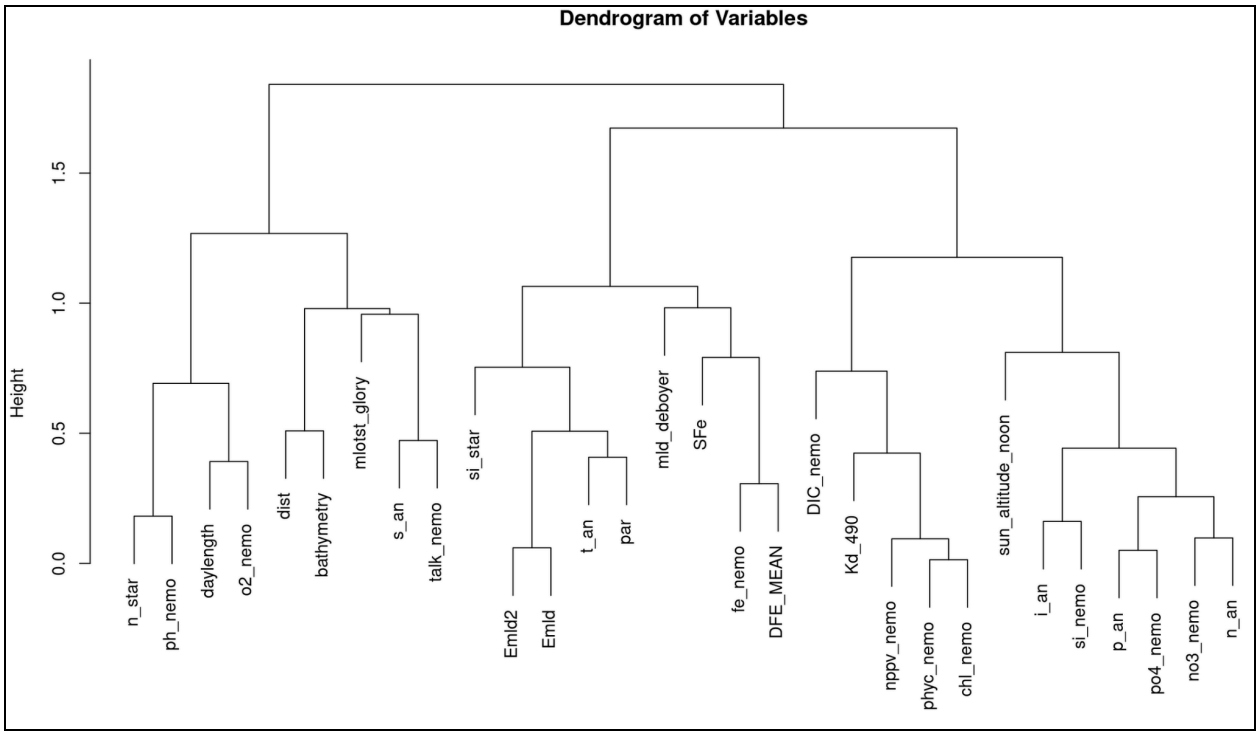

Figure 7. Dendrogram.

Dendrogram of predictor variables shows the correlation distance between each predictor variable. Similarly to the correlation heatmap, this shows the high correlation between nitrate, phosphate, and silicate, while environmental variables show little correlation with anything else.

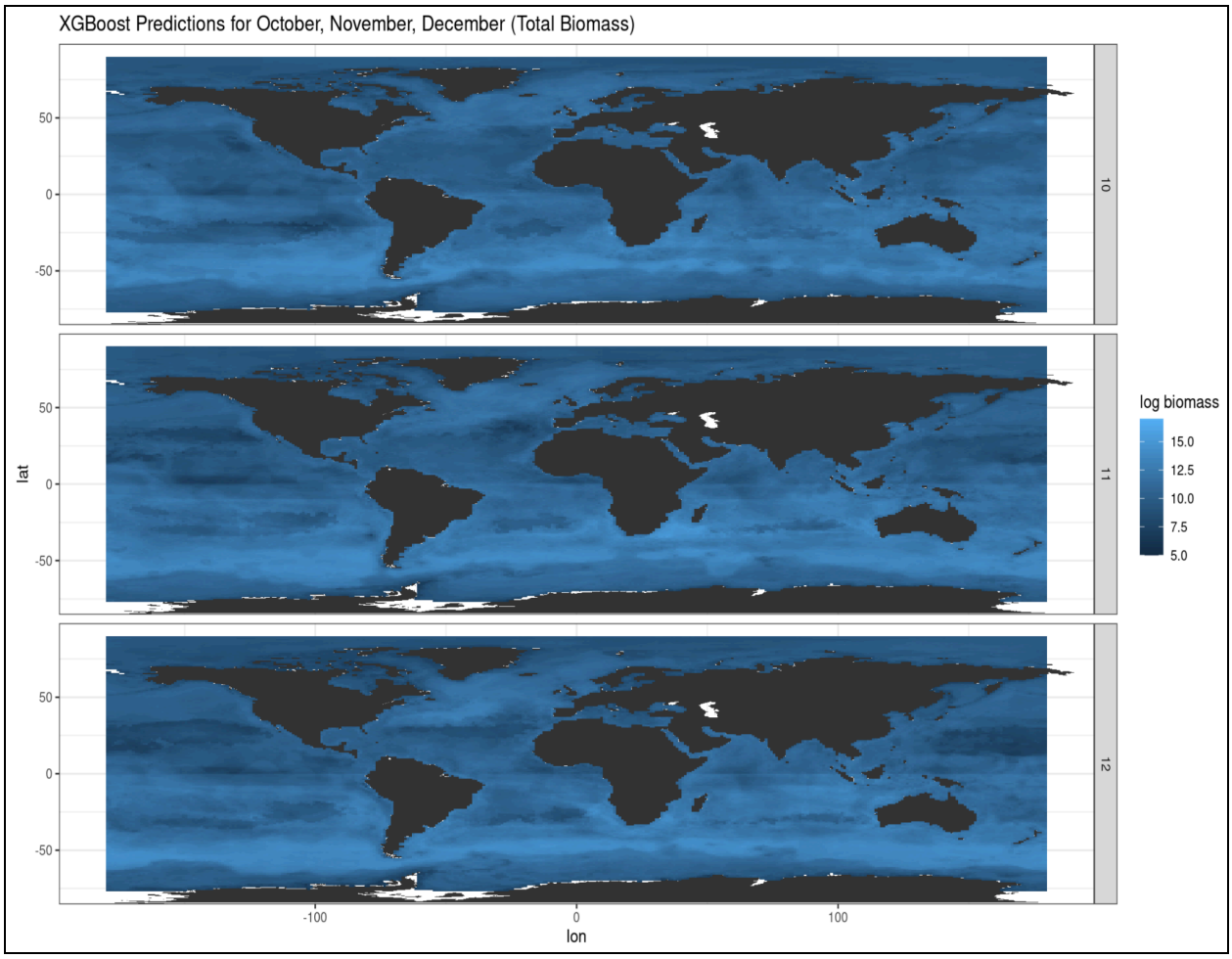

Figure 8. Northern autumn total coccolithophore biomass.
Total coccolithophore biomass (log scale) predictions for October, November, December using XGBoost model.

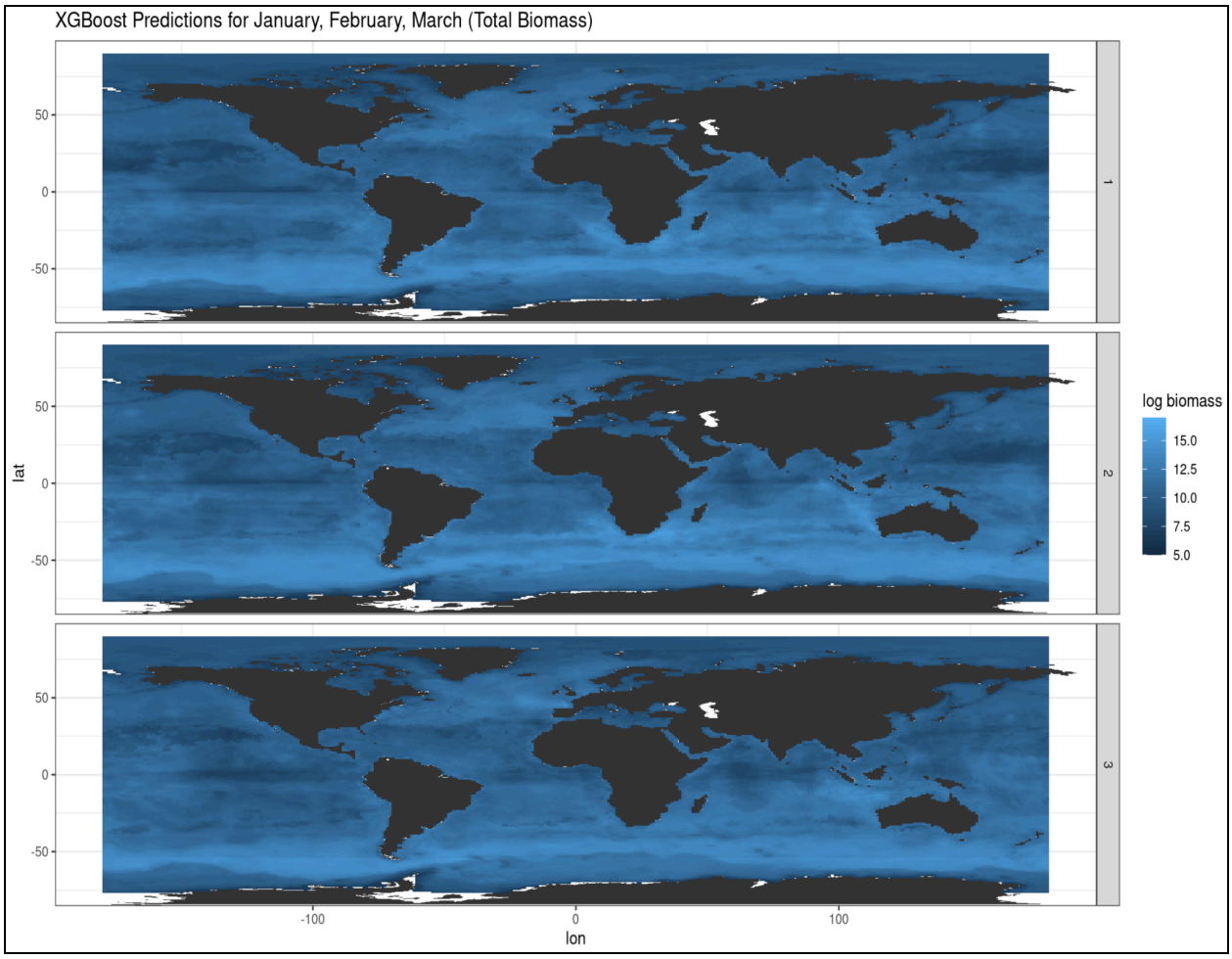

Figure 9. Northern winter total coccolithophore biomass.
Total coccolithophore biomass (log scale) predictions for January, February, March using XGBoost model.

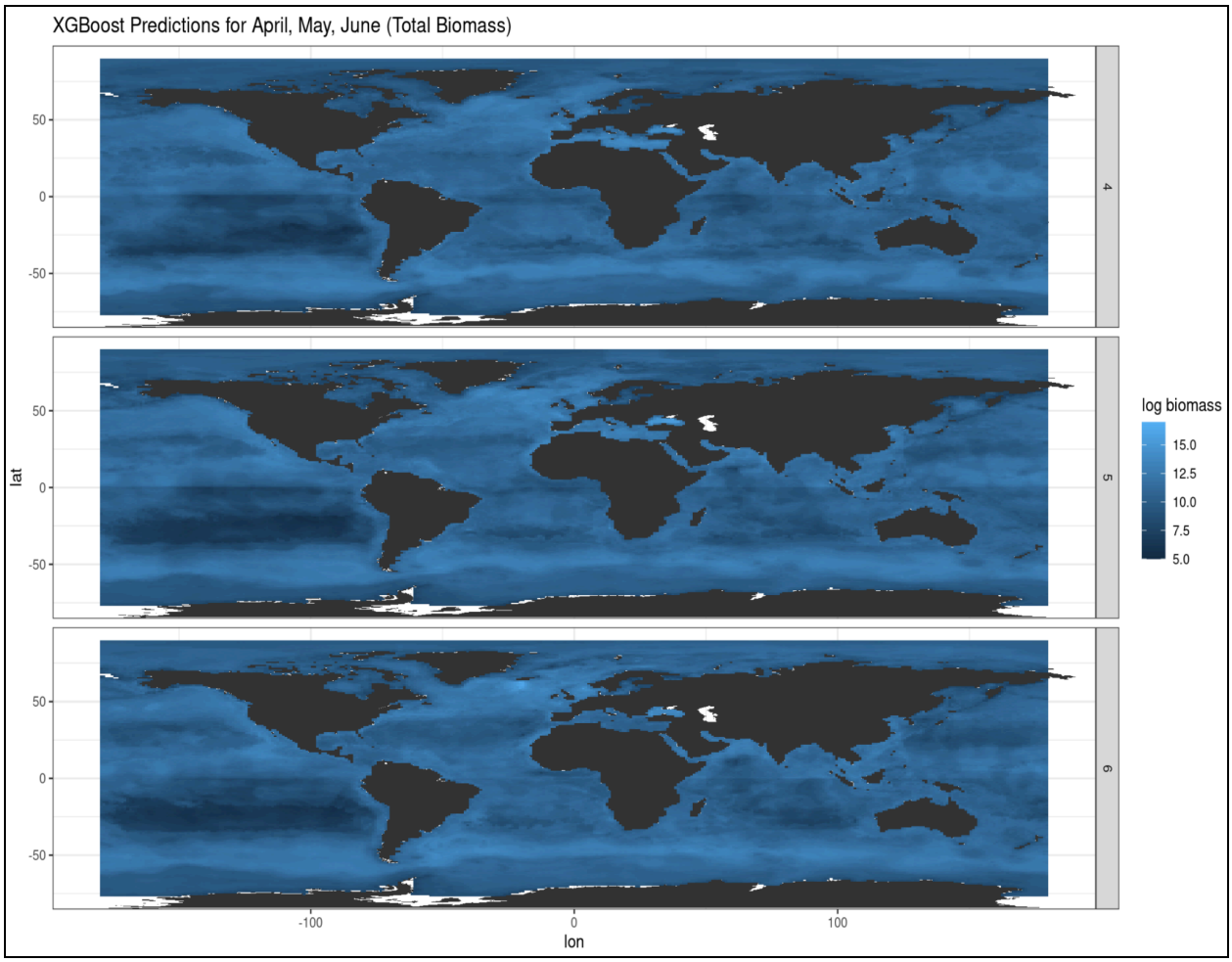

Figure 10. Northern spring total coccolithophore biomass.
Total coccolithophore biomass (log scale) predictions for April, May, June using XGBoost model.

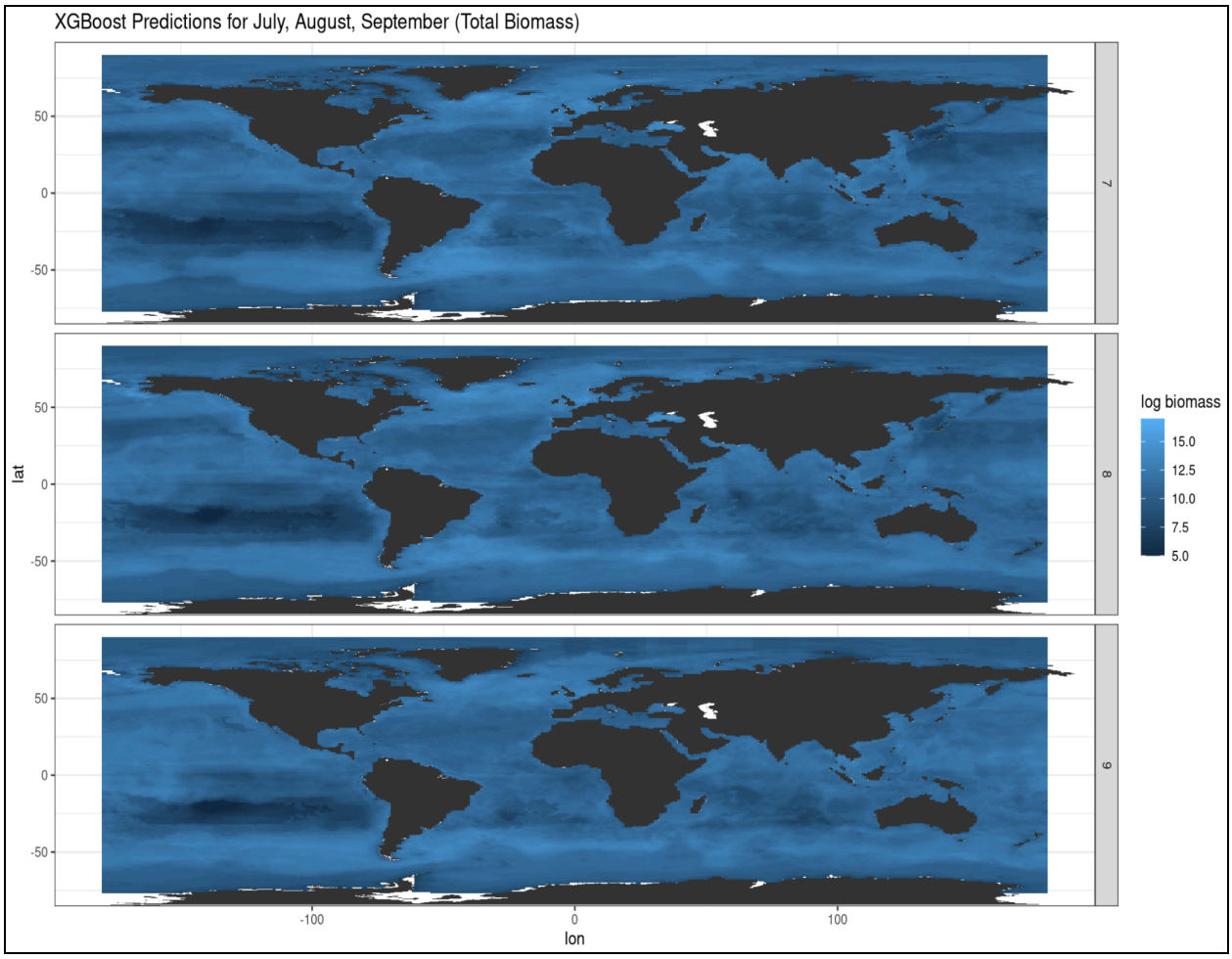

Figure 11. Northern summer total coccolithophore biomass.
Total coccolithophore biomass (log scale) predictions for July, August, September using XGBoost model.

Table 1. The environmental variables used in the development of the predictive models.
All predictor variables in the training dataset. These variables were joined with the log_biomass observations to form the training dataset for the model. Month, latitude, longitude, and pixel area were all disregarded when training the model, so that predictions would be based solely on environmental factors.

Variable Name	Description	Units	Source
month	Month the observation was recorded	N/A	Reagan et al., 2024
lat	Latitude of the observation, rounded to the nearest degree	Degrees	Reagan et al., 2024
lon	Longitude of the observation, rounded to the nearest degree	Degrees	Reagan et al., 2024
SFe	Soluble iron at location and month of observation	nmol L^{-1}	Tagliabue et al, 2016
sun_altitude_noon	Altitude of the sun at noon at location and month of observation	TBD	Reagan et al., 2024
daylength	Sunlight time at location and month of observation	Hours	Reagan et al., 2024
DFE_MEAN	Dissolved iron at location and month of observation	nmol L^{-1}	Tagliabue et al, 2016
bathymetry	Depth of ocean at location	m	Reagan et al., 2024
Kd_490	Diffuse light attenuation coefficient at 490 nm	m^{-1}	SeaWiFS satellite
par	Photosynthetically available radiation	µmol quanta m^{-2} s^{-1}	SeaWiFS satellite
t_an	Objectively analysed mean temperature at location and month of observation	℃	Reagan et al., 2024
s_an	Objectively analysed mean salinity at location and month of observation	N/A	Reagan et al., 2024
n_an	Objectively analysed mean nitrate at location and month of observation	µmol m^{-3}	Reagan et al., 2024
p_an	Objectively analysed mean phosphate at location and month of observation	µmol m^{-3}	Reagan et al., 2024
i_an	Objectively analysed mean silicate at location and month of observation	µmol m^{-3}	Reagan et al., 2024
no3_nemo	Nitrate at location and month of observation	µmol m^{-3}	NEMO project

po4_nemo	Phosphate at location and month of observation	µmol m^{-3}	NEMO project
si_nemo	Silicate at location and month of observation	µmol m^{-3}	NEMO project
fe_nemo	Iron at location and month of observation	µmol m^{-3}	NEMO project
chl_nemo	Chlorophyll at location and month of observation	µg m^{-3}	NEMO project
nppv_nemo	Net primary product at location and month of observation	µg C L^{-1} d^{-1}	NEMO project
phyc_nemo	Phytoplankton carbon at location and month of observation	mg C L^{-1}	NEMO project
o2_nemo	Oxygen dissolved at location and month of observation	mg C L^{-1}	NEMO project
talk_nemo	Total alkalinity at location and month of observation	µmol kg^{-1}	NEMO project
DIC_nemo	Dissolved inorganic carbon at location and month of observation	µmol kg^{-1}	NEMO project
ph_nemo	pH at location and month of observation	N/A	NEMO project
dist	Distance of location from land	km	TBD
mlotst_glory	Mixed layer depth	m	Glory model
mld_deboyer	Mixed layer depth	m	de Boyer Montégut et al., 2004
si_star	Silicate - Nitrate	µmol m^{-3}	i_an - n_an
n_star	Nitrate - 16*Phosphate	µmol m^{-3}	n_an - 16*p_an
Emld	Ratio with MLD Glory	N/A	par / (Kd_490*mlotst_glory)
Emld2	Ratio with MLD deBoyer	N/A	par / (Kd_490*mld_deboyer)
pixel_area	Area of 1°x1° block	km^2	N/A

Table 2. Monthly worldwide coccolithophore observations
The total number of coccolithophore observations in each month.

Month	Number of Observations
January	12,808
February	34,948
March	15,433
April	14,564
May	16,864
June	11,156
July	25,973
August	15,885
September	25,748
October	26,990
November	9,556
December	3,746

Table 3. Model predictive performance

R^2 for each model for each coccolithophore species abundance

Coccolithophore Species	Model Type	R^2
Total	Normalised Neural Network	0.54
Total	Random Forest	0.58
Total	Recursively partitioned decision tree	0.44
Total	Extreme Gradient Boosted Tree	0.59
Total	Generalised Linear Model with Elastic Net Regularisation	0.22
Coccolithus pelagicus	Normalised Neural Network	0.71
Coccolithus pelagicus	Random Forest	0.73
Coccolithus pelagicus	Recursively partitioned decision tree	0.52
Coccolithus pelagicus	Extreme Gradient Boosted Tree	0.72
Coccolithus pelagicus	Generalised Linear Model with Elastic Net Regularisation	0.52
Discosphaera tubifera	Normalised Neural Network	0.36
Discosphaera tubifera	Random Forest	0.48
Discosphaera tubifera	Recursively partitioned decision tree	0.36
Discosphaera tubifera	Extreme Gradient Boosted Tree	0.51
Discosphaera tubifera	Generalised Linear Model with Elastic Net Regularisation	0.24
Emiliania huxleyi	Normalised Neural Network	0.61
Emiliania huxleyi	Random Forest	0.64
Emiliania huxleyi	Recursively partitioned decision tree	0.54
Emiliania huxleyi	Extreme Gradient Boosted Tree	0.66
Emiliania huxleyi	Generalised Linear Model with Elastic Net Regularisation	0.39
Ophiaster hydroideus	Normalised Neural Network	0.28
Ophiaster hydroideus	Random Forest	0.35
Ophiaster hydroideus	Recursively partitioned decision tree	0.18
Ophiaster hydroideus	Extreme Gradient Boosted Tree	0.28
Ophiaster hydroideus	Generalised Linear Model with Elastic Net Regularisation	0.33

Data availability

Data were obtained from the NEMO community ocean model (https://www.nemo-ocean.eu/), the GLORY ocean data reanalysis using the CMEMS eddy-resolving model (doi 10.48670/moi-00021), the SeaWiFS project (https://oceancolor.gsfc.nasa.gov/), and the Aqua/MODIS satellite (https://oceancolor.gsfc.nasa.gov/)

References

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32.

Charalampopoulou, A., Poulton, A. J., Bakker, D. C. E., Lucas, M. I., Stinchcombe, M. C., & Tyrrell, T. (2016). Environmental drivers of coccolithophore abundance and calcification across Drake Passage (Southern Ocean). *Biogeosciences, 13*, 5917–5935.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)* (pp. 785–794). Association for Computing Machinery.

de Boyer Montégut, C., Madec, G., Fischer, A. S., Lazar, A., & Iudicone, D. (2004). Mixed layer depth over the global ocean: An examination of profile data and a profile-based climatology. *Journal of Geophysical Research, 109*, C12003.

Friedman, J. H., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software, 33*(1), 1–22.

Greenwell, B. M., & Boehmke, B. C. (2020). *Variable importance plots—An introduction to the R package vip*. *The R Journal, 12*(1), 343–366.

Holligan, P. M., Groom, S. B., & Harbour, D. S. (1993). What controls the distribution of the coccolithophore *Emiliania huxleyi* in the North Sea? *Fisheries Oceanography, 2*(3–4), 175–183.

Kuhn, M., & Silge, T. (2023). *Tidy modeling with R: A consistent framework for building machine learning models in R*. O'Reilly Media.

Leblanc, K., Hare, C. E., Feng, Y., Berg, G. M., DiTullio, G. R., Neeley, A., Benner, I., Sprengel, C., Beck, A., Sanudo-Wilhelmy, S. A., Passow, U., Klinck, K., Rowe, J. M., Wilhelm, S. W., Brown, C. W., & Hutchins, D. A. (2009). Distribution of calcifying and silicifying phytoplankton in relation to environmental and biogeochemical parameters during the late stages of the 2005 North East Atlantic Spring Bloom. *Biogeosciences, 6*, 2155–2179.

Naud, M. C., Sheward, R. M., Irwin, A. J., & Finkel, Z. V. (2025). Coccobase: an updated compilation of coccolithophore field observations. *Unpublished manuscript.*

Nissen, C., Vogt, M., Münnich, M., Gruber, N., & Haumann, F. A. (2018). Factors controlling coccolithophore biogeography in the Southern Ocean. *Biogeosciences, 15*, 6997–7024.

O'Brien, C. J., Peloquin, J. A., Vogt, M., Heinle, M., Gruber, N., Ajani, P., Andruleit, H., Arístegui, J., Beaufort, L., Estrada, M., Karentz, D., Kopczyńska, E., Lee, R., Poulton, A. J., Pritchard, T., & Widdicombe, C. (2013). Global marine plankton functional type biomass distributions: Coccolithophores. *Earth System Science Data, 5*, 259–276.

Pebesma, E., & Graeler, B. (2025). *gstat: Spatial and spatio-temporal geostatistical modelling, prediction and simulation* (Version 2.1-4) [Computer software manual]. NOAA National Centers for Environmental Information.

Reagan, J. R., Boyer, T. P., García, H. E., Locarnini, R. A., Baranova, O. K., Bouchard, C., Cross, S. L., Mishonov, A. V., Paver, C. R., Seidov, D., Wang, Z., & Dukhovskoy, D. (2024). *World Ocean Atlas 2023*. NOAA National Centers for Environmental Information.

Tagliabue, A., Aumont, O., Death, R., Dunne, J. P., Dutkiewicz, S., Galbraith, E., Misumi, K., Moore, J. K., Ridgwell, A., Sherman, E., Stock, C., Vichi, M., Völker, C., & Yool, A. (2016). How well do global ocean biogeochemistry models simulate dissolved iron distributions? *Global Biogeochemical Cycles*, *30*(2), 149-174.

Taylor, A. R., Brownlee, C., & Wheeler, G. (2017). Coccolithophore cell biology: Chalking up progress. *Annual Review of Marine Science, 9*, 283–310.

Therneau, T. M., & Atkinson, E. J. (1997). An introduction to recursive partitioning using the RPART routines. Mayo Foundation.

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S* (4th ed.). Springer.

Winter, A., & Siesser, W. G. (Eds.). (1994). *Coccolithophores*. Cambridge University Press.

Wright, M. N., & Ziegler, A. (2017). ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software, 77*(1), 1–17.