Honours Thesis

Modelling Coccolithophore data in the North Atlantic Ocean

Seoyeon (Cali) Park

B00768397

Fall 2021

Contents

1 Abstract

This thesis builds species models of coccolithophores using hierarchical generalized additive models (HGAM) and regression analysis using Least Absolute Shrinkage and Selection Operator (LASSO). Analysis will include many common model comparison techniques such as comparisons of adjusted R^2 and Akaiake Information Criterion scores of the training data, and Root Mean Squared Errors of both training and testing data, to examine which models can accurately predict species abundance. We start by extracting a subset of the Coccobase data repository and apply a series of models and analysis on corresponding results. Three of the best HGAMs were chosen based on comparison techniques mentioned above. The best LASSO model assessed was for the *Emiliania huxleyi* data. These components will be interpreted and discussed in depth.

2 Acknowledgements

I would like to express my gratefulness to all individuals who have contributed to aid in writing this honours thesis. I would like to acknowledge Dr. Andrew Irwin, whom without his support and patience would not have been possible. Meeting every week for discussions on room to improve and how to approach issues that have arisen during this process have been a huge support. Without Dr. Irwin, I would not have been able to gain knowledge and experience with applied statistical research with oceanography data.

I would like to thank the lab members, especially Joe, whom have initially taught me the background of LASSO regression, and Mohammad, always answering questions about statistics, both very kind and knowledgeable in their research. Another thanks to Crispin, who also suggested different approaches to summarizing the HGAM smooths, when I was stuck with a problem.

Special thanks to the professors that have taught me statistics throughout my undergraduate degree. I would like to thank each and every person that have taught me the world of statistics. Having a solid foundation from their shared knowledge have allowed me to want to pursue a further Master's degree in Statistics.

3 Introduction

Coccolithophores are a major type of calcifying phytoplankton commonly found across all oceans that occupy distinct niches from tropical to sub-polar seas. As climatic conditions change across the oceans from various reasons including global warming, it is important to understand how species abundance will change, and to develop species distribution models to predict how their future biogeography may change as a consequence of anticipated changes in climate. Many previous studies, using controlled laboratory experiments, show that variables such as light, nutrients, and temperature play an important role in determining interspecific competition and species abundance. Analysis of species abundance and its relation to environmental and physical variables using real data collected across the world's oceans is the next necessary step in accurately predicting species abundances and finding similarities and/or differences in the environmental conditions where coccolithophore species are observed, and is the main objective of this work.

Coccobase is a compilation of previously reported data that include both environmental and physical observations of coccolithophore species across the globe as well as their respective abundances [6]. Coccobase therefore provides an excellent opportunity to model coccolithophore abundances in the North Atlantic using real ocean data, which may in the future be extrapolated to the rest of the Earth's oceans.

Species models for *Emiliania huxleyi*, *Coccolithus pelagicus*, *Syracosphaera pulchra* are compared to models for the broad category of all coccolithophore observations, and the genera *Coccolithus* and *Syracosphaera*. These species were chosen based on the number of observations available. *E. huxleyi, C. pelagicus, S. pulchra* were the species with the highest number of observations after subsetting the data (Table 11).

The purpose of this research was to determine which environmental variables can be used to accurately predict species abundance, and to find similarities and/or differences in the environmental conditions where the coccolithophore species are observed, using appropriate modelling methods. The data was modelled using hierarchical generalized additive models (HGAMs) and GLMs with LASSO variable selection and regularization [3, 9]. These models were used to assess the differences in niche and to compare the predictive ability of the different modelling techniques. The best model for predictive accuracy and fit of data were determined based on common model comparison methods, using values of adjusted R^2 (R-sq adj), Akaike information criterion scores (AIC), and root mean squared errors (RMSE) [1].

4 Methods

4.1 Data collection and manipulation

The data used for this analysis comes from the Coccobase repository, available on Github [4]. Coccobase is built upon previously made compilations of data and includes many of the variables required for analysis. Coccobase has over 213,000 individual observations, carefully filtered by applying quality control measures to the pre-existing data gathered from the literature and earlier databases. There are 190 different coccolithophore taxa reported in Coccobase, but for this report, a subset of the data will be used [6].

The data included all observations of coccolithophores in the North Atlantic. (Table 11) The models include observations of coccolithophores, of the genera *Coccolithus* and *Syracosphaera*, as well as species specific observations of *Emiliania huxleyi*, *Coccolithus pelagicus*, and *Syracosphaera pulchra*. All the species-specific observations were subsetted based on the highest number of observations. The following table describes the number of observations in data corresponding to each species. ea

Table 1: Species and corresponding number of observations

Family/Species name	Number of observations
Coccolithophore	5823
Coccolithus species	803
Syracosphaera species	790
Emiliania huxleyi	1367
Coccolithus pelagicus	572
Syracosphaera pulchra	240

The variables used to predict species abundance were: phosphate, Si^*, N^*, salinity, temperature, and a discrete classifier for each species. Phosphate, is an essential element for nucleic acid, RNA/DNA synthesis of coccolithophores. [7] The higher the phosphate concentration, the more beneficial it would be for the species. Si^* is the difference between silicate and nitrate concentration, and N^* is the difference between nitrate and 16 times the phosphate concentration. Both differences were chosen because they show how species abundance

and species competition are correlated to the availability of environmental variables. Temperature is known to affect the abundance of coccolithophore species, hence it is sufficient to include it as one of the predictor variables [10]. Finally, salinity is a predictor because it influences strain specific growth. These variables will require an addition of the tensor product between longitude and latitude in the North Atlantic Ocean. Observations for species abundance was log transformed to follow an approximate normal distribution, since the histogram of the original scale appeared to be positively skewed.

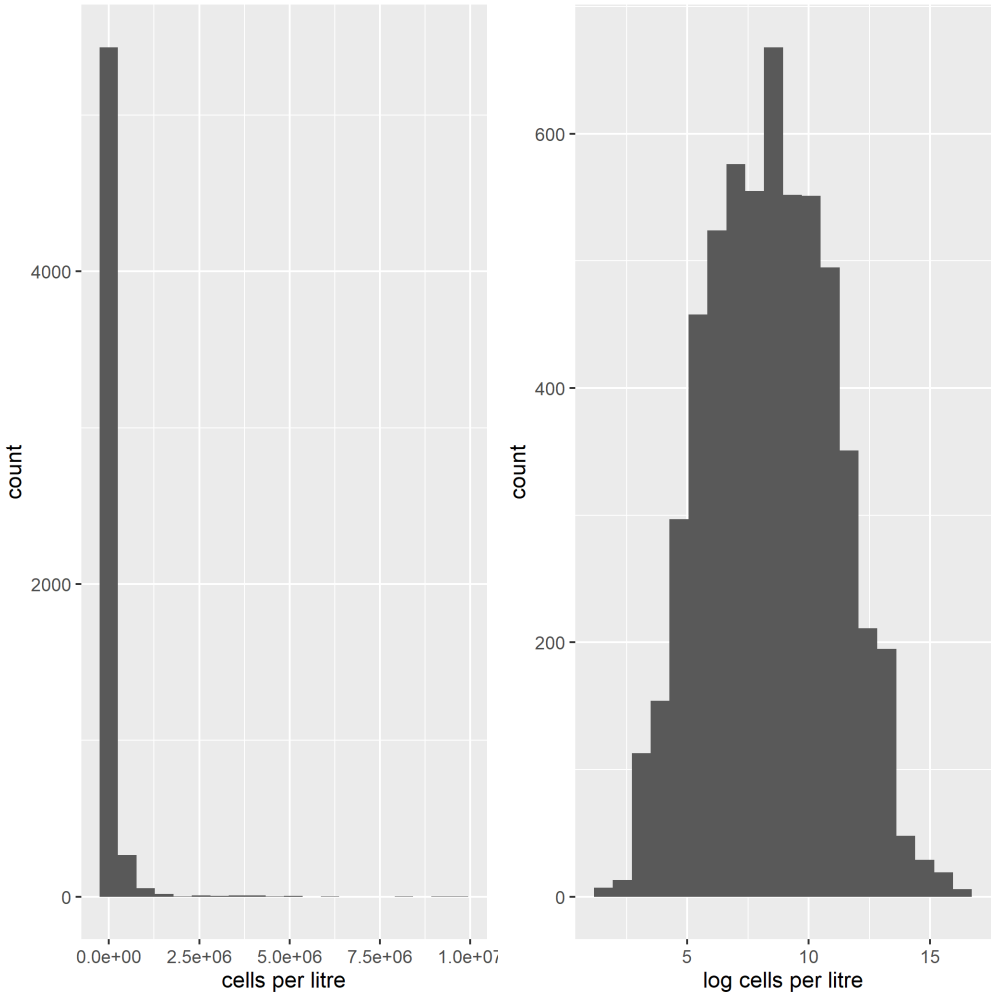

Figure 1: Histogram of species abundance counts. Left: Original scale. Right: Log transformed.

4.2 Brief introduction of generalized additive models (GAM) and hierarchical generalized linear models (HGLM)

Two modelling techniques, Generalized Additive Models (GAM) and Hierarchical Generalized Linear Models (HGLM), have been commonly used to model highly variable data. Certain aspects of these models are useful to construct the Hierarchical Generalized Additive Models, which was one of the two modelling techniques used in this study.

GAMs are very flexible models, where the only assumptions are that the functions are smooth (via splines) and that the response is an additive function of the smooths of each predictor. They have the following structure:

$$E[Y] = g^{-1}(\beta_0 + \sum_{i=1}^{I} f_i(x_i)) \tag{1}$$

where E(Y) is the expected value of the response Y with the corresponding link function g, and $f_i(x_i)$ is the sum of all predictors smoothed by the function fi with intercept β_0. The response is related upon the linear combination of smoothed $(s_i(x_i), i = 1, .., I)$ predictor variables, which usually allows for better fit of the data [3].

One of the advantages of using GAMs is that they are highly suitable for fitting non-linear data. As mentioned above, there are many ways to smooth the predictor variables via splines. Types of splines include Thin Plate Regression Splines (TPRS) and Cyclic Cubic Regression Splines (CRS), which should be chosen carefully based on the nature of the predictor variables.

Table 2: Description of "smoothers" used in analysis.

Smoothers	Description
Thin Plate Regression Splines (TPRS)	Used to smooth covariates in any number of dimensions
Cyclic Cubic Regression Splines (CRS)	Useful for fitting covariates that undergo cyclic effects (seasonal)
Random Effects (RE)	Smooths covariates without any assumptions on the structure.

In the analysis, phosphate, N^*, Si^*, salinity, temperature, and latitude used this smoother TPRS (although the temperature variable may suffice using the CRS smoother), where as longitude and scientific name used

the smoothers CRS and RE, respectively.

Another commonly used modelling technique, are HGLMs (or also called generalized linear mixed models) which are a flexible extension of generalized linear models (GLMs) [5]. They meet the assumptions of a generalized linear model with an additional component that allows the model to have hierarchical (structural or multi-level) relationships between the predictors, and as well have an overall relationship between the predictors and the response. This hierarchical structure leads to building more realistic models and reduces overfitting [2]. The assumptions of these modelling techniques will be beneficial to keep in mind for the development of the HGAM, explained in the next section.

4.3 Hierarchical Generalized Additive Models (HGAM)

Components of HGLM and GAM can be combined to make hierarchical generalized additive models (HGAM). HGAMs are constructed to model highly variable data, to understand the smooth hierarchical functional relationships between predictor groups and the response. The modelling efforts of this report closely followed the methods outlined in Pederson et al [9]. For this paper, only models G, GS, and GI are performed for analysis of coccolithophores in the North Atlantic Ocean. Model G are fitted on datasets including all coccolithophore observations, and datasets including the *Coccolithus* and *Syracosphaera* species. Model GI and GS are fitted on species specific data subsets. All models are fitted using the restricted (residual) maximum likelihood method.

Description of each model is summarized in the following table:

Table 3: Types of Models in Analysis

Type of Model	Description
Model G	A single global smoother for all observations
Model GS	A single global smoother plus group-level smoothers that have the same wiggliness
Model GI	A single global smoother plus group-level smoothers that have different wiggliness

Each model has the following structure coded into R:

- Model G: `gam(cells_per_litre = s(p_an, bs = "tp") + s(si_star, bs = "tp") + s(ni_star, bs = "tp") + s(s_an, bs = "tp") + s(t_an, bs = "tp") + s(scientific_name, bs = "re") + te(longitude, latitude, bs = c("cc","tp")), data)`

- Model GS (phosphate): `gam(cells_per_litre  s(p_an, bs = "tp") + s(si_star, bs = "tp") + s(p_an, bs = "tp") + s(s_an, bs = "tp") + s(t_an, bs = "tp") + te(longitude, latitude, bs = c("cc", "tp")) + s(p_an, scientific_name, bs = "fs"), data)`

- Model GI (phosphate): `gam(cells_per_litre  s(p_an, bs = "tp") + s(si_star, bs = "tp") + s(p_an, bs = "tp", m = 1) + s(s_an, bs = "tp") + s(t_an, bs = "tp") + te(longitude, latitude, bs = c("cc", "tp")) + s(p_an, scientific_name, bs = "fs"), data)`

4.4 Least Absolute Shrinkage and Selection Operator (LASSO)

The Least Absolute Shrinkage and Selection Operator (LASSO) can be incorporated into linear models to shrink some estimated parameters to 0. This allows for automatic variable selection and regularization for building simpler and more interpretable models.

Suppose the data is composed of N total observations. Under the assumption that each observation is independent (the same assumptions as ordinary least squares, OLS), the outcome of the data is arranged into an $n \times 1$ vector:

$$y_i = \begin{bmatrix} y_1 & y_2 & ... & y_n \end{bmatrix}^\top$$

where $i = 1, ..., n$.

Each predictor, x_i, is arranged into a $n \times p$ data matrix of observations:

$$x_{np} = \begin{bmatrix} x_{11} & x_{12} & ... & x_{1p} \\ x_{21} & x_{22} & ... & x_{2p} \\ ... & ... & ... & ... \\ x_{n1} & x_{n2} & ... & x_{np} \end{bmatrix}$$

where n is number of predictor variables in the data, and p is the total number of observations corresponding to each predictor variable.

In addition, suppose that there is an $n \times 1$ vector of parameters for each predictor β_n:

$$\hat{\beta} = \begin{bmatrix} \beta_1 & \beta_2 & ... & \beta_n \end{bmatrix}^\top$$

The standard linear model is:

$$y = X\beta + \epsilon$$

LASSO regression then becomes a modified least squares problem:

$$\hat{\beta_\lambda} = argmin_b \left\| Y - X\beta \right\|_2^2 + \lambda \left\| \beta \right\|_1$$

where $\left\| \beta \right\|_1 = \sum_{i=1} |\beta_i|$. The addition of the $\lambda \left\| \beta \right\|_1$ term adds regularization to the least squares problem, by adding a penalty to the minimization of $\left\| Y - X\beta \right\|_2^2$.

In addition, we see that for any $\lambda > 0$, there exists an $s_\lambda = \left\| \beta_\lambda \right\|_1$ where $\hat{\beta_\lambda} = argmin_b \left\| Y - X\beta \right\|_2^2$ such that $||\beta_1|| \leq s_\lambda$ [11].

4.5 Software

All data is manipulated and modelled using pre-existing R packages. Refer to the appendix for a full list of packages used. The coding was performed with R/RStudio. The full R project can be accessed via Github at: **https://github.com/calipark1213/Honours-Coccobase** [8].

4.6 Training and Test data

Each of the species subsets were further partitioned into training and testing data. To allow both sets to have a good representation of data and to avoid overfitting, total data were split randomly into 60% training and 40% test data. The number of observations outlined in table 1 are further split into the following table:

Table 4: Training and Testing set

Genus/Species name	Train	Test
Coccolithophore	3493	2329
Coccolithus species	482	321
Syracosphaera species	474	316
Emiliania huxleyi	820	547
Coccolithus pelagicus	343	229
Syracosphaera pulchra	144	96

4.7 Model Comparison Methods

Common model comparison techniques was used to analyze the differences between HGAM and LASSO regression models. The R-squared adjusted values and AIC values of the training data will be compared. R-squared adjusted is a measure of how much the variation of the full data is measured by the model, whereas AIC values evaluates how well a model fits the data it was generated from. Furthermore, the RMSE values of both training and testing data will be compared, since it measures the standard deviation of the residuals. For LASSO, models will be compared using variable selection and RMSE values. RMSE values were calculated on a log transformed scale, so the model values should be interpreted by taking the exponent of the RMSE value, and be interpreted as a factor out on average from the true values in the original scale. By taking the value on a log scale, the values are less sensitive to large outliers.

5 Results

5.1 HGAMS

5.1.1 R-squared adjusted values

As outlined in table 5, for models G and GS, the highest R-squared adjusted value corresponded to the dataset that included all of the *Coccolithus* genus. The lowest value of R-squared adjusted for model G

corresponded to the species specific dataset for *Syracosphaera pulchra*, with a value of 0.497 (or deviating by a factor of $e^{0.497} = 1.6438$ on the original scale). The second lowest R-squared adjusted value was for the dataset that included all of the species of coccolithophores in the North Atlantic.

For models GS and GI, the group-level smoothers with same/different wiggliness, respectively, did not differ by a significant amount, for species *Emiliania huxleyi* and *Coccolithus pelagicus*. For model GI, the highest R-squared value was for the model with the individual penalty on the salinity variable for the *Coccolithus pelagicus* species.

Table 5: Adjusted R-squared values of all models (Training data)

Model G					
Coccolithophore	0.591				
Coccolithus species	0.718				
Syracosphaera species	0.641				
Emiliania huxleyi	0.643				
Coccolithus pelagicus	0.674				
Syracosphaera pulchra	0.497				
Model GS	Phosphate	Si*	N*	Salinity	Temperature
Emiliania huxleyi	0.633	0.633	0.633	0.633	0.633
Coccolithus pelagicus	0.639	0.639	0.639	0.639	0.639
Syracosphaera pulchra	0.492	0.492	0.492	0.492	0.492
Model GI	Phosphate	Si*	N*	Salinity	Temperature
Emiliania huxleyi	0.634	0.633	0.639	0.633	0.633
Coccolithus pelagicus	0.639	0.639	0.637	0.65	0.641
Syracosphaera pulchra	0.492	0.492	0.505	0.492	0.496

5.1.2 AIC values

The best models out of all the models computed for model G, GS, and GI, were the models that included the training data corresponding to the species *Emiliania huxleyi*. As outlined in table 6, the lowest AIC scores corresponded to the *Emiliania huxleyi* models. For model G, this value was -591.8112, and for model GS, the lowest AIC value corresponded to the group level smooth of salinity with *Emiliania huxleyi* as the species.

Lastly, for model GI, the lowest AIC value corresponded to the group level smooth of of N* with *Emiliania huxleyi* as the species.

Table 6: AIC values of all models (Training data)

Model G					
Coccolithophore	-425.736				
Coccolithus species	-175.147				
Syracosphaera species	-351.161				
Emiliania huxleyi	-591.811				
Coccolithus pelagicus	-90.676				
Syracosphaera pulchra	-81.999				
Model GS	Phosphate	Si*	N*	Salinity	Temperature
Emiliania huxleyi	-570.196	-569.632	-571.776	-571.850	-569.631
Coccolithus pelagicus	-62.0486	-62.049	-62.049	-62.049	-62.049
Syracosphaera pulchra	-77.033	-77.032	-77.034	-77.032	-77.033
Model GI	Phosphate	Si*	N*	Salinity	Temperature
Emiliania huxleyi	-573.891	-569.844	-587.358	-571.846	-570.704
Coccolithus pelagicus	-62.049	-62.049	-59.2310	-48.448	-65.229
Syracosphaera pulchra	-75.423	-77.033	-77.759	-77.033	-77.173

5.1.3 RMSE values

Finally, to measure how well the fitted training models will predict in general, the RMSE of the predictions based on the test data were calculated. The RMSE values shared the same conclusion as the AIC values. The best models out of all the models computed for model G, GS, and GI, were the models that included the testing data corresponding to the species *Emiliania huxleyi*. As outlined in table 7, the lowest RMSE scores for the *Emiliania huxleyi* models corresponded to the smallest AIC value. For model G, this value was 0.18261, and for model GS, the lowest AIC value corresponded to the group level smooth of phosphate with *Emiliania huxleyi* as the species. Same as the conclusion for the AIC values, for model GI, the lowest RMSE value corresponded to the group level smooth of of N* with *Emiliania huxleyi* as the species.

Table 7: RMSE values of all models (Training data, Test data)

Model G					
Coccolithophore	0.220 0.227				
Coccolithus species	0.183 0.246				
Syracosphaera species	0.149 0.296				
Emiliania huxleyi	0.159 0.183				
Coccolithus pelagicus	0.186 0.226				
Syracosphaera pulchra	0.164 0.228				
Model GS	Phosphate	Si*	N*	Salinity	Temperature
Emiliania huxleyi	0.163 0.182	0.163 0.182	0.163 0.182	0.163 0.182	0.163 0.182
Coccolithus pelagicus	0.201 0.216	0.201 0.216	0.201 0.216	0.201 0.216	0.201 0.216
Syracosphaera pulchra	0.165 0.221	0.165 0.221	0.165 0.226	0.165 0.221	0.165 0.221
Model GI	Phosphate	Si*	N*	Salinity	Temperature
Emiliania huxleyi	0.162 0.181	0.163 0.182	0.161 0.180	0.163 0.182	0.162 0.181
Coccolithus pelagicus	0.201 0.216	0.201 0.214	0.202 0.214	0.197 0.214	0.201 0.217
Syracosphaera pulchra	0.164 0.197	0.165 0.221	0.161 0.164	0.165 0.221	0.164 0.187

The top 3 models based on the R-squared adjusted, AIC, and RMSE value were: Model G of *Coccolithus* species, Model GS with the group-level smooth of salinity for *Emiliania huxleyi*, and model GI with the group-level smooth of N* for *Emiliania huxleyi*.

5.2 LASSO regression

The same subsetted data were fitted using a different modelling technique, using LASSO. The hopes were that the best models based on HGAMs would also be the best choice using LASSO regression. Using the glmnet package from R, the LASSO models were fitted using the best lambda value calculated by the lambda.min function. From there, the coefficients of the model estimates were extracted. The top 3 LASSO models will also be chosen based on variable selection and RMSE values. The results are outlined below.

5.2.1 Coefficients of terms

LASSO models using *Coccolithus* species, *Emiliania huxleyi*, *Coccolithus pelagicus*, and the full data shrunk one of the environmental variables to 0. This is seen in table 8, where the models for *Coccolithus* and full data have shrunk the salinity term to 0, and the models for *Emiliania huxleyi* and *Coccolithus pelagicus* data

have shrunk the scientific name term to 0. The model for *Syracosphaera pulchra* have shrunk 3 terms to 0 with the coefficients being closest to 0, but with sample size taken into account this model should not be included in the final selection. For each LASSO model, the lambda.min function was used to extract the best lambda value. The following plot shows Log(λ) against the Mean-Squared Error component. Note that (1) - (6) is the plot for *Coccolithophore*, *Coccolithus* species, *Syracosphaera* species, *Emiliania huxleyi*, *Coccolithus pelagicus*, and *Syracosphaera pulchra*, respectively.

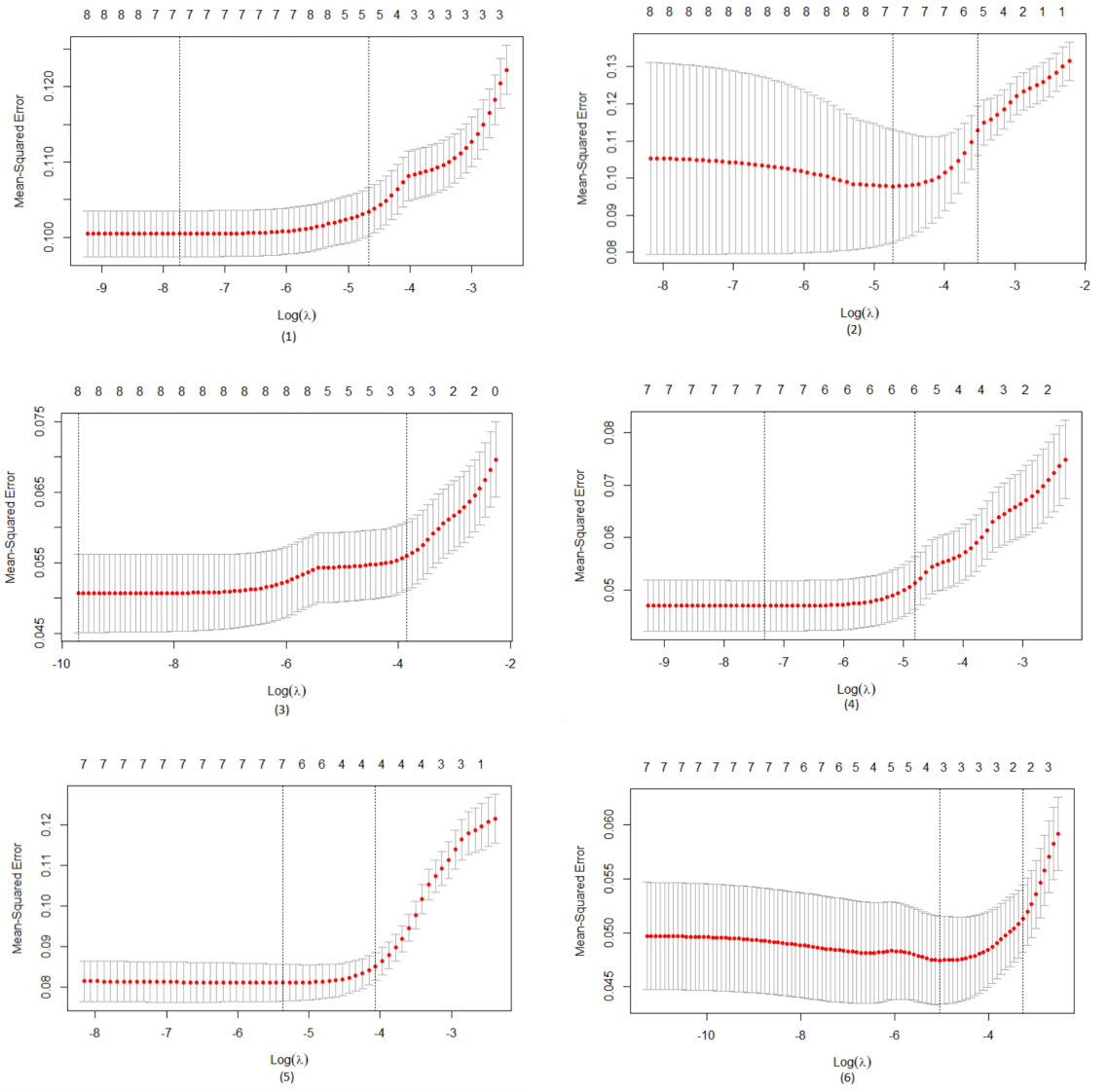

Figure 2: Plot of Log(λ) against Mean-Squared Error corresponding to each LASSO model.

Table 8: Table of coefficients via LASSO regression (Training data)

	Coccolithophore	*Coccolithus* species	*Syracosphaera* species	*Emiliania huxleyi*	*Coccolithus pelagicus*	*Syracosphaera pulchra*
Intercept	0.486	-2.493	-26.848	-7.833	-1.722	-1.833E1
Phosphate	-0.487	-0.276	-1.606	-1.231	-0.097	.
Si*	-0.087	-0.073	-0.172	-0.108	-0.065	-8.699E-2
N*	-0.039	0.012	-0.159	-0.088	0.014	-1.609E-1
Salinity	.	.	0.781	0.268	-0.141	5.572E-1
Temperature	0.039	0.069	-0.024	-0.005	0.097	2.395E-5
Scientific name	-0.003	0.010	-0.001	.	.	.
Longitude	0.004	-0.013	-0.019	0.009	-0.012	.-7.292E-3
Latitude	0.022	0.055	0.021	0.018	0.065	.

Table 9: Minimum lambda value corresponding to each model fitted with training data

Family/Species name	Minimum lambda value
Coccolithophore	3.9716E-4
Coccolithus species	9.616E-3
Syracosphaera species	1.2768E-4
Emiliania huxleyi	9.386E-5
Coccolithus pelagicus	7.276E-4
Syracosphaera pulchra	1.341E-3

Since the models were fitted using the cv.glmnet function, the minimum lambda (or best lambda) value were calculated for each of the regression models. The above table shows the values corresponding to each species data.

5.2.2 RMSE values

The following table summarizes the RMSE values of each test data fitted by LASSO regression:

Table 10: RMSE values of all LASSO models (Training data, Test data)

Family/Species name	RMSE
Coccolithophore	0.316 0.313
Coccolithus species	0.287 0.308
Syracosphaera species	0.218 0.245
Emiliania huxleyi	0.213 0.223
Coccolithus pelagicus	0.275 0.288
Syracosphaera pulchra	0.207 0.205

From this table, we can see that the lowest RMSE value corresponds to the LASSO regression model fitted using the *Syracosphaera pulchra* data, although one should note that due to the small number of observations and the LASSO regression model shrinking most of the coefficients to zero, the next lowest RMSE value should be chosen. Therefore, the best model via LASSO regression is the model fitted onto the *Emiliania huxleyi* data.

Combining the conclusions from HGAMs and LASSO regression, the best models based on the comparison techniques outlined above were all using the *Emiliania huxleyi* data. Actual vs predicted plots of the best models are shown below:

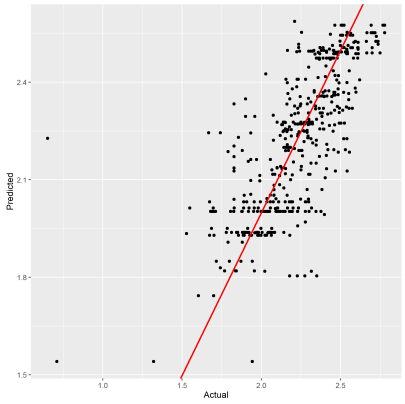

Figure 3: Actual VS Predicted model of *Emiliania huxleyi* data. Model GS with group-level smooth on salinity.

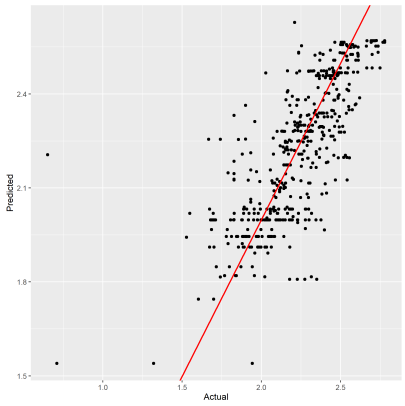

Figure 4: Actual VS Predicted model of *Emiliania huxleyi* data. Model GI with group-level smooth on N*.

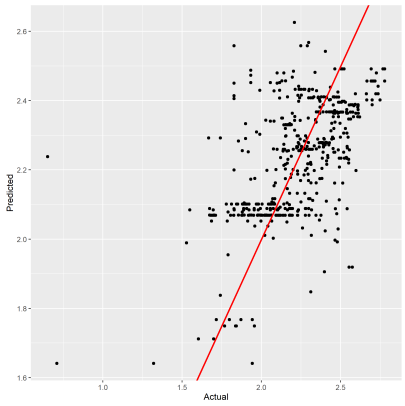

Figure 5: Actual VS Predicted model of *Emiliania huxleyi* data. LASSO model.

From the actual vs predicted plots, majority of the points are clustered around the one-to-one line. With an exception of a few outliers, the overall relationships in all three plots show a good fit.

6 Discussion and conclusions

In this project, methods using hierarchical generalized additive models as well as LASSO regression were used to build species models in the North Atlantic ocean using R/Rstudio. Initially, the data were compiled using the Coccobase data repository, and subsetted using the longitude and latitude, and further partitioned into the 6 different species specific data groups using the highest number of observations.

The best models selected for the HGAMs were model G using the *Coccolithus* species data, and model GS

with group smoother on salinity, model GI with group smoother on N^* using the data for *Emiliania huxleyi*. Using these results, the species distribution models can be further interpreted and summarized using the weighted mean and variances of each of the plotted smooths shown in figures 12, 13, and 14 (Appendix). If an individual were to require a more general interpretation, a weighted average of the individual models' predictions can be used, which would imply model averaging rather than model selection.

The best models using LASSO regression was the model using the data for *Emiliania huxleyi*. Although the LASSO coefficients for the *Syracosphaera pulchra* data shrunk two terms, phosphate and latitude to 0, this model was exempted from the model selection by taking into account the sample size for these species. LASSO regression was mainly useful for variable selection. For example, the salinity term for coccolithophore annd *Coccolithus* species shrunk to 0, indicating that the term could be removed for future analysis. Although I was expecting the correlated variables such as phosphate, nitrate, and silicate to be shrunk to 0, this was not the case for this analysis, as the models were not particularly sparse. Unfortunately, model selection for LASSO was not based on AIC, but it would be useful to calculate some numerical value for comparisons. In the future, it would be beneficial to determine the variables for prediction using LASSO regression first, then running the HGAMs for faster computational time, as well as determining the differences in species niche.

To improve the models, it may be beneficial to take into account the times of each observations, and build a time-series model of the distribution of species. This would be another interesting approach to determine if or how the species distributions change over time. A sensitivity analysis should be conducted given that some species data (e.g. *Syracosphaera pulchra*) used had a small number of observations in comparison to other species (e.g. *Emiliania huxleyi*). This will be beneficial, as number of observations may influence the overall fit of the models.

References

[1] H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19:716–723, 1974.

[2] Andrew Gelman. Multilevel (hierarchical) modeling: What it can and cannot do. *Technometrics*, 2006.

[3] Antoine Guisan, Thomas C Edwards, and Trevor Hastie. Generalized linear and generalized additive models in studies of species distributions: Setting the scene. *Ecological Modelling*, 2002.

[4] Andrew Irwin and Seoyeon Park. Marine microbial macroecology, 2021.

[5] Y. Lee and J. A. Nelder. Hierarchical generalized linear models. *Journal of the Royal Statistical Society: Series B (Methodological)*, 1996.

[6] M. Naud, R. Sheward, A. J. Irwin, and Z. V. Finkel. Coccobase: an updated compilation of coccolithophore field observations. Pre-print, 2020.

[7] Colleen J. O'Brien, Meike Vogt, and Nicolas Gruber. Global coccolithophore diversity: Drivers and future change. *Progress in Oceanography*, page 27–42, 2016.

[8] Seoyeon Park. Honours - coccobase.

[9] Eric J. Pedersen, David L. Miller, Gavin L. Simpson, and Noam Ross. Hierarchical generalized additive models in ecology: An introduction with mgcv. *PeerJ*, 2019.

[10] Dionysios E. Raitsos, Samantha J. Lavender, Yaswant Pradhan, Toby Tyrrell, Philip C. Reid, and Martin Edwards. Coccolithophore bloom size variation in response to the regional environment of the subarctic north atlantic. *Limnology and Oceanography*, 2006.

[11] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58:267–288, 1996.

7 Appendix

Table 11: Full species list and corresponding number of observations

Species names	Number of observations	Species names	Number of observations	Species names	Number of observations
Acanthoica acanthos	32	*Coccolithophore*	1064	*Helicosphaera wallichii*	4
Acanthoica quattrospina	47	*Coccolithus pelagicus*	572	*Helladosphaera cornifera*	105
Acanthoica spp.	35	*Coccolithus* spp.	230	*Homozygosphaera spinosa*	7
Algirosphaera robusta	58	*Coccosphaera* spp.	39	*Lohmannosphaera* spp.	2
Alisphaera gaudii	1	*Corisphaera gracilis*	8	*Michaelsarsia adriaticus*	4
Alisphaera unicornis	18	*Corisphaera* spp.	13	*Michaelsarsia elegans*	13
Anthosphaera spp.	7	*Coronosphaera mediterranea*	150	*Michaelsarsia* spp.	12
Braarudosphaera spp.	1	*Cyrtosphaera aculeata*	6	*Oolithotus fragilis*	25
Calcidiscus leptoporus	43	*Discosphaera tubifera*	11	*Ophiaster formosus*	12
Calcidiscus leptoporus/Coccolithus pelagicus	1	*Emiliania huxleyi*	1367	*Ophiaster hydroideus*	171
Calcidiscus spp.	57	*Emiliania huxleyi/Gephyrocapsa* spp.	1	*Ophiaster* spp.	5
Calciopappus caudatus	116	*Florisphaera profunda*	3	*Palusphaera vandelii*	9
Calciopappus rigidus	2	*Gephyrocapsa ericsonii*	19	*Pappomonas* spp.	1
Calciopappus spp.	10	*Gephyrocapsa muellerae*	34	*Papposphaera* spp.	4
Calciosolenia brasiliensis	3	*Gephyrocapsa oceanica*	3	*Pontosphaera syracusana*	3
Calyptrolithina divergens	2	*Gephyrocapsa ornata*	1	*Reticulofenestra parvula*	4
Calyptrosphaera sphaeroidea	90	*Gephyrocapsa* spp.	49	*Reticulofenestra* spp.	1
Calyptrosphaera spp.	14	*Gladiolithus flabellatus*	1	*Rhabdosphaera clavigera*	51
Canistrolithus spp.	1	*Helicosphaera carteri*	114	*Rhabdosphaera hispida*	6
Coccolithaceae	281	*Helicosphaera* spp.	1	*Rhabdosphaera* spp.	7

Species names	Number of observations	Species names	Number of observations
Rhabdosphaera xiphos	8	*Syracosphaera* spp.	226
Scyphosphaera apsteinii	5	*Syracosphaera tumularis*	16
Syracosphaera anthos	12	*Thoracosphaera heimii*	7
Syracosphaera bannockii	13	*Umbellosphaera* spp.	7
Syracosphaera borealis	9	*Umbellosphaera tenuis*	22
Syracosphaera corolla	15	*Umbilicosphaera sibogae*	32
Syracosphaera coronata	4		
Syracosphaera dentata	18		
Syracosphaera dilatata	14		
Syracosphaera halldalii	15		
Syracosphaera histrica	9		
Syracosphaera marginiporata	18		
Syracosphaera molischii	97		
Syracosphaera nana	1		
Syracosphaera nodosa	11		
Syracosphaera noroitica	5		
Syracosphaera ossa	14		
Syracosphaera prolongata	8		
Syracosphaera pulchra	240		
Syracosphaera rotula	45		

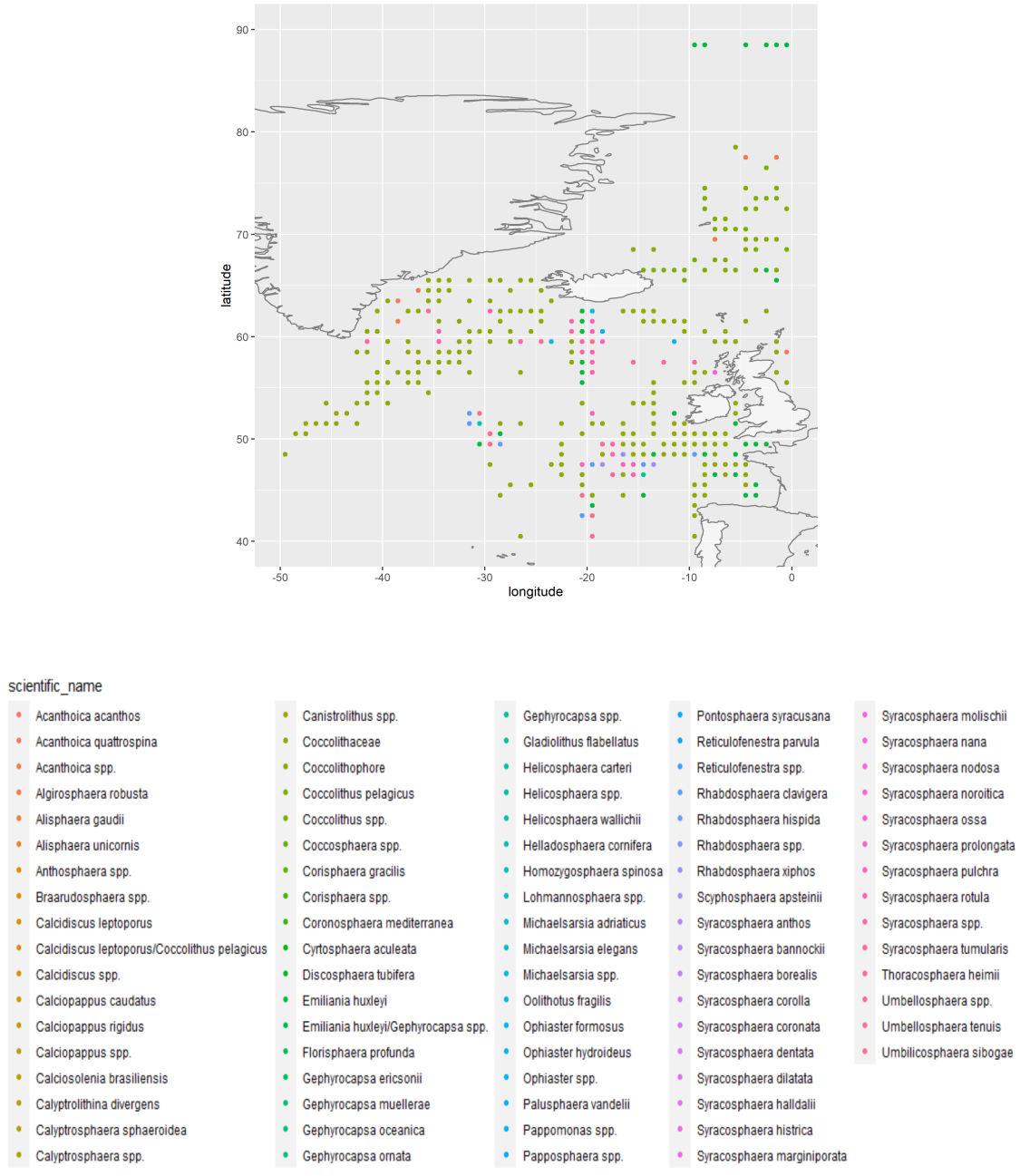

Figure 6: Distribution of all species in full data

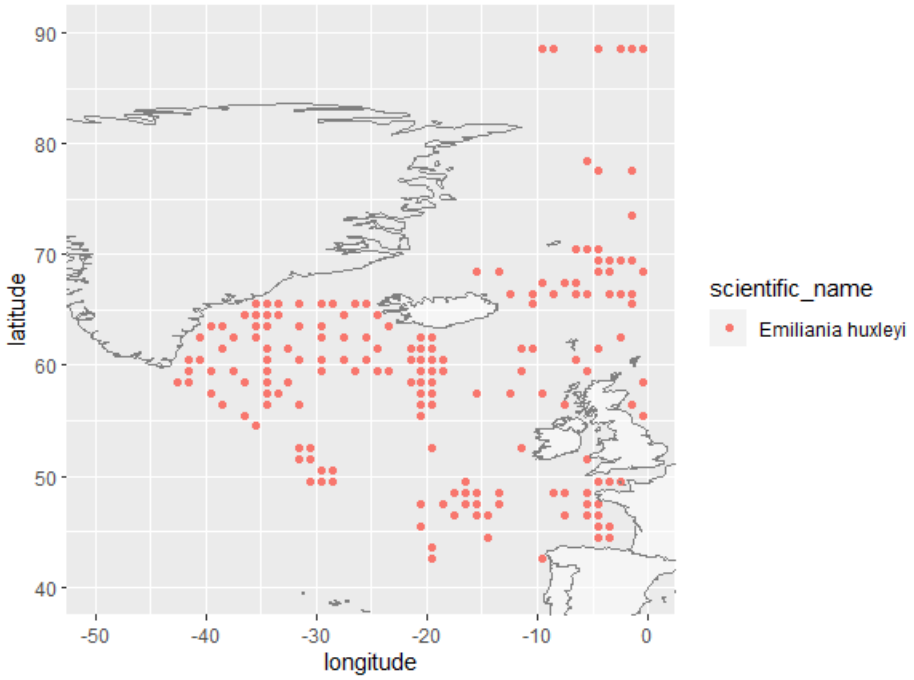

Figure 7: Distribution of *Emiliania huxleyi*

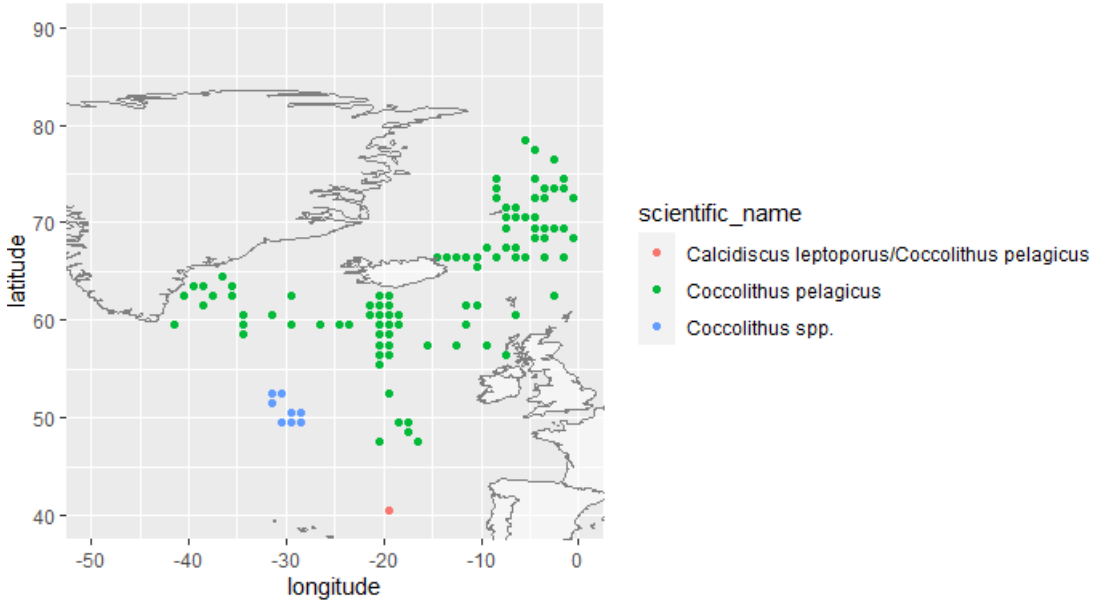

Figure 8: Distribution of *Coccolithus* species

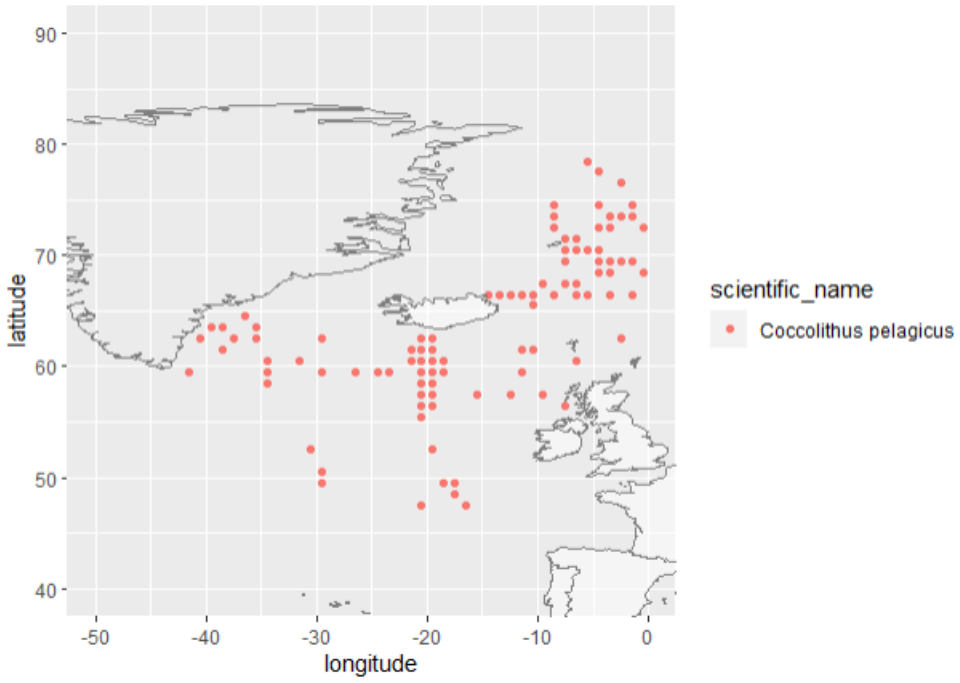

Figure 9: Distribution of *Coccolithus pelagicus*

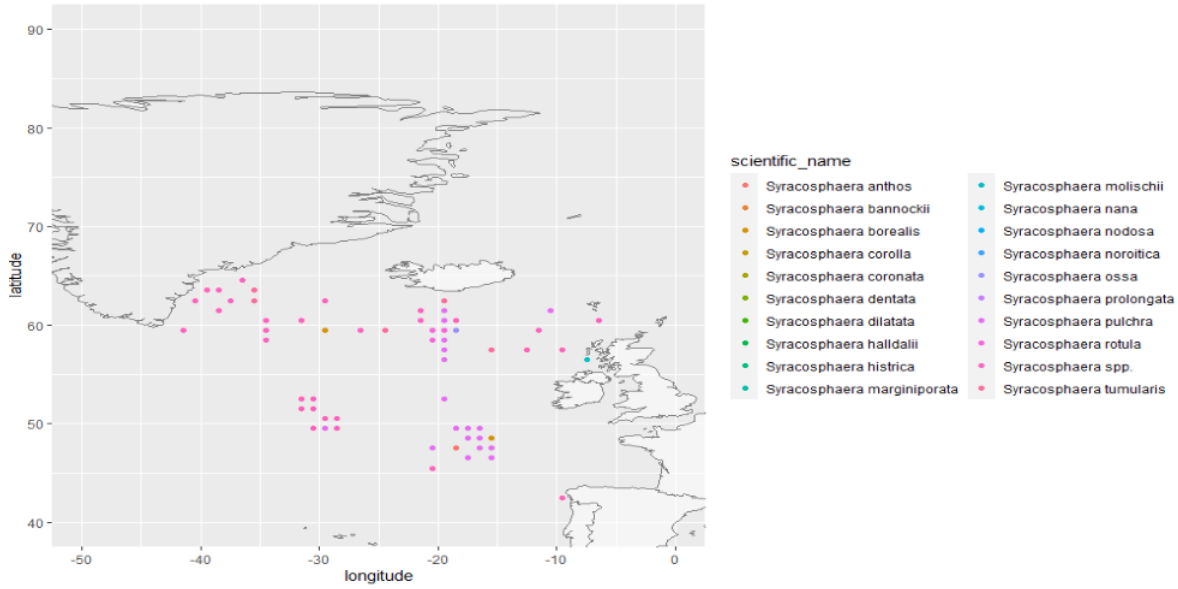

Figure 10: Distribution of *Syracosphaera* species

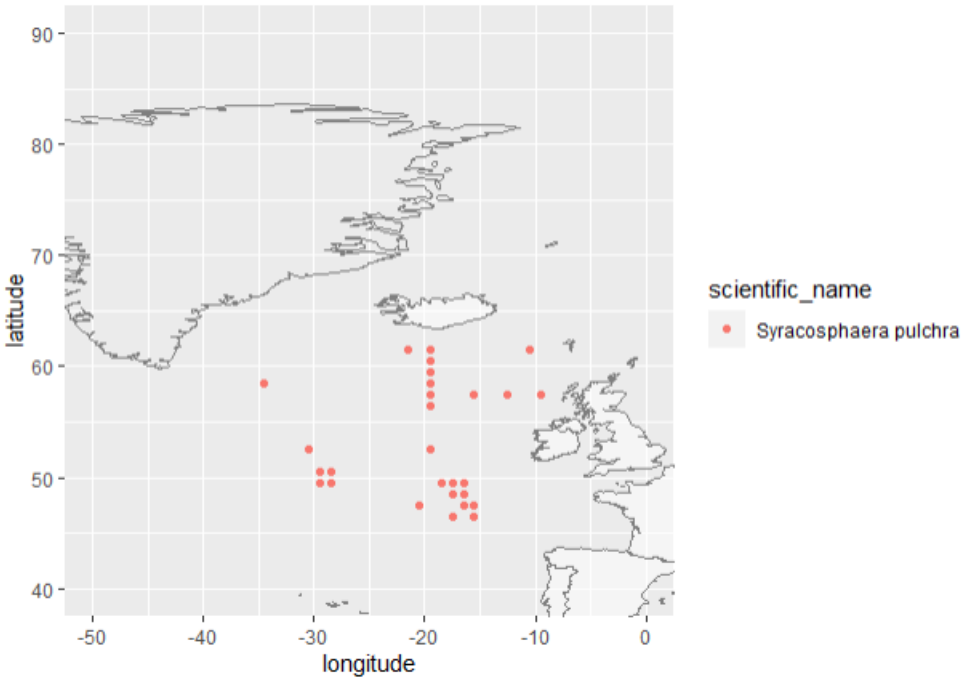

Figure 11: Distribution of *Syracosphaera pulchra*

Next pages will include some GAM plot outputs of training data using "seWithMean = T". These GAM plots correspond to the models that were selected to be the "best" models based on AIC and RMSE values.

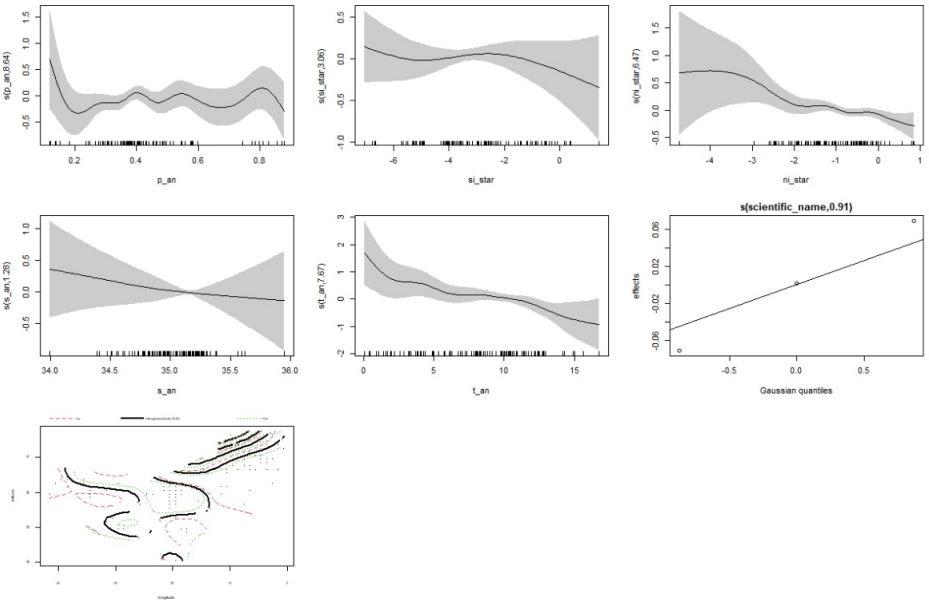

Figure 12: *Coccolithus* species data. Model G

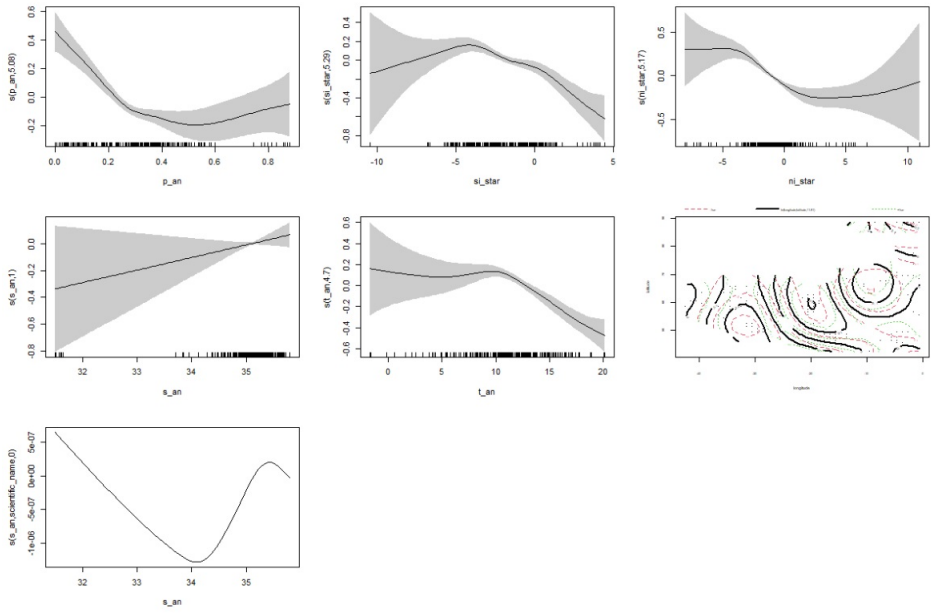

Figure 13: *Emiliania huxleyi* data. Model GS with specific smooth on salinity.

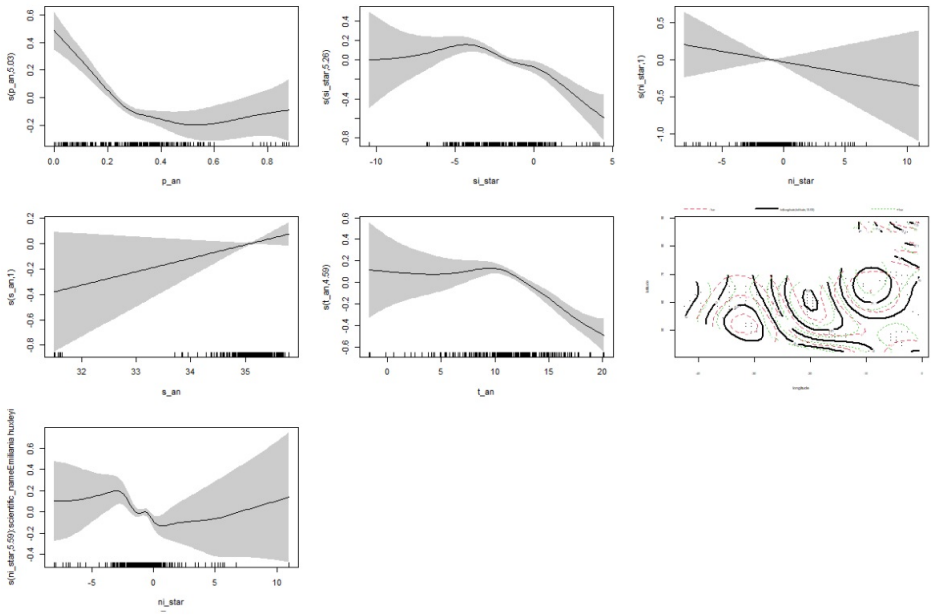

Figure 14: *Emiliania huxleyi* data. Model GI with specific smooth on N*.

All code used to create plots and complete analyses can be found at: **https://github.com/calipark1213/Honours-Coccobase**