Distribution and Elemental Composition of Picoplankton in the North Pacific Ocean

by

Jonathan Bradet-Legris

Submitted in partial fulfillment of the requirements
for the degree of Master of Science

at

Dalhousie University
Halifax, Nova Scotia
January 2021

Contents

List of Tables

List of Figures

Abstract

Marine picoplankton account for a considerable portion of primary production in the Ocean, particularly in the oligotrophic regions. Picoplankton are dominated by three major groups: *Prochlorococcus*, *Synechococcus* and picoeukaryotes, whose populations are controlled by many complex interacting factors. We analyze a rich dataset from a cruise in the North Pacific Ocean to model the distribution and elemental composition of each picoplankton group using only environmental data as predictors. Linear regression, generalized additive models, and random forests were used to make models of phytoplankton abundances and carbon biomasses. Elemental composition for each phytoplankton group was modeled using Bayesian linear regression by regressing elemental C, N, and P concentrations on picoplankton biovolumes. Our species distribution models show temperature and salinity are consistently the most important predictors to explain variation in abundance and biomass. Along the full transect, nutrient concentrations (PO_4, Fe, Mn, Cu) provide useful insights on sharp population shifts over short distances and our results support the claim that iron and phosphorus are limiting nutrients in the North Pacific Ocean's oligotrophic gyre. Our elemental quota model show that the composition and cellular C:P, N:P and C:N ratios varies substantially among the three picoplankton groups. Average carbon content for *Prochlorococcus*, *Synechococcus* and picoeukaryotes were 167 $fg\ C/\mu m$ (95% CR: $6.89-457.8$), 538 $fg\ C/\mu m$ (95% CR: $307.2-771.5$), and 297 $fg\ C/\mu m$ (CR: $13.74-804.6$) respectively. This model provides a method for estimating individual elemental content of each phytoplankton group, which is otherwise unmeasurable directly from field samples.

1 Introduction

Primary production in marine picoplankton ($\lesssim 2\mu m$ in diameter) is dominated by three phytoplankton functional groups: *Prochlorococcus*, *Synechococcus* and picoeukaryotes. Flombaum et al. [2013] estimates that *Prochlorococcus* and *Synechococcus* are responsible for 8.5% and 16.7% of global net primary production, respectively. The prokaryotes *Prochlorococcus* and *Synechococcus* are photosynthetic marine cyanobacteria that often dominate oligotrophic regions of the ocean (Ting et al. [2002]). *Prochlorococcus* are very small, with diameters typically in ranges $0.5\mu m - 0.8\mu m$, and dominate in waters with low nutrient concentrations (Bertilsson et al. [2003]). These features make them exceptionally well suited for the low nutrient conditions of the tropical and subtropical oceans, where they are the most numerically abundant photosynthetic organisms (Ribalet et al. [2015]). *Synechococcus* are slightly larger with diameters of $0.7\mu m - 1.6\mu m$ (Paulsen et al. [2015]) and are ubiquitous along a wider latitudinal range than *Prochlorococcus*, capable of being found in waters as cold as $2°C$ (Shapiro and Haugen [1988]). Picoeukaryotes refer to a diverse multi-phyletic group of photosynthetic eukaryotes that are smaller than $3\mu m$ in diameter, although typically still larger than the two previously mentioned groups.

Environmental conditions vary widely throughout the area of our study in the North Pacific Ocean. The subarctic gyre, situated in the northernmost region of the Pacific Ocean, is nutrient rich and high in primary production due to upwelling bringing nutrients up from the deep ocean. The subtropical gyre south of the subarctic gyre is a convergent body of water resulting in downwelling of nutrients, making these waters oligotrophic (Reid [1962]). This sharp contrast in nutrient concentrations between these two adjacent bodies of water, along with other physical processes give rise to the Transition Zone Chlorophyll Front (TZCF): a dynamic boundary between low and high chlorophyll levels in the ocean. The TZCF is known to be good indicator for primary productivity and separating sub-

arctic and subtropical phytoplankton communities (Polovina et al. [2017]). The TZCF is located between $32°N$ and $42°N$ and migrates seasonally about 1000km, reaching its southernmost position in the winter and northernmost position in the summer (Roden [1991]). The dataset used for this study was gathered on a single cruise along the 158°W meridian, collecting data between 22°N and 43°N latitude. The dataset contains high frequency abundance measurements of *Prochlorococcus*, *Synechococcus*, and picoeukaryotes along with various physical and chemical variables (see Table 2.1 for a summary of all variables).

Phytoplankton distribution is controlled by complex interaction of a number factors. Bottom-up control factors (e.g., temperature, nutrients) are known to be well correlated with phytoplankton abundances globally (Flombaum et al. [2013], Guo et al. [2013]). Picoplankton mortality can be strongly controlled by top-down controls, notably grazers and viruses (Guo et al. [2014], Guo et al. [2013]). Given our rich and diverse environmental dataset on the nutrient gradient of the North Pacific, we aim to determine if these bottom-up controls are sufficient to determine phytoplankton distributions, implicitly capturing top-down factors controlling populations. Thus, our first research question is: what physical-chemical environmental variables can best determine the concentration of picoplankton communities in the North Pacific Ocean? The different functional groups experience predation in differing amounts, even in the same location (Guo et al. [2013]), while competing with each other for resources. Ocean conditions in our study consistently have limiting nutrients relative to the physiological needs of their phytoplankton populations, and the identity of the limiting nutrient changes depending on the region of the ocean (Moore et al. [2013]). We hypothesize that the sharp changes in community structure occurring along the gradient can be explained by the combination of gradual changes in the suite of physical-chemical conditions. We will determine the relation between environmental conditions and abundance of each phytoplankton functional groups using linear models, generalized additive

models, and random forests, to cover a range between simple interpretable models and more complex models.

Phytoplankton play an important role in the biological carbon pump: a part of the carbon cycle where atmospheric CO_2 is absorbed by phytoplankton and a portion of it eventually sinks to the deep ocean to be sequestered in sediment. Total particulate and cellular sources of organic carbon, nitrogen, and phosphorus in the ocean typically follow a molar elemental ratio of 106C : 16N : 1P, known as the Redfield ratio (Redfield et al. [1963]). The Redfield ratio is an average and can vary considerably by region, depth and season. C:N:P stoichiometry also varies considerably between species, phytoplankton functional groups as a whole, and as a response to environmental conditions such as nutrient limitations (Liefer et al. [2019]; Heldal et al. [2003]; Gundersen et al. [2002]). Samples collected during the cruise enabled measurements of total particulate CNP. Our second research questions is: can we model cellular carbon, nitrogen and phosphorus of our three picoplankton groups as a function of abundances and cell volumes. We hypothesize that phytoplankton of the picoplanktonic size range and organic debris should account for nearly all organic macromolecules in oligotrophic region of the North Pacific. We will use Bayesian linear regression to model particulate organic C,N, and P. The model's posterior distributions allow us to estimate the an expected value and credible region for elemental quotas of C, N, and P in *Prochlorococcus*, *Synechococcus* and picoeukaryotes.

2 Methods

2.1 Data

The focus of this study is on data obtained from the "Gradients 2" cruise. The cruise departed from Honolulu, Hawaii, and did a round trip roughly along the 158°W meridian, collecting data between 22° and 43° north latitude. The cruise took place in 2017 from May 28 to June 13. We also look at datasets from two sister cruises Gradients 1 and Gradients 3, which took place along a similar trajectory in 2016 and 2019 respectively.

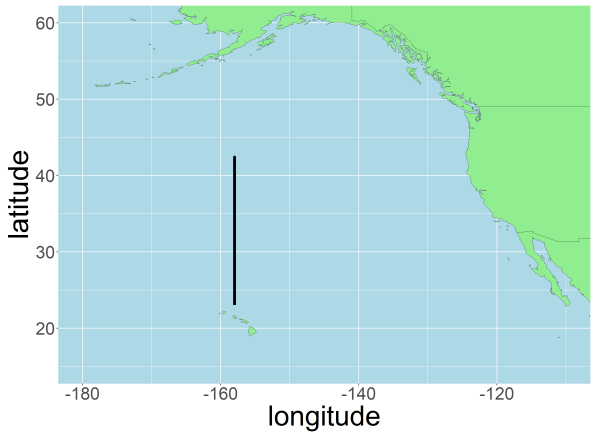

Figure 2.1: Map of Gradients 2 cruise trajectory (black line).

The Gradients 2 dataset contains high frequency (every $\sim$ 3 minutes) data, of cell abundance (cells/μL) of *Prochlorococcus*, *Synechococcus* and picoeukaryotes, collected by SeaFlow, a surface level underway flow cytometry instrument. Sea surface temperature (SST), salinity (SAL), phosphate concentration (PO4) and photosynthetically active radiation (PAR) measurements are also provided with similar high frequency. Phosphate measurements were only taken on the northward transect. The dataset frequently contained small gaps, less than 30 minutes, of missing data that were filled in with linear interpolation. A few large gaps were also present and left empty since linear interpolation would no longer be appropriate, given their larger size. Surface trace metal measurements were much lower in frequency but still taken over the whole latitude range. Data was grouped

and averaged into 0.01° latitude bins.

short name	n obs	units	description
pro	3555	$cells/\mu L$	*Prochlorococcus* abundance
syn	3555	$cells/\mu L$	*Synechococcus* abundance
pico	3555	$cells/\mu L$	picoeukaryotes abundance
pro_d	3555	μm	*Prochlorococcus* average spherical diameter
syn_d	3555	μm	*Synechococcus* average spherical diameter
pico_d	3555	μm	picoeukaryotes average spherical diameter
pro_C	3555	$pg/\mu L$	*Prochlorococcus* estimated biomass
syn_C	3555	$pg/\mu L$	*Synechococcus* estimated biomass
pico_C	3555	$pg/\mu L$	picoeukaryotes estimated biomass
SST	3555	$°C$	sea surface temperature
SAL	3510	PSU	salinity
PO4	1655	$nmol/L$	Phosphate concentration
PAR	2946	$\mu mol\ m^{-2}s^{-1}$	photosynthetically active radiation
Fe	129	$nmol/L$	iron concentration in seawater
Mn	140	$nmol/L$	manganese concentration in seawater
Cu	140	$nmol/L$	copper concentration in seawater
C	99	$\mu mol/L$	particulate organic carbon concentration
N	99	$\mu mol/L$	particulate organic nitrogen concentration
P	99	$\mu mol/L$	particulate organic phosphorus concentration

Table 2.1: Names, descriptions and number of observations for the variables used in the Gradients 2 dataset. Interpolation of small missing segments and binning data into 0.01°L already applied. Variables used in random forests are imputed to 3555 observations for those models, to match picoplankton abundances.

2.2 Models

We aim to both predict phytoplankton abundance accurately and understand the effect of each factor on abundance. We use three methods to model phytoplankton abundance: linear models, generalized additive models (GAMs) and random forests. The linear and generalized additive models provide a simple interpretable output describing predictors and their effect on abundance, while random forests can provide a more accurate model but with a more complex interpretation of its variables. All analyses were conducted with R (R Core Team [2020]) and all plots were made with `ggplot2` (Wickham [2016]).

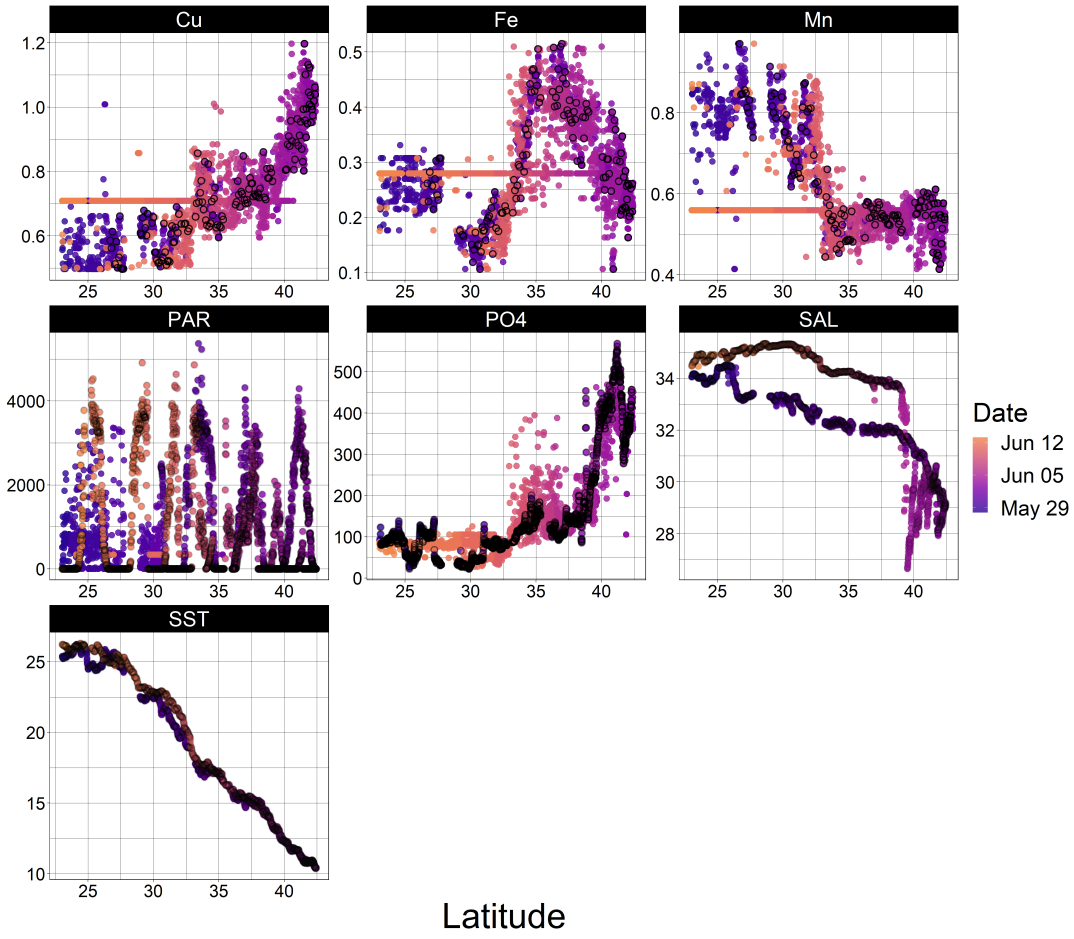

Figure 2.2: Imputed and raw data of all predictors used as input for random forests. Raw data are circled by a dark border to differentiate from imputed data.

2.2.1 Linear Models

Linear models are used for the initial analysis as they are very interpretable and can identify the most important predictors. For the linear models only, predictors were standardized to have 0 mean and identical variance to make model coefficients comparable. Since linear regression requires no missing data, we limit these models to the most abundant predictors: temperature, salinity, PAR and PO4. The model equation for all three picoplankton groups is:

$$\log(Abun) = \beta_0 + \beta_1(SST) + \beta_2(SAL) + \beta_3(PO_4) + \beta_4(PAR) + \epsilon, \quad \epsilon \sim N(0, \sigma) \quad (1)$$

where *Abun* is the abundance of *Prochlorococcus*, *Synechococcus*, or picoeukaryotes. No interaction terms between predictors are included in the model. We assume error terms ϵ to be iid and normally distributed with mean 0 and con-

stant variance. Since all predictors are standarized to a 0 mean, the intercept β_0 represents the log abundance of picoplankton at average environmental conditions for each predictor. Models were made using the base R `lm` function R Core Team [2020].

2.2.2 Generalized Additive Models

Liang et al. [2019] and Irwin and Finkel [2017] show that phytoplankton growth rates have a non-linear relationship with temperature and nutrient concentrations. Variation in phytoplankton growth rates is known empirically to be linked to variation in biomass. We use GAMs to capture these non-linear relationships, which fit smooth functions of the environmental predictors to the data in lieu of constant coefficients used linear models. The model equation for all three picoplankton groups is:

$$\log(Abun) = s(SST) + s(SAL) + s(PO4) + \epsilon, \quad \epsilon \sim N(0, \sigma) \tag{2}$$

where $s(\cdot)$ is a smooth function of the data. Smooth functions in this model are made of cubic splines with 9 knot positions. Flexibility of the smooth functions in GAMs are penalized by a smoothing parameter to avoid potential overfitting, which we choose with the help of k-fold cross validation ($k = 10$). Model does not include interaction terms between any of the predictors. GAMs serve well as an intermediate between linear models and random forests, as they are more complex than linear models yet much easier to interpret than machine learning algorithms such as random forests. We limit our GAMs to the same predictors as used in the linear models. All models are made using the `mgcv` R package [Wood, 2011].

2.2.3 Random Forests

In situ observations often have a limiting factor controlling phytoplankton populations and there are some sharp changes in abundance along the transect, suggesting there could be equally sharp changes in the relationship between abundance and

environmental predictors. Random forests provide a nonparametric approach for fitting flexible non-linear relationships to the data, making them more appropriate than GAMs for modeling these potential step-wise relationships between predictors and abundance. Random forests can deal with missing values by means of imputing them, allowing us to include sparse trace metal data in our models without having to discard predictors with higher frequency measurements used in the previous models. Random forests models and data imputations were done with the `randomForest` R package [Liaw and Wiener, 2002] and imputed data were only used for random forests.

We model picoplankton in a number of different ways using random forests, using SST, SAL, PO_4, PAR, Fe, Mn, and Cu as predictors. The three metals were selected from a set of six metals by LASSO using the `glmnet` package (Friedman et al. [2010]). To validate our models, we split our into test and training sets, chosen by splitting the data into alternating sets of 24 hours. These testing and training sets ensure the model is trained along the unique conditions throughout the entire transect of the cruise and that training data includes both day and night data on picoplankton. As with our previous models, we first model individual abundances of each picoplankton group. Phytoplankton abundances are not comparable across species since their nutrient requirements vary because of their different sizes. Thus, comparing abundances between the different phytoplankton groups can be misleading, making it hard to produce a meaningful form of relative abundance by summing up their abundances. Instead of relative abundances, we model the log of pro/syn and syn/pico abundance ratios. These two ratios still provide a relative measure of abundances allowing us to account implicitly for competition between the phytoplankton groups that co-occur in high quantities. To obtain a more comparable measure of individual picoplankton populations, we estimate carbon biomasses from abundances and average cell volumes using conversions obtained from our elemental quota models in our second analysis. We

then model these carbon biomass estimates against the same environmental predictors used in the individual abundance random forest models.

The relationships modeled between picoplankton abundances and the environmental predictors from the Gradients 2 data may not necessarily be representative of picoplankton abundances in general. We use random forests to model abundances from the gradients 1, 2 and 3 cruises using temperature, salinity and average cell diameter as a possible proxy for nutrient conditions. Only three predictors are used since sufficient phosphate and metals data are not available in the two other cruise datasets. These models will allow us to see how relationships between abundances and environmental predictors differ in similar datasets collected at different times.

2.3 Interpreting Models

Linear model coefficients and GAM smooth functions provide a simple interpretable description of the effect of each predictor on phytoplankton abundances. Although random forests perform very well at fitting a model to the data, they are not as interpretable as the two previous models. We therefore use some extra tools to understand the output of our random forests: variable importance measure and accumulated local effects (ALE) plots to describe the individual effect and importance of each predictor. Importance measures are computed using the `randomForestExplainer` package (Paluszynska et al. [2020]), and Accumulated local effects are computed using the `iml` package (Molnar et al. [2018]).

2.3.1 Accumulated Local Effects (ALE)

To visualize the individual effects of predictors from our random forests we use accumulated local effects (ALE) as suggested by Molnar. ALE plots hold a few advantages over other commonly used methods such as partial dependence (PD)

plots and marginal plots (M plots), namely, they are unbiased. For a two variable example, PD plots are calculated as:

$$f_{1,PD}\left(x_1\right) \equiv \mathbb{E}\left[f\left(x_1, X_2\right)\right] = \int p_2\left(x_2\right) f\left(x_1, x_2\right) dx_2,$$

where x_1 is the variable of interest, $f(\cdot)$ is the model, and $p_2(x_2)$ is the marginal density function of x_2. There are two main problems that occur in the likely case that the variables are correlated. The first problem is that using the marginal distribution of x_2 will likely extrapolate beyond the scope of the data, for certain combinations of (x_1, x_2). M plots solve this by using the conditional density $p_{2|1}(x_2 \mid x_1)$ instead of the marginal density $p_2(x_2)$, and are calculated by:

$$f_{1,M}\left(x_1\right) \equiv \mathbb{E}\left[f\left(X_1, X_2\right) \mid X_1 = x_1\right] = \int p_{2|1}\left(x_2 \mid x_1\right) f\left(x_1, x_2\right) dx_2.$$

A second problem with PD plots is that the interpretation of correlated variables remains entangled: the effect of changing x_1 will include the effect of changing x_2 as well. ALE plots address this second problem as well, and are calculated by:

$$\begin{aligned} f_{j,ALE}\left(x_j\right) &= \int_{x_{\min,j}}^{x_j} \mathbb{E}\left[f^j\left(X_j, \mathbf{X}_{\setminus j}\right) \mid X_j = z_j\right] dz_j - C \\ &= \int_{x_{\min,j}}^{x_j} \int p_{\setminus j|j}\left(x_{\setminus j} \mid z_j\right) f^j\left(z_j, x_{\setminus j}\right) dx_{\setminus j} dz_j - C, \end{aligned}$$

where $f^j\left(x_j, \mathbf{x}_{\setminus j}\right) \equiv \frac{\partial f\left(x_j, \mathbf{x}_{\setminus j}\right)}{\partial x_j}$ is the *local effect* of x_j on $f(\cdot)$. X_j is the variable of interest, $X_{\setminus j}$ is the set of all other variables, $x_{\min,j}$ is a value chosen near the lower bound of the effective support of $p_j(\cdot)$, and C is a constant chosen to vertically center the function at 0. The local effect isolates the effect of x_j from the other variables at a particular value $x_j = z_j$, and is then averaged across all values of $x_{\setminus j}$ weighted by the conditional density of $X_{\setminus j}$ given $X_j = z_j$. This averaged local effect is then integrated across all values of x_j up to z_j to account for the combined effect of all previous local effects, thus calculating the accumulated local effect. ALE plots use the conditional density similarly to M plots to avoid extrapolation,

while using local effects allows ALE plots to omit the effect of other variables from x_j (Apley and Zhu [2020]).

2.3.2 Importance Measures

There are various methods to evaluate a notion of how important a certain variable is to a random forest. In this study, we use and compare three different methods: mean minimal depth (MMD), mean square error (MSE) increase, and node purity increase. Mean minimal depth is the average depth across all trees at which a predictor is first used as the splitting criteria for a node. MSE increase is the change in mean squared error on out-of-bag data, from before and after randomly permuting the predictor of interest. For regression random forests, node purity increase is calculated as the total decrease residual sum of squares from splitting on the variable of interest and then averaged across all trees (Liaw and Wiener [2002]).

2.4 Elemental Quota Models

Particulate organic carbon, nitrogen and phosphorus concentrations were regularly measured along the Gradients 2 transect. We model these biomasses as a function of abundances of the three phytoplankton groups. There are a few problems with using the data as is: first, the biomass samples are unfiltered and can contain organic particles larger than can be measured by the flow cytometer. To solve this, we restrict our model to data taken in oligotrophic waters below 35°N latitude. Irwin et al. [2006] show that smaller phytoplankton dominate large cells in oligotrophic waters and that the picoplankton size fraction ($< 2\mu m$) accounts for nearly all biomass in low nutrient conditions. Second, biomass samples were taken from much larger volumes of water samples while seaflow abundance is determined from small microliter samples, making measurements more susceptible to the ocean's patchiness. Abundance therefore suffered from higher variance and outliers which could lead to selecting erroneous data points to use in the model. To

solve this, abundance is smoothed using a one hour running median before selecting data points matching the location of biomass samples. Third, non-photosynthetic bacteria and organic debris contribute to the total particulate organic matter measurements, but were not counted by the SeaFlow instrument. To solve this problem, we assume this extra organic mass to be relatively constant throughout the transect and add an intercept to the model. We were unable to verify this assumption due to lack of relevant data and it may be a source of considerable uncertainty.

We use bayesian linear regression to model organic biomass of carbon, nitrogen and phosphorus as a function abundance and volume of the three phytoplankton groups. Linear models are chosen so that the model coefficients provide an interpretable conversion ratio from volume to biomass. The model formula is:

$$C = \beta_0 + \beta_1(pro * vol) + \beta_2(syn * vol) + \beta_3(pico * vol) + \epsilon, \quad \epsilon \sim N(0, \sigma) \qquad (3)$$

where C is the concentration of carbon, nitrogen, or phosphorus in $pmol/\mu L$ and *pro*, *syn* and *pico* are the abundances of the three phytoplankton groups in $cells/\mu L$. Coefficients β_{1-3} are for the interaction terms between phytoplankton abundances and their respective spherical volumes (vol) and have units $cells/\mu L\ \mu m^3$. The interaction terms represent cell biovolumes: the quantity in moles of C, N, or P per unit of volume. The intercept β_0 represents organic debris, assumed to be constant due to not having any measurements pertaining to debris. All parameters are restricted to positive values, by specifying non-negative uninformative priors for each of them, in order keep their interpretation meaningful, since they all represent a quantity of organic matter. Bayesian regression provides a simple way to estimate average cell biovolumes and range of possible values via the parameters' posterior distributions. We assess our bayesian model using two diagnostic measures: n_eff, the effective number of independent MCMC samlpes for a particular parameter, and Rhat, a measure of how well chains have converged to a stable posterior distribution. We use an adapted R^2 for Bayesian regression as

defined by Gelman et al. [2019] to evaluate model goodness of fit. Models were fit using `rstan` (Stan Development Team [2020]).

Model coefficients can be used to create separate volume to particulate organic matter biomass conversions for each phytoplankton group. We assume elemental cell quotas and the way they vary with cell volume are fixed over the whole transect, and estimate particulate organic biomass for the entire dataset. Despite only training the model on data up to 35° latitude predicted carbon biomass estimates were used in section 3 for the carbon biomass random forest models.

3 Picoplankton Distribution Models

3.1 Results

3.1.1 Linear Models

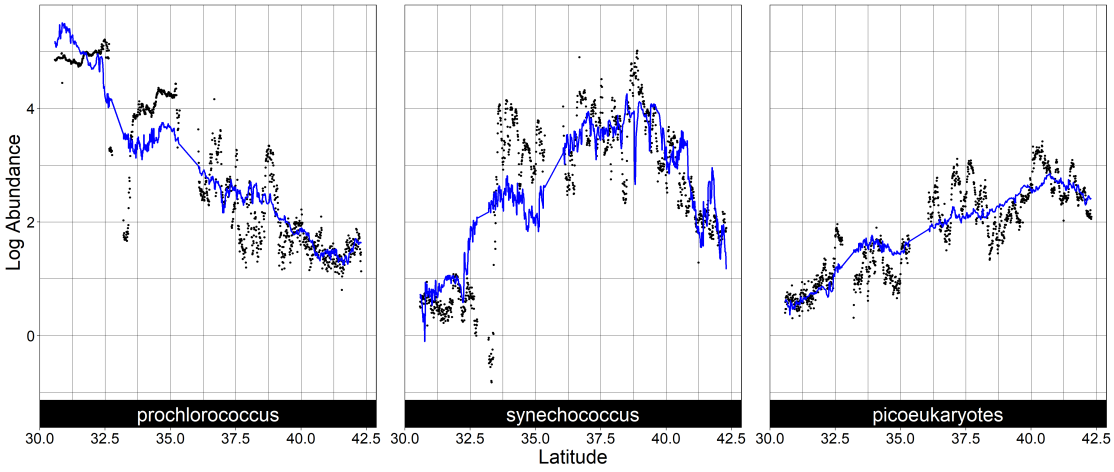

Figure 3.1: Linear models of phytoplankton log abundance. Blue line: predicted log abundance fit, black dots: raw data averaged over 0.01° latitude bins.

Underway phosphate measurements were only available on the northward transect and PAR measurements were not available south of 32.5°N, resulting in a restricted latitude range for these first models (Figure 3.1). The *Prochlorococcus* model had the best adjusted R^2 value at 0.761 All predictors in the linear models are highly significant (p-value < 0.001), with the exception of PAR in the *Synechococcus* model, and PO4 in the picoeukaryotes model, which were significant with a p-value < 0.05, (Table 3.2). Diagnostic plots for the linear model residuals can be found in the Appendix (Figures 6.2, 6.3, and 6.4). Lag-1 autocorrelation in residuals for *Prochlorococcus*, *Synechococcus* and picoeukaryotes are 0.956, 0.975, and 0.947 respectively. These autocorrelations result in an effective sample size of 24, 14, and 29 for *Prochlorococcus*, *Synechococcus* and picoeukaryotes respectively, calculated as suggested by Reiher and Huzzen [1967]. The largest coefficient in all models is always SST followed by SAL. Note that in each model, SST and SAL are always of the opposite sign, despite SST and SAL being positively correlated with a Pearson correlation coefficient of 0.879. PAR and PO4 coefficient effect sizes are

relatively small compared to their corresponding SST effect size.

Linear model	Adjusted R^2	AIC
Prochlorococcus	0.837	1710.8
Synechococcus	0.636	2540.9
picoeukaryotes	0.756	1015.6

Table 3.1: Adjusted R^2 and AIC values of linear models in Figure 3.1. AIC values comparable to corresponding GAM AIC values in Table 3.3.

Coefficient	pro	s.e.	p-value	syn	s.e.	p-value	pico	s.e.	p-value
Intercept	4.2740	0.0400	0.000	2.0979	0.0594	8.47e-181	1.2414	0.0288	1.84e-234
SST	2.363	0.0610	1.36e-204	-3.768	0.090	1.25e-224	-1.442	0.044	1.33e-163
SAL	-0.497	0.0821	1.93e-09	2.675	0.122	2.00e-88	0.594	0.059	7.84e-23
PO4	0.085	0.0306	5.73e-03	-0.561	0.045	5.63e-33	0.019	0.022	3.99e-01
PAR	-0.133	0.0165	1.89e-15	0.062	0.025	1.11e-02	0.062	0.012	1.80e-07

Table 3.2: Coefficients from linear models in Figure 3.1. All predictors have been standarized to have mean 0 and unit variance so that coefficients are comparable to each other. P-values may be misleading due to high autocorrelation in data.

3.1.2 GAMs

GAMs were also fit using only data from the northward transect, but PAR was removed as a predictor due to its very small effect size in the linear models (Table 3.2). Predicted log abundance matches the main patterns in the data, including the sharp transition in *Synechococcus* log abundance near 33° latitude, which the linear model failed to capture (Firgure 3.2).

GAM models were trained on the exact same data as linear models making their adjusted R squared and AIC calues comparable (see Tables 3.1 and 3.3). Adjusted R squared and AIC values were improved in every model. Diagnostic plots for the GAM residuals can be found in the Appendix (Figures 6.5, 6.6, and 6.7). Error distributions follow a mostly normal distribution with some slight deviations in the tails of the distributions most likely due to some of the steep shifts in abundances missed by the models. Lag-1 autocorrelation in residuals for *Prochlorococcus*, *Synechococcus* and picoeukaryotes are 0.948, 0.949, and 0.873 respectively. These autocorrelations result in an effective sample size of 29, 28, and 72 for *Prochlorococcus*, *Synechococcus* and picoeukaryotes respectively, calculated as suggested by Reiher and Huzzen [1967].

GAM smooth functions for temperature (Table 3.3) have the largest effect size for all three phytoplankton groups, compared to salinity and PO4. Each temperature curve has a distinctive peak where temperature contributes the most to abundance (Figure 3.3). Smooth functions for salinity in picoeukaryotes and *Prochlorococcus* have a similar shape to their temperature counterparts, although smaller in effect size. The PO4 smooth function for picoeukaryotes barely deviates from zero, while the PO4 smooth functions for the two other groups show a slight decrease at low PO4 concentrations.

GAM	Adjusted R^2	AIC
Prochlorococcus	0.864	1523.2
Synechococcus	0.843	1665.5
picoeukaryotes	0.900	89.9

Table 3.3: Adjusted R^2 and AIC values of generalized additive models. AIC values comparable to corresponding linear model AIC values in Table 3.1.

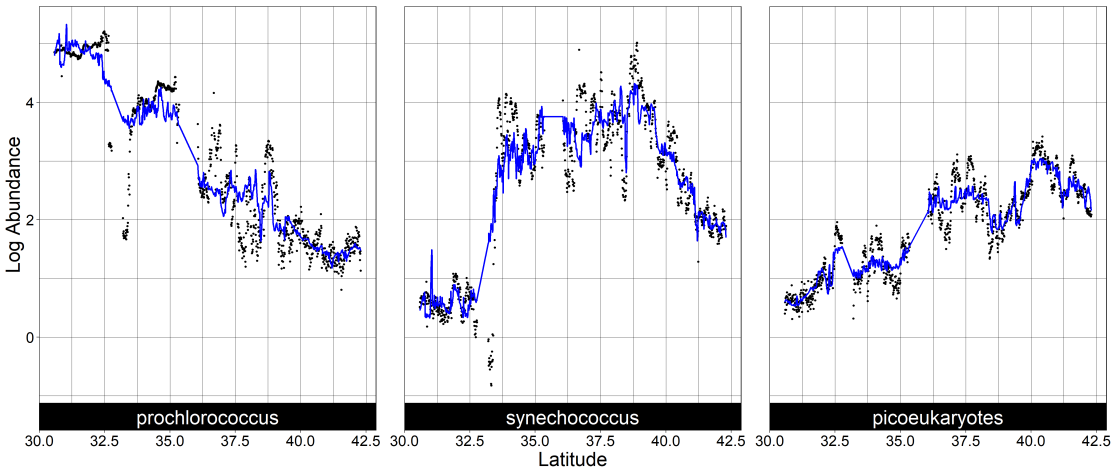

Figure 3.2: Generalized additive models of phytoplankton log abundance. Deviance explained by model: pro 86.6%, syn 84.5%, pico 90.1%.

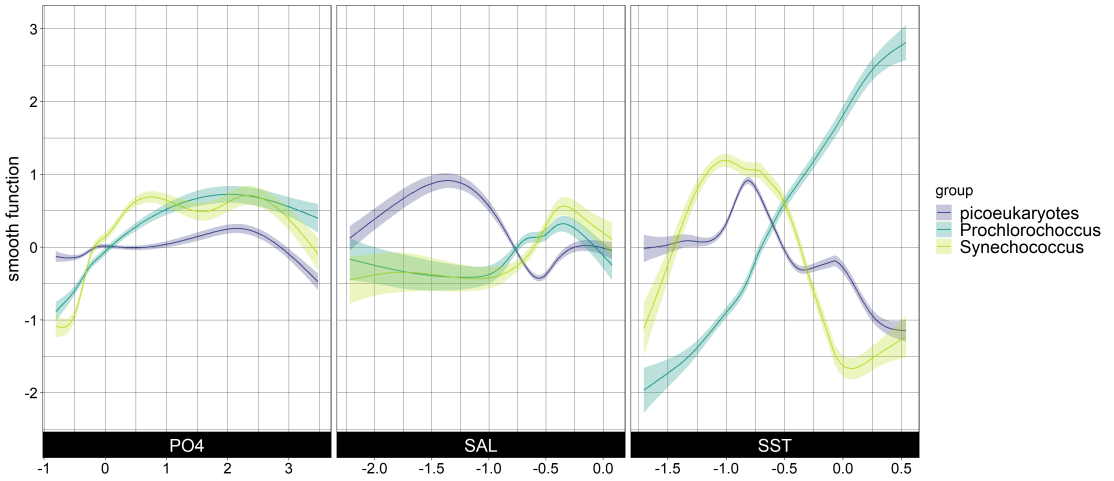

Figure 3.3: GAM smooth functions, with standard error (shaded area), of models from Figure 3.2.

3.1.3 Random Forests

Random forests captures log abundance of all three phytoplankton groups exceptionally well. Models explain 85% to 88% of the variance and even capture some of the slight differences in abundance between northward and southward transects of the cruise (Figure 3.4). Accumulated local effects (ALE) curves in Figure 3.5 show temperature and salinity have the largest effect on log abundance (note the larger range on the scales of SST and SAL). Temperature ALE curves show a clear optimum for each species: high temperature for *Prochlorococcus*, intermediate temperatures for *Synechococcus* and low temperatures for picoeukaryotes.

Phosphate ALE curves at very low phosphate concentrations show a positive effect on *Prochlorococcus* abundance and a negative effect on the two other picoplankton groups. *Synechococcus* thrive better than the two other groups on intermediate phosphate levels and picoeukaryotes has an affinity for the highest phosphate concentrations. PAR ALE curves are all flat, this is consistent with the near zero coefficients in the linear models. Iron and especially manganese ALE curves demonstrate a clear stepwise behavior. Cu ALE curves have little effect size with the only discernible pattern being a peak at intermediate copper concentrations for *Synechococcus*.

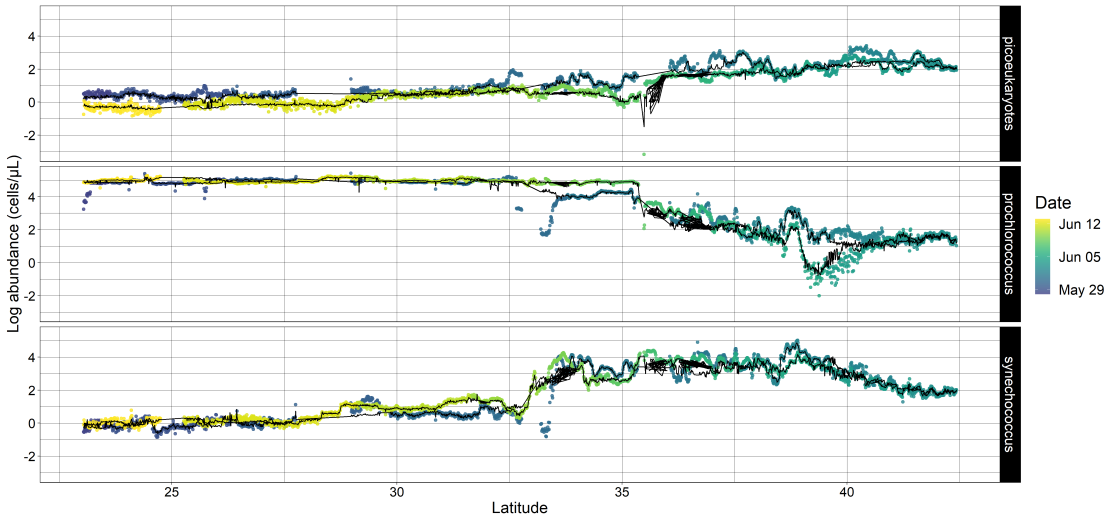

Figure 3.4: Random forest models of log abundance of *Prochlorococcus* (pro), *Synechococcus* (syn), and Picoeukaryotes (pico). Black lines: random forest model fits, dots: cruise raw data binned in 0.01 latitude intervals, colored by time of collection. Percent variance explained on test data: pro 85.7%, syn 87.0%, pico 88.0%.

All three importance measures agree relatively well with each other amongst all phytoplankton groups (Figure 3.6). Temperature is always the most important predictor in the three importance plots, which reflects well the fact that temperature ALE curves had the largest effect sizes out of all the other predictors. Phosphate and salinity are the next two most important predictors, with phosphate being more important in all three measures for *Synechococcus*. PAR is

unequivocally the least important predictor and this is consistent with all previous results on PAR. All three mportance measures of Mn and Fe for *Synechococcus* were comparable to salinity and PO_4. Copper has peculiarly high importance in node purity increase for *Prochlorococcus*, seemingly contradicting its relatively small effect size in its associated ALE curve.

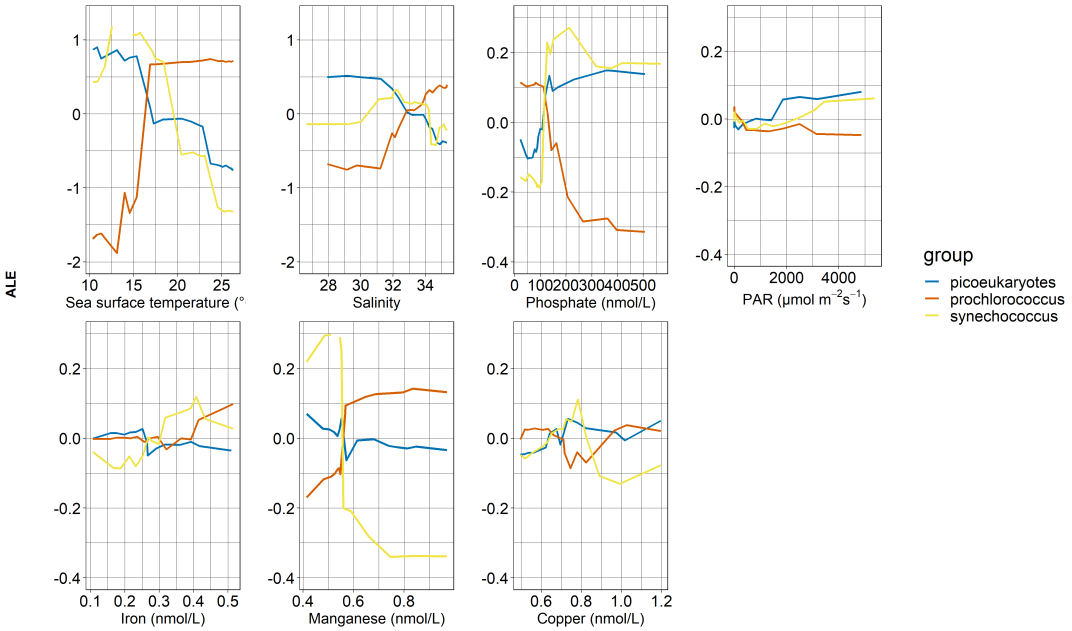

Figure 3.5: Accumulated local effects (ALE) plots of environmental covariates for each random forests in Figure 3.4.

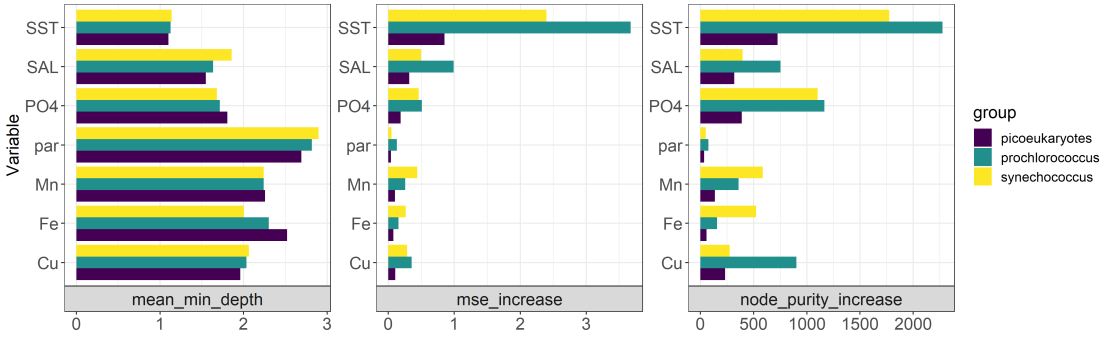

Figure 3.6: Three selected importance measures computed from the random forests in Figure 3.4.

3.1.4 Abundance Ratios

As with abundance models, our random forests predict log abundance ratios with high explained variance (Figure 3.7). Log abundance ratios show somewhat stepwise behavior, with syn/pico ratios having only one stepwise increase while pro/syn ratios show two stepwise decreases along the gradient. ALE curves (Figure 3.8) essentially provide the same information as ALE curves for individual phytoplankton groups (Figure 3.5): Temperature and salinity have the largest effect size and their ALE curves show that each picoplankton group has clear optimal range for temperature and phosphate concentrations. Stepwise behavior in iron and manganese is also maintained for *Prochlorococcus* and *Synechococcus* as shown in the corresponding curves for pro/syn in Figure 3.8. One notable difference is that the ALE curve for copper now show a clear peaks in their relationship, both showing the advantage of *Synechococcus* and intermediate phosphate concentrations. Temperature consistently remains the most important predictor across all measures (Figure 3.9). Copper has particularly high importance for pro/syn ratios while iron was the most important metal for syn/pico ratios.

3.1.5 Comparing Gradients Cruises 1, 2, and 3

We bring in the Gradients 1 and Gradients 3 data set to see if the relationship between phytoplankton abundance and the environmental covariates are the same across three different years in the same geographic location of the ocean. Differences in absolute values of effect sizes of ALE curves (3.10) can be found for all predictors between the three cruises. This reflects the different levels of abundances in phytoplankton amongst the three cruises (Need to add abundance raw data for G1 and G3 in appendix for reference). Sharp shifts in temperature ALE curves for *Prochlorococcus* and *Synechococcus* occur mostly at different thresholds amongst the three cruises. This is consistent with the sharp shifts in abundance for *Prochlorococcus* and *Synechococcus* occurring at different latitudes/temperatures

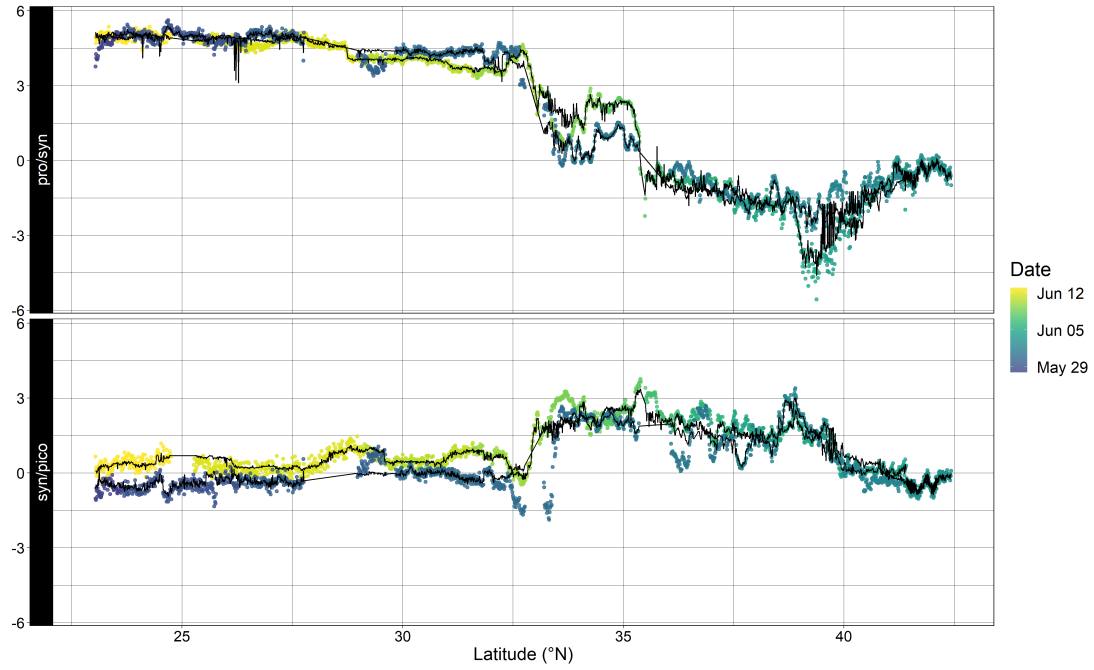

Figure 3.7: Random forest models of pro / syn and syn / pico abundance log ratios. Black lines: random forest model fits, dots: cruise raw data binned in 0.01 latitude intervals, colored by time of collection. Percent variance explainedon test data: pro/syn 96.0% syn/pico 68.0%.

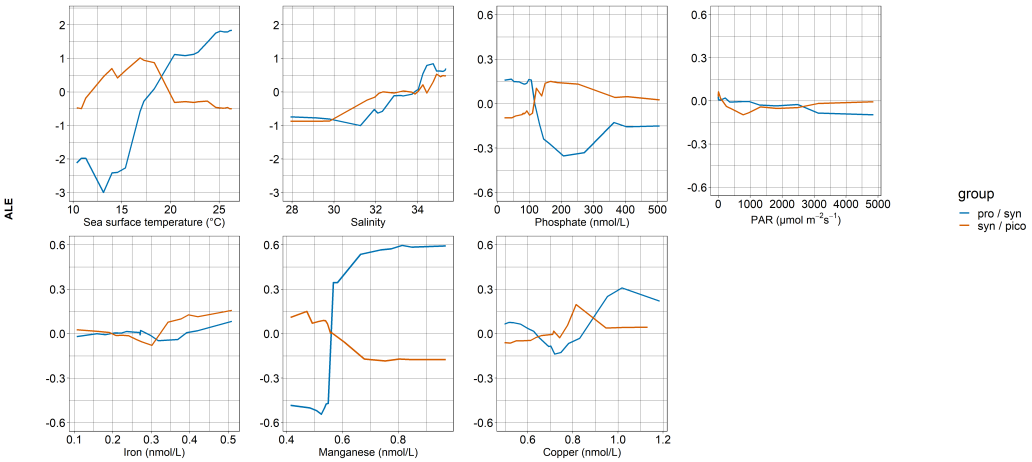

Figure 3.8: Accumulated local effects (ALE) plots of environmental covariates abundance log ratio random forests in Figure 3.7.

in the three cruises.

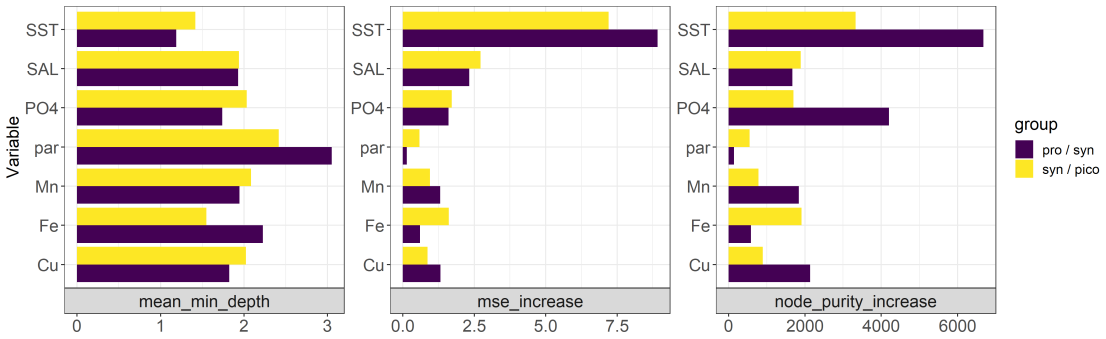

Figure 3.9: Importance measures computed from the abundance log ratio random forest in Figure 3.7. Mse_increase and node_purity_increase for syn\pico has been scaled up by a factor of 5 for visibility.

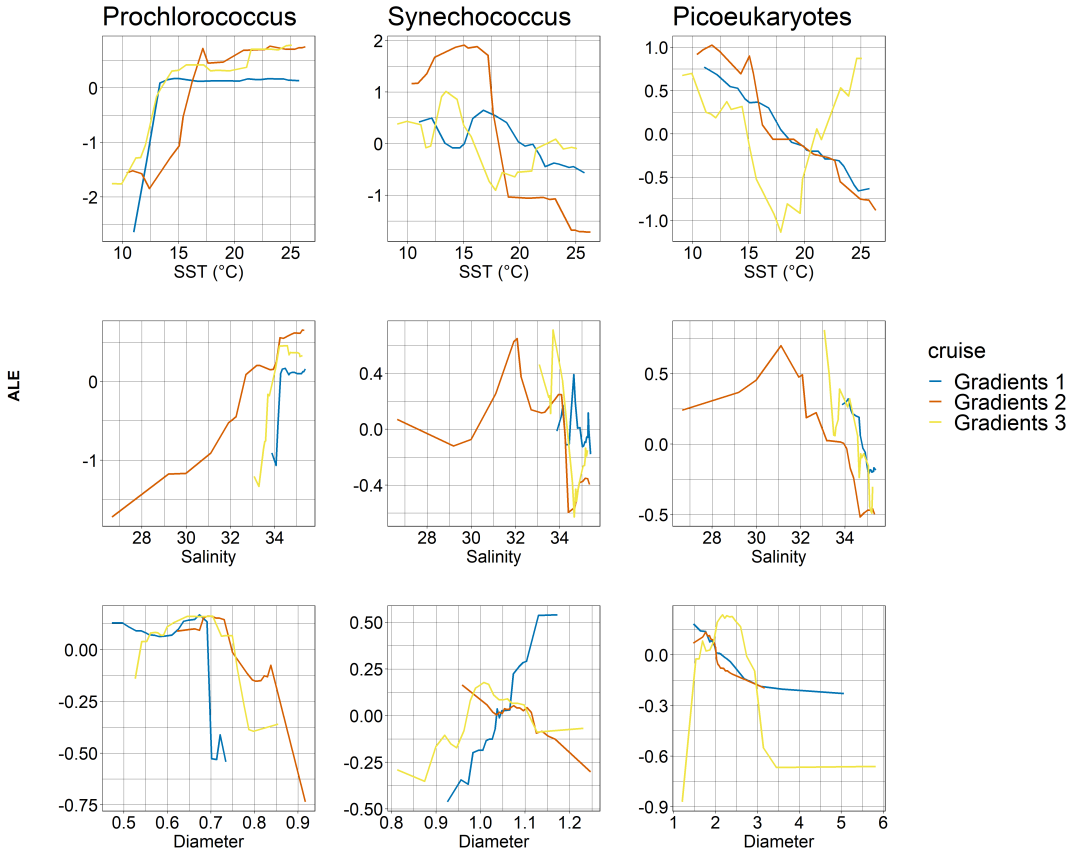

Figure 3.10: ALE plots of random forests modeled on the three Gradients cruises for *Prochlorococcus*, *Synechococccus* and picoeukaryotes.

3.1.6 Carbon Biomass

Carbon biomasses were estimated using elemental carbon models from section 4 (see methods section 2.4 also). Carbon biomass random forests perform very sim-

ilarly log abundance counterparts with explained variances of: 87.2%, 85.4%, and 88.7% for *Prochlorococcus*, *Synechococcus* and picoeukaryotes respectively (Figure 3.11). ALE plots in Figure 3.12 shows temperature and salinity having a considerably larger effect size than other predictors. Temperature and phosphate ALE curves show the same essential relationship as with their corresponding abundance model ALE curves. *Prochlorococcus* have higher ALE curve values at high temperature and low phosphate than the two other groups, *Synechococcus* has higher ALE curves at intermediate temperature and phosphate levels, and picoeurakyote ALE curves are greatest at low temperatures and high phosphate levels. Iron, manganese and copper ALE curves (Figure 3.12) also show higly similar relationships to their abundance counterparts (Figure 3.5). PAR ALE curves are nearly flat, as with all previous random forest models. Importance measures in Figure 3.13 all show temperature as the most important followed by salinity and phosphate. Metals' imortance measure for all three groups were mostly all comparable in size to phosphate importances.

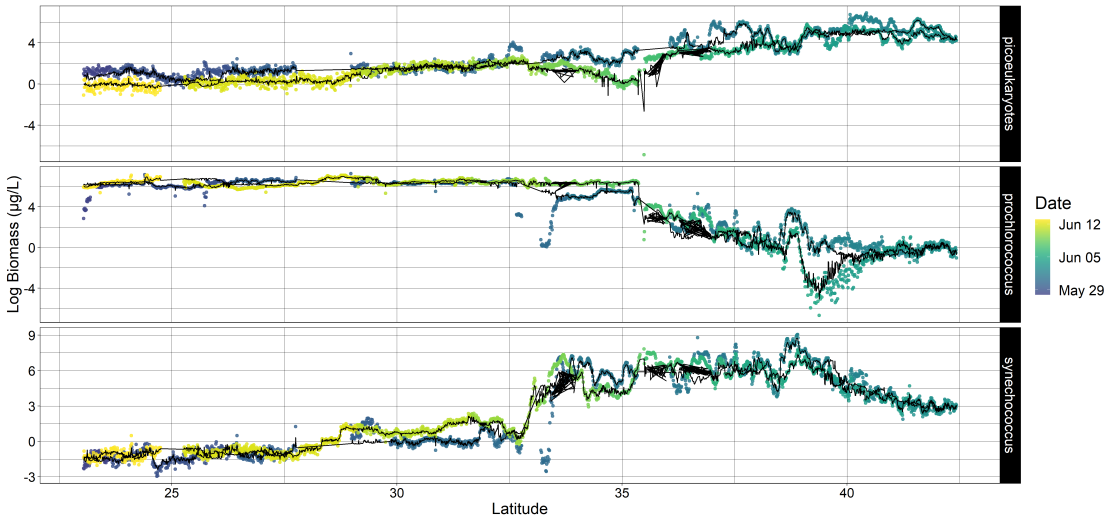

Figure 3.11: Random forest models on phytoplankton log carbon biomass of *Prochlorococcus* (pro), *Synechococcus* (syn), and Picoeukaryotes (pico). Black lines: random forest model fits, dots: cruise raw data binned in 0.01 latitude intervals, colored by time of collection. Percent variance explained on test data: pro 87.2%, syn 85.4%, pico 88.7%.

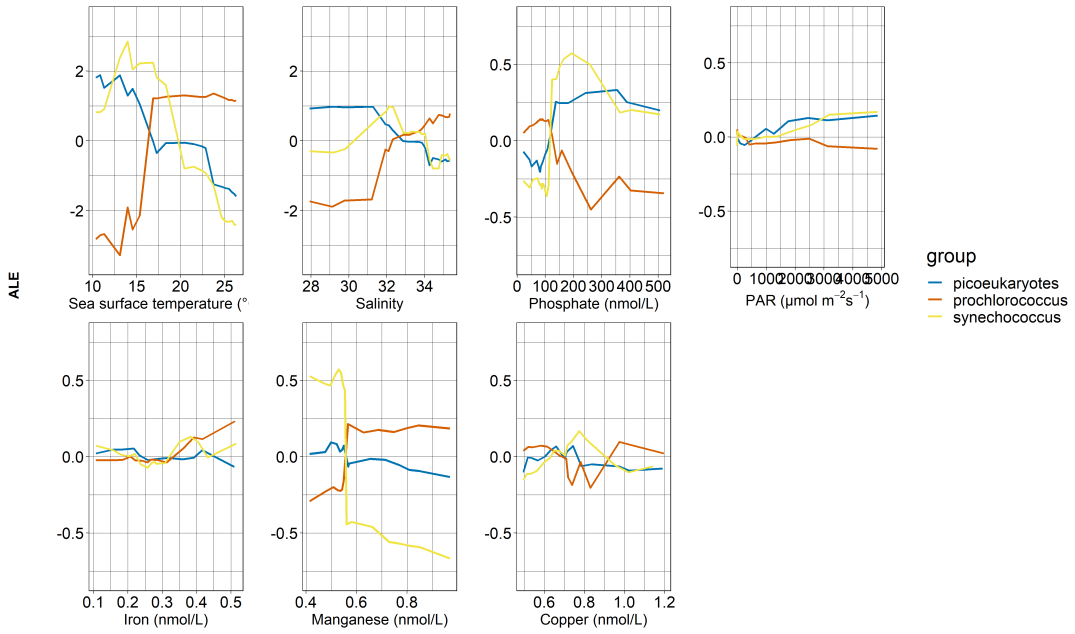

Figure 3.12: Accumulated local effects (ALE) plots of environmental covariates for each random forest in Figure 3.11.

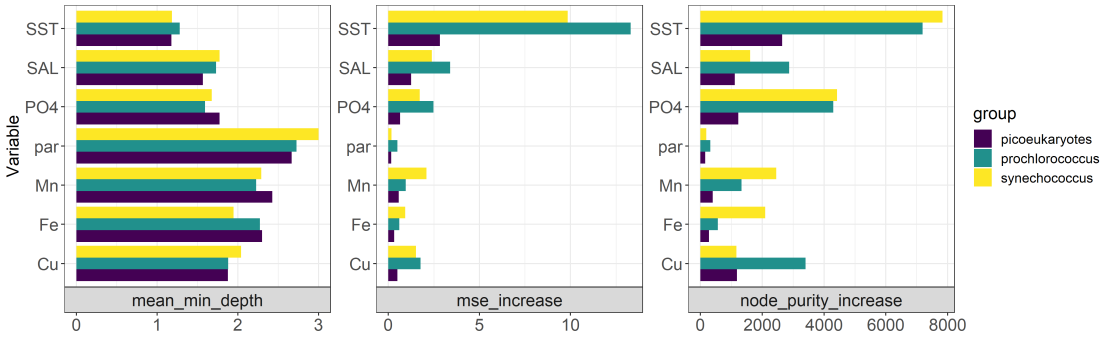

Figure 3.13: Importance measures computed from the random forests in Figure 3.11.

3.2 Discussion

We've modeled the abundances and biomasses of the three dominant photosynthetic picoplankton groups along the gradient in the North Pacific using a series of statistical models. We showed that temperature and salinity have a non-linear relationship to abundance and that they are the main predictors of abundance (Tables 3.2 & 3.6). Nutrients in general tend to have a smaller step-wise relationship to abundance and our results suggested that iron and phosphate are likely limiting nutrients in the lower latitudes. Nutrients' effects on abundances were

markedly different from their effect on carbon biomass concentrations, especially copper concentrations which had a role in determining *Prochlorococcus* and *Synechococcus* biomass but no effect on their abundances.

3.2.1 Linear Models of Picoplankton Abundance

Our linear models captured some of the main features explaining abundances of the three main picoplankton groups, but the models contain a number of inadequacies suggesting better models are needed. The low adjusted R^2 values (Table 3.1) for the linear models suggest linear relationships are not appropriate to capture the relationship between the environment and abundances. In particular, the *Synechococcus* linear model (Figure 3.1) misses the sharp increase in log abundance near 33° latitude, a main goal of this study. Although PAR is significant (t-test, $\alpha < 0.05$) in two of the models (Table 3.2), its effect size in all three models is very small, suggesting light levels play little role in predicting current cell abundance in this part of the surface ocean. Picoeukaryotes are larger cells and much more abundant in the nutrient rich waters of the North Pacific Ocean where the competitive advantage of small cells acquiring more nutrients, due to their lower surface area-to-volume ratios, are no longer important. Thus, we could reasonably expect that phosphate plays an important role in picoeukaryote abundance, yet PO4's coefficient in the picoeukaryotes model is insignificant with a minimal effect size (Table 3.2). This suggests linear relationships are most likely not appropriate to model PO4's effect or perhaps that temperature better reflects nutrient conditions in the ocean. Temperature has the largest effect size in all three picoplankton models and yet it's also known to have a non-linear relationship with growth rate (Liang et al. [2019]). Having temperature with such a large and significant effect despite its inadequate linear relationship emphasizes just how important temperature is as a predictor. As previously noted, the coefficient for SST is always opposite sign of SAL even though these two predictors are positively correlated (Figure 6.1). This may be the models' attempt in adjusting to a

relationship with temperature that should be non-linear. Lastly, p-values for the model coefficients are most likely highly inflated making the above interpretations hard to validate. Our errors are fairly well normally distributed but the high autocorrelation in the data reduces our effective sample sizes to 24, 14, and 29, for *Prochlorococcus*, *Synechococcus* and picoeukaryotes respectively, making perhaps only the picoeukaryotes model results viable.

3.2.2 Generalized Additive Models of Picoplankton Abundance

Our fairly simple GAMs with just three predictors capture more than 85% of the variance in abundance for all three picoplankton groups. Like the linear models, GAMs maintain temperature as the most important predictor with the largest effect size. Phosphate concentration smooth functions' effect remains small, showing just a small decrease in abundance for *Prochlorococcus* and *Synechococcus* below 100 $nmol/L$ and slight increase in the highest concentrations. The temperature smooth function for picoeukaryotes shows a number of local peaks despite the smoothing penalty. This jagged relationship with temperature is most likely a result of its strong correlation with other environmental predictors (see Figure 6.1) including those we could not include in this model, whose effect would be predominantly absorbed into temperature. Errors in our GAMs are also fairly well normally distributed but still suffer from high autocorrelation, although slightly lower tham those of the linear models. Thus the effective sample sizes of 29, 28, and 72, for *Prochlorococcus*, *Synechococcus* and picoeukaryotes respectively, make our GAM results much more viable than the linear models. Overall, GAMs are much more effective than linear models at capturing variation in phytoplankton populations using just the most readily available data, as validated by their lower AIC values (Table 3.3). However, the unusual temperature smooth functions warrant further investigation and there isn't enough metals data to verify their importance using a GAM. This highlights the need to use non-linear models capable of incorporating our low frequency nutrient measurements, hence our choice to use random forest

models.

3.2.3 Absolute Abundance and Abundance Ratio Random Forests

Random forests performed best in fitting the data while also showing the clearest picture of the effect and importance of the environmental predictors. Log abundance and log abundance ratio random forests all have high explained variance ranging from $85\% - 88\%$ (Figures 3.4 & 3.7), still suggesting potential over-fitting of the models. ALE curves for individual abundance models (Figure 3.5) are very similar in shape to their GAM smooth function counterparts (Figure 3.3) and account for the largest effect sizes of all predictors. Phosphate concentration ALE curves however, show a much clearer relationship with abundance than GAM smooth curves for phosphate. Phosphate ALE curves show a step-wise increase in effect at roughly 100 $nmol/L$ for *Synechococcus* and picoeukaryotes. Iron ALE curves show a similar increase in the $0.3 - 0.4$ $nmol/L$ range for *Prochlorococcus* and *Synechococcus*. Moore et al. [2013] note that phosphate and iron are generally limiting nutrients for phytoplankton in the oligotrophic gyre of the North Pacific Ocean. These step-wise increases to abundance for phosphate and iron support the idea that they were limiting nutrients up until reaching the threshold concentrations mentioned above. Manganese ALE curves show a strong stepwise relationship for all 3 groups at roughly 5.5 $nmol/L$, with concentrations above the threshold having a positive on *Prochlorococcus* and negative effect on *Synechococcus* and picoeukaryotes. Twining and Baines [2013] and Moore et al. [2013] show phytoplankton to have consistently lower metal quotas for manganese than for iron. Since manganese concentrations are almost always higher than iron concentrations, manganese is most likely not a limiting nutrient. Note that the concentrations for the step-wise changes at 100 $nmol$ P$/L$, $0.3 - 0.4$ $nmol$ Fe$/L$, and 5.5 $nmol$ Mn$/L$, all occur at roughly 33°N latitude (see Figure 2.2): the same region at which *Synechococcus* abundance sharply increase and where there is a sudden drop in *Prochlorococcus* abundance. Thus these sudden changes in nutrient

concentrations may be good environmental indicators of when sharp changes in *Prochlorococcus* and *Synechococcus* abundances occur.

To account for the effect of competition and relative advantage between picoplankton groups for accessing resources in different conditions, we've modeled the log ratios of their abundances. We considered ratios between *Prochlorococcus* and *Synechococcus*, who are the two main competitors in lower latitudes, and ratios between *Synechococcus* and picoeukaryotes, who are the two main competitors in the higher latitudes. ALE curves for log abundance ratios in Figure 3.8 largely communicate the same changes in abundances as ALE curves for individual log abundance models (Figure 3.5). This could suggest that effects competition between groups are already implicitly captured in individual models. One new piece of information however comes from the copper ALE curves, which are much more clearly resolved in Figure 3.8. They show that *Synechococcus* profit the most from intermediate concentrations of copper, while *Prochlorococcus* benefits from both higher and lower concentrations of copper, relative to *Synechococcus* abundance.

3.2.4 Comparing Random Forests on Gradients Cruises 1, 2 & 3

We modeled picoplankton distributions from two other similar cruises to see if abundances maintain the same relationship to their environmental conditions. Due to limits on common data between the three datasets, we had to limit predictors to temperature and salinity, and used average cell diameter as well to gain potential insight on nutrient conditions. Almost all temperature ALE curves in Figure 3.10 show either very sharp increases/decreases or wavy shapes with multiple peaks, a pattern that seems unusual knowing that temperature tends to have a more gradual relationship between temperature and phytoplankton growth rates (Liang et al. [2019]). For example, the Gradients 2 temperature ALE curve for *Synechococcus* (Figure 3.10), shows an extremely sharp decrease around $18°C$, much more steep than its counterpart in Figure 3.5. This is most likely caused by temperature

having to account for the effects of its correlating nutrients, that could not be included in this model. Thus, varying nutrient conditions may explain some of the differences in temperature relationships to phytoplankton abundance between the three cruises. Diameter ALE curves in Figure 3.10 show differing size ranges and effects between the three cruises, supporting our assumption that nutrient conditions were varying between the cruises, but limitations in data collection prevent us from further verifying this.

3.2.5 Carbon Biomass Random Forests

Carbon biomass estimates provides a measurable quantity that is comparable across the different picoplankton groups, and that accounts for both abundance and cell size. Our biomass ALE curves (Figure 3.12) reveal notably similar relationships to those of our abundance model ALE curves (figure 3.5). The close similarity in model fits and ALE curve relationships suggest biomass and abundance are equally predictable using our set of environmental predictors. Copper ALE curves in all our random forest models (log abundance, lag abundance ratios, and log biomass) all show a slight preference for intermediate copper concentrations for *Synechococcus*. Copper does play a few important biological roles in phytoplankton (Twining and Baines [2013]). Mann et al. [2002] show that *Synechococcus* in particular were generally resistant to copper toxicity in high copper concentrations. This may be why *Synechococcus* appears to benefit more from higher concentrations of copper.

3.2.6 Model limitations

Light is essential for phytoplankton to photosynthesize and consequently grow, yet in all our models, PAR is consistently the least important predictor. Flombaum et al. [2013] show that average monthly PAR is an important predictor of average monthly global populations of *Prochlorococcus* and *Synechococcus*. One reason for

PAR being unimportant may be that we are using surface PAR measurements which greatly overestimate irradiation levels below the surface that phytoplankton are actually receiving. Since we modeled surface level populations, light may also essentially always be available in excess relative to phytoplanktons' other needs for growth. Having access to measurements or estimates of the mixed layer depth may be a better proxy of average available light near the surface to use in our models. Another possible reason is that phytoplankton growth rates have a lagged response to light exposure. Since abundance data collected varies over both time, space, and consequently nutrient conditions as well, it would be impractical to determine the correct lag used to adjust PAR data.

4 Analysis of Elemental C, N, P Cell Content in Picoplankton

4.1 Results

All parameters in macromolecular quota models converged with R hat values within 0.01 of 1 and sufficient effective sample size (n_eff). Model parameter outputs and parameter distribution plots can be found in appendix Tables 6.1, 6.2 & 6.3, and appendix Figures 6.8, 6.9, & 6.10. The summarized results in Tables 4.1 - 4.3 are averaged results for the models up to 35° as shown in Figure 4.1. Quota models for C,N,P had a Bayesian R^2 of 0.34, 0.64, 0.64 respectively.

The large majority of organic C, N, and P in oligotrophic waters is attributed to debris, especially in carbon, which represents 61.7% of organic carbon (Table 4.1). Excluding debris, picoeukaryotes account for the largest proportion of phosphorus, while *Prochlorococcus* accounts for the most carbon and nitrogen (see Table 4.1 and Figure 4.2). However, when accounting for cell volume, Table 4.2 shows that *Synechococcus* has the highest cell density of carbon, nitrogen and phosphorus by a fairly large margin compared to the next most abundant group. *Synechococcus* on average have 81.2% more carbon than picoeukaryotes, 215% more nitrogen than *Prochlorococcus* and 106% more phosphorus than picoeukaryotes. Table 4.3 shows that total elemental ratios match closely to Redfield ratios of carbon, nitrogen and phosphate. In contrast, elemental ratios in the composition of phytoplankton and debris vary substantially from Redfield. Ratios range from: 58.18 to 170.14 for C:P, 7.06 to 41.22 for N:P and 3.86 to 8.64 for C:N (Table 4.3).

Above 35°N latitude, our extrapolated model predicts an overall increase in C, N, and P biomass, attributed to a large increase in *Synechococcus* and picoeukaryotes (see Figure 4.2). *Prochlorococcus* biomass decreases to nearly zero in the higher latitudes.

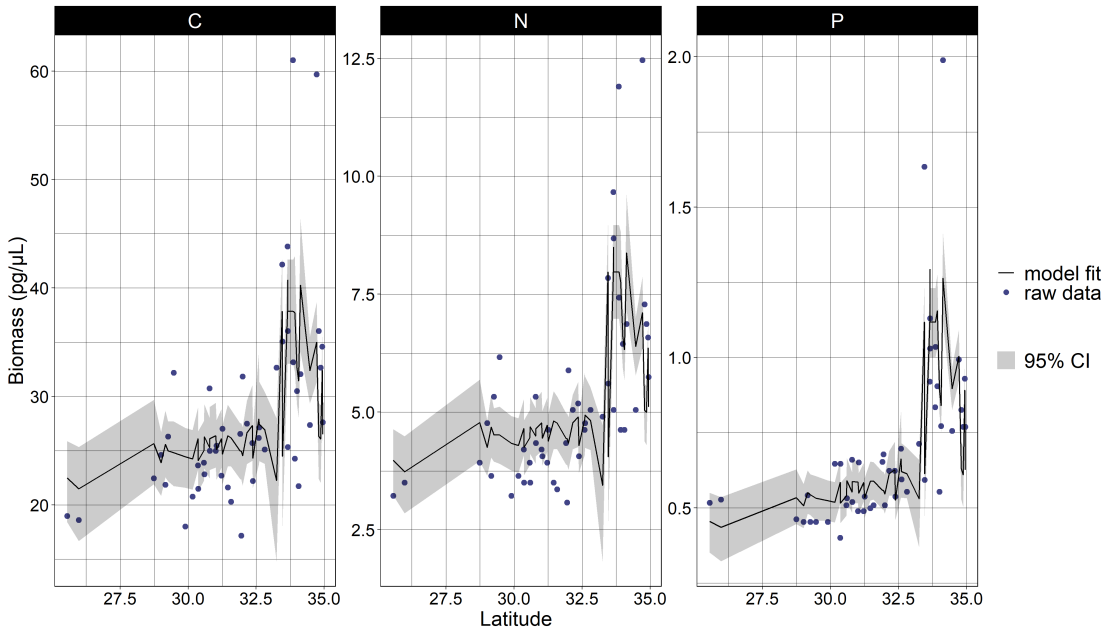

Figure 4.1: C, N, P quota models fit to data. Grey area represents a 95% credible region on the posterior mean fit.

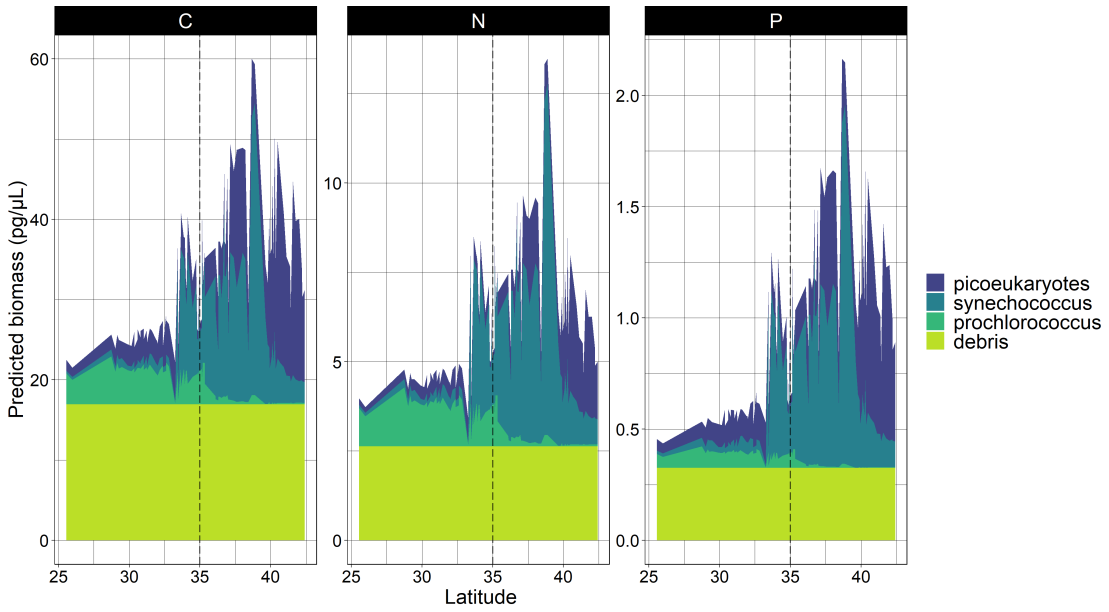

Figure 4.2: Predicted biomass for carbon, nitrogen and phosphorus, broken down by contribution from phytoplankton groups and debris. The model was trained on data south of 35°N latitude (dashed line).

	pro	95% CR	syn	95% CR	pico	95% CR	debris	95% CR
C	0.1476	0.0060–0.4001	0.1267	0.0779–0.1739	0.1087	0.0051–0.2965	0.6170	0.2648–0.8435
N	0.2207	0.0152–0.5252	0.1727	0.1234–0.2242	0.0841	0.0030–0.2515	0.5226	0.1513–0.7853
P	0.1040	0.0034–0.3186	0.2079	0.1666–0.2507	0.1802	0.0167–0.3935	0.5080	0.1875–0.7446

Table 4.1: Average contribution (%), of each phytoplankton group and debris, to total predicted biomass for carbon, nitrogen and phosphorus. Rows sum up to ~ 1.

	$fg\ C/um^3$	95% CR	$fg\ N/um^3$	95% CR	$fg\ P/um^3$	95% CR
pro	167.7	6.89–457.8	46.17	3.22–109.2	2.73	0.09–8.38
syn	538.4	307.2–771.4	145.62	98.98–194.8	23.90	18.44–29.38
pico	297.2	13.74–804.6	41.77	1.51–126.3	11.58	1.01–25.51

Table 4.2: Average mass of C, N, P per cubic micrometer of cell volume for each phytoplankton group.

	C:P	95% CR	N:P	95% CR	C:N	95% CR
pro	170.14	7.05–4343.00	41.22	2.32–1135.07	3.86	0.19–67.11
syn	58.18	32.30–90.02	13.46	8.72–19.87	4.31	2.24–7.39
pico	60.58	3.32–764.37	7.06	0.29–92.72	8.64	0.38–219.57
debris	133.88	52.48–405.97	17.91	4.71–58.84	7.47	2.91–27.51
total data	111.17		17.61		6.36	
Redfield	106		16		6.625	

Table 4.3: Median elemental ratios for each phytoplankton group, debris, and total data.

4.2 Discussion

Our linear models provide an estimate of biomass for each picoplankton group, describing how cell biomass changes with volume. However, we expect this linear relationship to break for cell volumes moving further away from the mean volume and outside the range of our data, especially since the ratio between the volume of the inside of the cell and cell membrane will change with size and these two components differ in composition.

Worden et al. [2004] propose 237 $fg\ C/\mu m^3$ as robust estimate for carbon biomass in *Prochlorococcus* and *Synechococcus* and potentially picophytoplankton in general. Our average estimate for *Prochlorococcus* and picoeukaryotes of 168 $fg\ C/\mu m^3$ and 297 $fg\ C/\mu m^3$ respectively, are relatively close to this esti-

mate. However, *Synechococcus* carbon content is estimated to be substantially higher at 538 $fg\ C/\mu m^3$. Such high values are not unheard of: Verity et al. [1992] find an average of 470 $fg\ C/\mu m^3$ for *Synechococcus* in their culture studies relating cell volume to carbon content. Bertilsson et al. [2003] find that cellular C quotas in *Synechococcus* are consistently higher when they are P-limited, which could be the case in these oligotrophic waters, so the high density of *Synechococcus* may in fact be a stress response. C:N:P ratios vary widely among phytoplankton groups and debris (Table 4.3), demonstrating the different resource requirements of each phytoplankton groups. Bertilsson et al. [2003] find elemental ratios ranging from $121 - 165$ for C:P and $21 - 33$ for N:P in nutrient replete conditions for *Prochlorococcus* and *Synechococcus* and even higher ratios in P-limited conditions. Our own C:P ratio for *Prochlorococcus* of 168 falls just at the edge of this range while its N:P ratio of 42.1 is also slightly above this range (see Table 4.3). In contrast, The C:P and N:P of *Synechococcus* (58.2 and 13.5, respectively) are well below these ranges. The particularly high densities for carbon, nitrogen and phosphorus of *Synechococcus*, paired with its low C:P and N:P ratios, indicate that *Synechococcus* is thriving in oligotrophic waters. Despite this, Figure 4.2 shows that up to about 33°N latitude, *Synechococcus* account for the smallest fraction of C,N and P biomass. This suggests that *Synechococcus* biomass is being controlled by some external factor other that nutrient availability.

As with abundance, biomass of *Synechococcus* and picoeukaryotes increases markedly in the higher latitudes, while *Prochlorococcus* biomass falls to baseline levels. A noteworthy difference between abundance and biomass for picoeukaryotes is that picoeukaryote carbon biomass plummets around 39° latitude (Figure 4.2), replaced mostly by a peak in *Synechococcus* and even a small peak in *Prochlorococcus*. Looking at the raw data, this seems to be caused by a combination of both a decrease in average volume and abundance. This drop in biomass may be slightly exaggerated though, as picoeukaryote volumes in that range are smaller than what

the model was trained on in lower latitudes. *Synechococcus* biomass may also be overestimated in the higher latitudes since our model is assuming the high C,N,P densities of *Synechococcus* in oligotrophic waters to be ubiquitous throughout the entire range of the study. In reality, it's not clear that *Synechococcus* would maintain its high elemental C,N,P quotas while competing with picoeukaryotes in high nutrient conditions. Lastly, it's possible that concentration of organic debris could change in the higher latitudes was well. In future studies, these issues could be addressed by measuring carbon, nitrogen and phosphate contents with samples filtered to the same size fraction as measurements gathered from flow cytometry.

5 Conclusion

Phytoplankton populations are controlled by a variety of complex interacting factors. Determining what affects their populations will be key in understanding how they will react to expected future changes in ocean biogeochemistry due to climate change. We conducted an analysis on a rich environmental dataset from a cruise in the North Pacific Ocean. Our goal was to model the distribution of the three dominant groups of photosynthetic picoplankton using just environmental conditions, i.e. "bottom-up" factors, and furthermore, determine their macromolecular compositions for carbon nitrogen and phosphorous content.

Our abundance models show the necessity of using non-linear methods to predict phytoplankton abundances. Using progressively more complex models demonstrated how much additional insight is gained from using non-linear and machine learning models compared to the simpler and easier to interpret linear models. Temperature and salinity are identified as the most important predictors and this result is supported later on by the non-linear models as well. However, GAMs clearly demonstrate that variance in picoplankton abundances are better explained using a non-linear relationship for temperature, salinity, and phosphate. Random forests had the additional advantages of having even more flexibility in modeling non-linear relationships and being able to include sparsely available nutrients data via imputation of missing data. Random forest models were thus able to show the lesser, albeit still important roles of nutrient concentrations that couldn't be identified with the previous models. Phosphate, iron, and manganese concentrations were the particularly important nutrients in explaining picoplankton abundances. Nutrient concentrations often have a step-wise relationship to abundance and biomass. These findings support previous research finding that phosphorus and iron concentrations are limiting factors controlling phytoplankton abundance in the North Pacific Ocean's oligotrophic gyre. These stepwise effects of nutrient concentrations offers insight the sharp changes in populations that

temperature doesn't fully account for. Picoplankton carbon biomasses maintain essentially the same relationship as thier corresponding abundances with environmental predictors. Even though carbon biomasses are an estimated value, which adds uncertainty to modeling them, they can be modeled just as well as abundances.

Estimates for total CNP content and C:P, N:P and C:N ratios for phytoplankton and organic debris were determined by Bayesian linear regression. These estimates varied considerably among the different phytoplankton groups and debris, but were all close to or within the range of previous studies' results. Estimates for carbon per unit volume for *Synechococcus* were particularly high while *Prochlorococcus* and picoeukaryotes estimates matched closely to known average carbon content. Typically, estimating carbon biomass from field samples relies on using cell volume to carbon conversion or average carbon per cell values based on controlled culture experiments. However, cellular elemental content and ratios in the different phytoplankton groups are known to have a fair degree of plasticity. Our elemental content models provide a relatively simple method to estimate cellular carbon, nitrogen, and phosphorus that relies directly on sample data rather than relying on reference values.

Abundance models comparing cruises show that although temperature and salinity are the most important predictors for modeling abundance, they are inadequate on their own to explain differences in populations between these cruises on different years. Having more data on nutrient concentrations from multiple datasets would be necessary to test whether or not differing nutrient conditions can account for the discrepancies in phytoplankton abundance between cruises. We had to limit our elemental quotas model to the oligotrophic region of the North Pacific due to abundance data not being able to account for larger cells in the highly productive northern latitudes. Abundance measurements and organic

matter measurements filtered to the same size fraction would allow for a greater latitudinal coverage of our model. Elemental C,N,P cellular content is expected to change along with resource availability so extending the range of our model would allow us to see how picoplankton elemental composition might change in high nutrient conditions.

6 Appendix

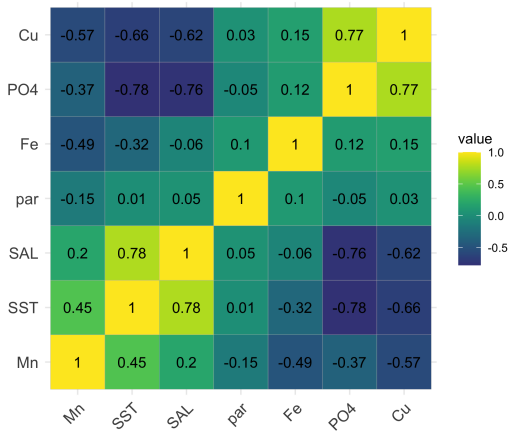

Figure 6.1: Correlation matrix of predictors from imputed data used in individual abundance random forests (Figure 3.4), abundance ratio random forests (Figure 3.7) and biomass random forests (Figure 3.11).

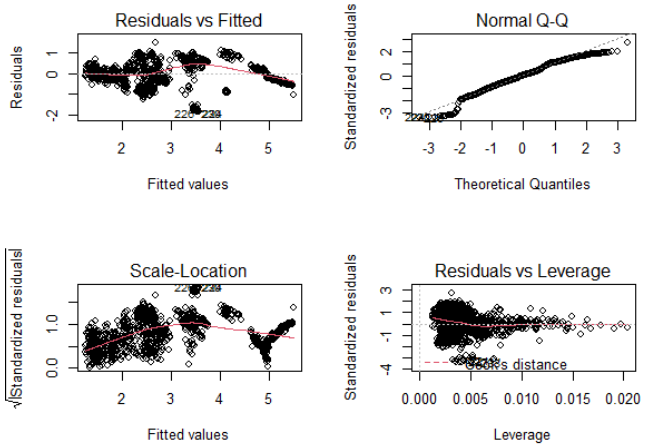

Figure 6.2: Diagnostic plots for *Prochlorochoccus* GAM model (Figure 3.2).

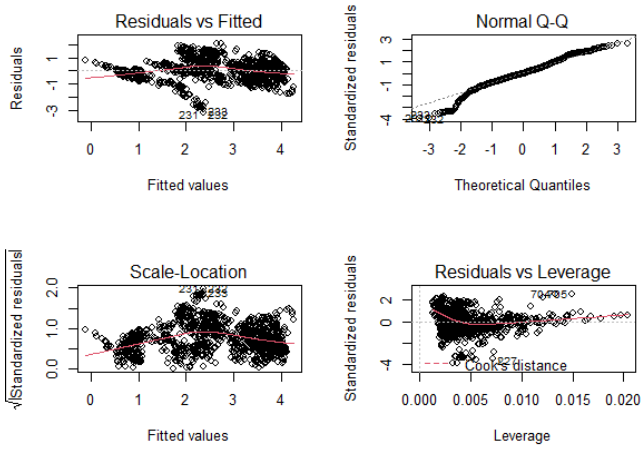

Figure 6.3: Diagnostic plots for *Synechococcus* GAM model (Figure 3.2).

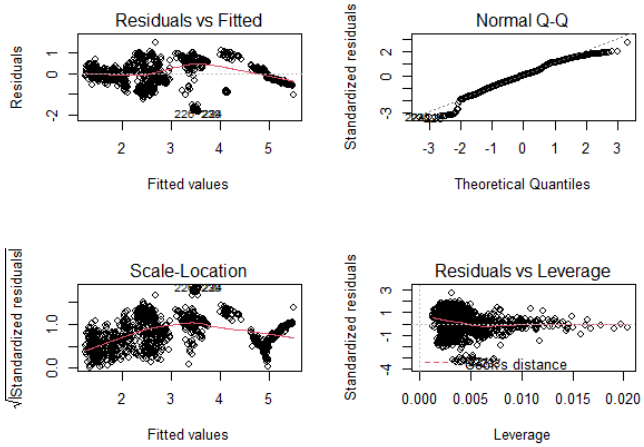

Figure 6.4: Diagnostic plots for picoeukaryotes GAM model (Figure 3.2).

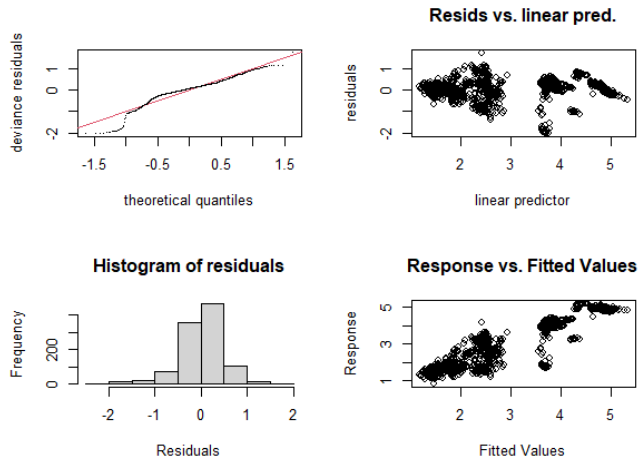

Figure 6.5: Diagnostic plots for *Prochlorochoccus* GAM model (Figure 3.2).

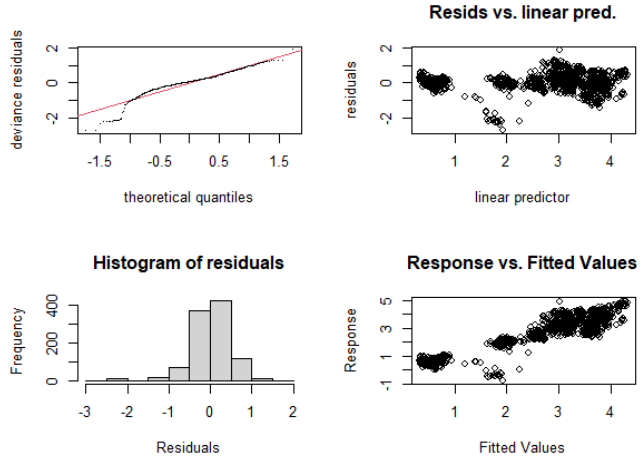

Figure 6.6: Diagnostic plots for *Synechococcus* GAM model (Figure 3.2).

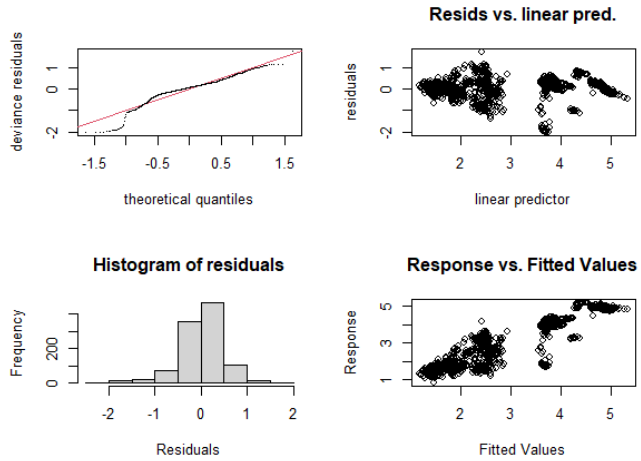

Figure 6.7: Diagnostic plots for picoeukaryotes GAM model (Figure 3.2).

	mean	se_mean	sd	2.5%	25%	50%	75%	97.5%	n_eff	Rhat
beta_pro_vol	0.014	0.00026	0.0103	0.00057	0.0057	0.012	0.020	0.038	1594	1.0007
beta_syn_vol	0.045	0.00022	0.0098	0.02557	0.0382	0.045	0.051	0.064	1949	1.0005
beta_pico_vol	0.025	0.00042	0.0183	0.00114	0.0103	0.021	0.036	0.067	1911	1.0007
intercept	1.410	0.01011	0.3609	0.58527	1.1833	1.458	1.680	1.978	1275	1.0018
sigma	0.652	0.00152	0.0717	0.52802	0.6020	0.647	0.697	0.810	2238	1.0006
lp__	-14.7	0.0547	1.77	-19.09	-15.58	-14.341	-13.44	-12.33	1041	1.0023

Table 6.1: Stan ouptut for carbon elemental quota model (Figure 4.1).

	mean	se_mean	sd	2.5%	25%	50%	75%	97.5%	n_eff	Rhat
beta_pro_vol	0.0033	0.000055	0.0021	0.00023	0.0017	0.0030	0.0047	0.0078	1425	0.9992
beta_syn_vol	0.0104	0.000040	0.0017	0.00707	0.0092	0.0104	0.0115	0.0139	1918	1.0026
beta_pico_vol	0.0030	0.000048	0.0024	0.00011	0.0011	0.0024	0.0042	0.0090	2502	1.0012
intercept	0.1878	0.001760	0.0629	0.05226	0.1453	0.1947	0.2349	0.2931	1275	0.9999
sigma	0.1152	0.000281	0.0130	0.09341	0.1059	0.1141	0.1230	0.1448	2142	1.00151
lp__	59.83	0.049608	1.793	55.6878	58.898	60.171	61.155	62.236	1306	0.9997

Table 6.2: Stan ouptut for nitrogen elemental quota model (Figure 4.1).

	mean	se_mean	sd	2.5%	25%	50%	75%	97.5%	n_eff	Rhat
beta_pro_vol	0.000088	0.0000019	0.000073	0.0000029	0.00003	0.00007	0.00013	0.00027	1458	0.9997
beta_syn_vol	0.000771	0.0000020	0.000088	0.0005955	0.00071	0.00077	0.00083	0.00095	2048	0.9996
beta_pico_vol	0.000374	0.0000051	0.000209	0.0000325	0.00022	0.00035	0.00051	0.00082	1692	1.0006
intercept	0.010502	0.0000963	0.003142	0.0037729	0.00845	0.01077	0.01279	0.01584	1065	0.9999
sigma	0.006059	0.0000138	0.000658	0.0049508	0.00559	0.00600	0.00647	0.00747	2286	1.0033
lp__	186.8941	0.0536948	1.775701	182.46531	185.943	187.284	188.194	189.239	1094	1.0029

Table 6.3: Stan ouptut for phosphorus elemental quota model (Figure 4.1).

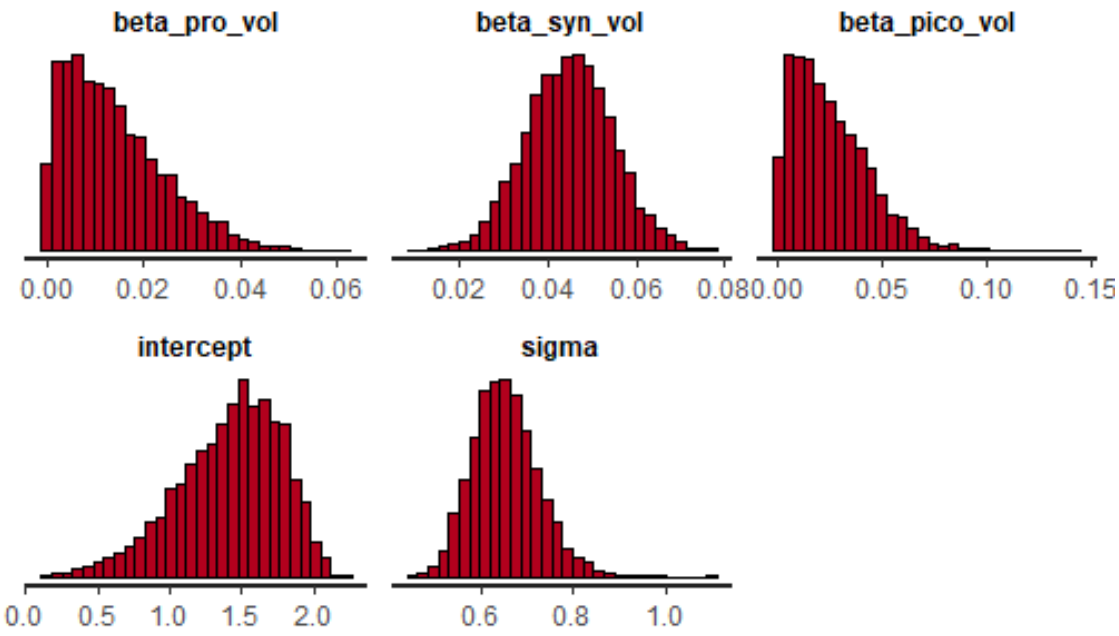

Figure 6.8: Posterior distributions of parameters (Table 6.1) for carbon elemental quota model (Figure 4.1).

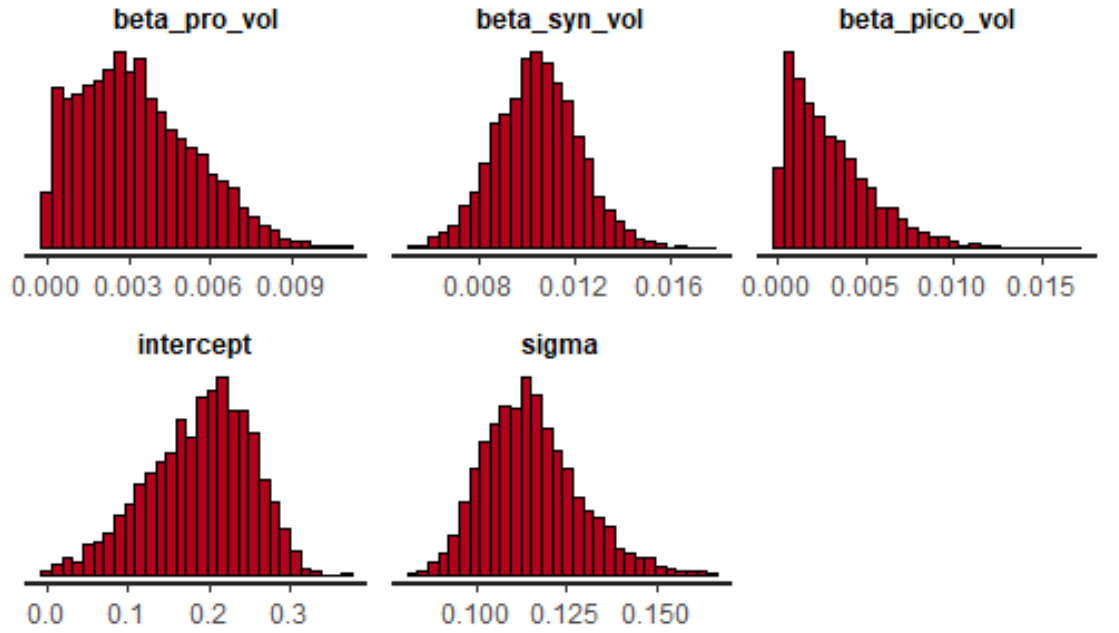

Figure 6.9: Posterior distributions of parameters (Table 6.2) for nitrogen elemental quota model (Figure 4.1).

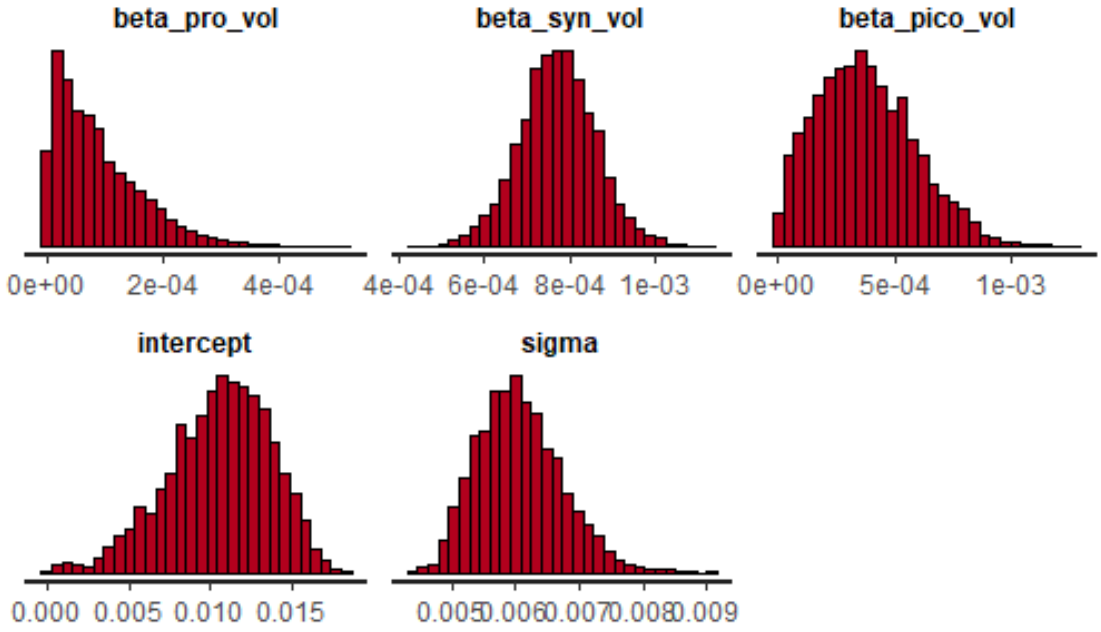

Figure 6.10: Posterior distributions of parameters (Table 6.3) for phosphorus elemental quota model (Figure 4.1).

References

P. Flombaum, J. L. Gallegos, R. A. Gordillo, J. Rincon, L. L. Zabala, N. Jiao, D. M. Karl, W. K. W. Li, M. W. Lomas, D. Veneziano, C. S. Vera, J. A. Vrugt, and A. C. Martiny. Present and future global distributions of the marine cyanobacteria *prochlorococcus* and *synechococcus*. *Proceedings of the National Academy of Sciences*, 110(24):9824–9829, 2013. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1307701110.

Claire S. Ting, Gabrielle Rocap, Jonathan King, and Sallie W. Chisholm. Cyanobacterial photosynthesis in the oceans: the origins and significance of divergent light-harvesting strategies. *Trends in Microbiology*, 10(3):134–142, 2002. ISSN 0966-842X. doi: 10.1016/S0966-842X(02)02319-3.

S. Bertilsson, O. Berglund, D. M. Karl, and S. W. Chisholm. Elemental composition of marine prochlorococcus and synechococcus: Implications for the ecological stoichiometry of the sea. *Limnology and Oceanography*, 48(5):1721–1731, 2003. ISSN 1939-5590. doi: 10.4319/lo.2003.48.5.1721.

Francois Ribalet, Jarred Swalwell, Sophie Clayton, Valeria Jiménez, Sebastian Sudek, Yajuan Lin, Zackary I. Johnson, Alexandra Z. Worden, and E. Virginia Armbrust. Light-driven synchrony of prochlorococcus growth and mortality in the subtropical pacific gyre. *Proceedings of the National Academy of Sciences*, 112(26):8008–8012, 2015. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1424279112.

Maria Lund Paulsen, Karen Riisgaard, T. Frede Thingstad, Michael St John, and Torkel Gissel Nielsen. Winter spring transition in the subarctic atlantic: microbial response to deep mixing and pre-bloom production. *Aquatic Microbial Ecology*, 76(1):49–69, 2015. ISSN 0948-3055. doi: 10.3354/ame01767. Publisher: Inter Research.

Lynda P. Shapiro and Elin M. Haugen. Seasonal distribution and temperature tolerance of synechococcus in boothbay harbor, maine. *Estuarine, Coastal and Shelf Science*, 26(5):517–525, 1988. ISSN 0272-7714. doi: 10.1016/0272-7714(88)90004-2.

Joseph L. Reid. On circulation, phosphate-phosphorus content, and zooplankton volumes in the upper part of the pacific ocean1. *Limnology and Oceanography*, 7(3):287–306, 1962. ISSN 1939-5590. doi: 10.4319/lo.1962.7.3.0287.

Jeffrey J. Polovina, Evan A. Howell, Donald R. Kobayashi, and Michael P. Seki. The transition zone chlorophyll front updated: Advances from a decade of research. *Progress in Oceanography*, 150:79–85, 2017. ISSN 0079-6611. doi: 10.1016/j.pocean.2015.01.006.

G. I. Roden. Biology, oceanography, and fisheries of the north pacific transition zone and subarctic frontal zone, 1991.

C. Guo, H. Liu, L. Zheng, S. Song, B. Chen, and B. Huang. Bottom-up and top-down controls on picoplankton in the east china sea. *Biogeosciences Discussions*, 10:8203–8245, 2013. doi: 10.5194/bgd-10-8203-2013.

C. Guo, H. Liu, L. Zheng, S. Song, B. Chen, and B. Huang. Seasonal and spatial patterns of picophytoplankton growth, grazing and distribution in the east china sea. *Biogeosciences*, 11(7):1847–1862, 2014. ISSN 1726-4170. doi: https://doi.org/10.5194/bg-11-1847-2014. Publisher: Copernicus GmbH.

C. M. Moore, M. M. Mills, K. R. Arrigo, I. Berman-Frank, L. Bopp, P. W. Boyd, E. D. Galbraith, R. J. Geider, C. Guieu, S. L. Jaccard, T. D. Jickells, J. La Roche, T. M. Lenton, N. M. Mahowald, E. Marañón, I. Marinov, J. K. Moore, T. Nakatsuka, A. Oschlies, M. A. Saito, T. F. Thingstad, A. Tsuda, and O. Ulloa. Processes and patterns of oceanic nutrient limitation. *Nature Geoscience*, 6(9):701–710, 2013. ISSN 1752-0894. doi: 10.1038/ngeo1765.

A. C. Redfield, B. H. Ketchum, and F. A. Richards. The influence of organisms on the composition of sea-water. *The sea: ideas and observations on progress in the study of the seas*, 1963. URL `http://www.vliz.be/en/imis?module=ref&refid=28944&printversion=1&dropIMIStitle=1`.

Justin D. Liefer, Aneri Garg, Matthew H. Fyfe, Andrew J. Irwin, Ina Benner, Christopher M. Brown, Michael J. Follows, Anne Willem Omta, and Zoe V. Finkel. The macromolecular basis of phytoplankton C:N:P under nitrogen starvation. *Frontiers in Microbiology*, 10, 2019. ISSN 1664-302X. doi: 10.3389/fmicb.2019.00763. Publisher: Frontiers.

M. Heldal, D. J. Scanlan, S. Norland, F. Thingstad, and N. H. Mann. Elemental composition of single cells of various strains of marine *prochlorococcus* and *synechococcus* using x-ray microanalysis. *Limnology and Oceanography*, 48(5):1732–1743, 2003. ISSN 1939-5590. doi: 10.4319/lo.2003.48.5.1732.

Kjell Gundersen, Mikal Heldal, Svein Norland, Duncan A. Purdie, and Anthony H. Knap. Elemental c, n, and p cell content of individual bacteria collected at the bermuda atlantic time-series study (BATS) site. *Limnology and Oceanography*, 47(5):1525–1530, 2002. ISSN 1939-5590. doi: 10.4319/lo.2002.47.5.1525.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2020. URL `https://www.R-project.org/`.

Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. ISBN 978-3-319-24277-4. URL `https://ggplot2.tidyverse.org`.

Yue Liang, Julie A. Koester, Justin D. Liefer, Andrew J. Irwin, and Zoe V. Finkel. Molecular mechanisms of temperature acclimation and adaptation in marine diatoms. *The ISME Journal*, 13(10):2415–2425, 2019. ISSN 1751-7370. doi: 10.1038/s41396-019-0441-9. Number: 10 Publisher: Nature Publishing Group.

Andrew J. Irwin and Zoe V. Finkel. Phytoplankton functional types: a trait perspective. *bioRxiv*, page 148312, 2017. doi: 10.1101/148312. Publisher: Cold Spring Harbor Laboratory Section: New Results.

S. N. Wood. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society (B)*, 73(1):3–36, 2011.

Andy Liaw and Matthew Wiener. Classification and regression by randomforest. *R News*, 2(3):18–22, 2002. URL `https://CRAN.R-project.org/doc/Rnews/`.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22, 2010. URL `http://www.jstatsoft.org/v33/i01/`.

Aleksandra Paluszynska, Przemyslaw Biecek, and Yue Jiang. *randomForestExplainer: Explaining and Visualizing Random Forests in Terms of Variable Importance*, 2020. URL `https://CRAN.R-project.org/package=randomForestExplainer`. R package version 0.10.1.

Christoph Molnar, Bernd Bischl, and Giuseppe Casalicchio. iml: An r package for interpretable machine learning. *JOSS*, 3(26):786, 2018. doi: 10.21105/joss.00786.

Christoph Molnar. *5.3 Accumulated Local Effects (ALE) Plot | Interpretable Machine Learning*. URL `https://christophm.github.io/interpretable-ml-book/ale.html`.

Daniel Apley and Jingyu Zhu. Visualizing the effects of predictor variables in black box supervised learning models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2020. doi: 10.1111/rssb.12377.

Andrew J. Irwin, Zoe V. Finkel, Oscar M. E. Schofield, and Paul G. Falkowski. Scaling-up from nutrient physiology to the size-structure of phytoplankton communities. *Journal of Plankton Research*, 28(5):459–471, 2006. ISSN 0142-7873. doi: 10.1093/plankt/fbi148. Publisher: Oxford Academic.

Andrew Gelman, Ben Goodrich, Jonah Gabry, and Aki Vehtari. R-squared for bayesian regression models. 73(3):307–309, 2019. ISSN 0003-1305. doi: 10.1080/00031305.2018.1549100. Publisher: Taylor & Francis.

Stan Development Team. RStan: the R interface to Stan, 2020. URL `http://mc-stan.org/`. R package version 2.21.2.

Barbara S. Reiher and Carl S. Huzzen. Some comments on the effective sample size of second order markov processes. 12(4):63–74, 1967. ISSN 0020-6024. doi: 10.1080/02626666709493551. Publisher: Taylor & Francis.

Benjamin S. Twining and Stephen B. Baines. The trace metal composition of marine phytoplankton. *Annual Review of Marine Science*, 5(1):191–215, 2013. doi: 10.1146/annurev-marine-121211-172322.

Elizabeth L. Mann, Nathan Ahlgren, James W. Moffett, and Sallie W. Chisholm. Copper toxicity and cyanobacteria ecology in the sargasso sea. *Limnology and Oceanography*, 47(4):976–988, 2002. ISSN 1939-5590. doi: 10.4319/lo.2002.47.4.0976.

Alexandra Z. Worden, Jessica K. Nolan, and B. Palenik. Assessing the dynamics and ecology of marine picophytoplankton: The importance of the eukaryotic component. *Limnology and Oceanography*, 49(1):168–179, 2004. ISSN 1939-5590. doi: 10.4319/lo.2004.49.1.0168.

Peter G. Verity, Charles Y. Robertson, Craig R. Tronzo, Melinda G. Andrews, James R. Nelson, and Michael E. Sieracki. Relationships between cell volume and the carbon and nitrogen content of marine photosynthetic nanoplankton. *Limnology and Oceanography*, 37(7):1434–1446, 1992. ISSN 1939-5590. doi: 10.4319/lo.1992.37.7.1434.