1. Introduction

Background.

The consequences of a changing climate are many and far-reaching. One consequence that has gained significant attention over the past decade is ocean deoxygenation, a decrease in levels of dissolved O_2. The impacts of ocean deoxygenation are only now becoming fully apparent: e.g. changes to ocean biogeochemistry, macro- and microorganism death due to hypoxia, and increased ocean production of nitrous oxide (N_2O), a greenhouse gas (Keeling et al., 2010).

Thus, we wanted to answer the questions: what factors predict ocean deoxygenation over time? Given these factors, are we able to accurately predict ocean oxygenation in new data? Oceanic levels of O_2 determine global populations' access to food and other resources, and the ability to predict this information may guide policies by (1) generating potentially-useful information on factors important to preventing or slowing deoxygenation, and (2) demonstrating potential future trends.

Data cleaning and preparation.

We are using data from the California Cooperative Oceanic Fisheries Investigations (CalCOFI), which we downloaded from Kaggle (link). CalCOFI's data represent the longest oceanographic time series in the world, spanning the years 1949 to 2016. Measured off the coast of California, the dataset includes larval and egg abundance data on 250+ species, as well as oceanographic data such as water temperature, salinity, chlorophyll, and oxygenation.

After downloading and importing the data, I cleaned and prepared it. Next, I restricted the dataset to observations after 2008. The original timeseries was quite long (1949-2016, with 800k+ observations) and would have significantly increased computational effort; this also provides a more current (i.e. approximately prior-decade) interpretation of results. For this same reason of computational efficiency, I also took a 10% random sample from the resulting dataset of ~46k, leaving me with ~4,600 observations.

Next, I removed from consideration features with large amounts of missing data. We had selected a preliminary list of predictors, but several of these had large proportions of data missing. Finally, for the 14 features that remained (see Models section for the final list), I removed any missing values for predictors.

After a final dataset was created for analytic purposes, training and testing datasets were created using `caret::createDataPartition()`, with 80 percent of data dedicated to training and 20 percent to testing. The training set was used for all subsequent analyses, including comparison of candidate models, with the testing set only used to evaluate the final chosen mode.

Using the training set, I explored outliers for all variables contained in the dataset. While some variables had extremely skewed distributions (e.g. right-skewed phaeophytin, ammonium, and nitrite concentrations), none had outliers visible from boxplots that warranted consideration of correction. Exploratory outlier visualizations are not shown in this document, but code for the boxplots is included in an appendix at the bottom of the submitted RMarkdown file.

2. Exploratory analysis/visualization

2.a. Correlation.

A correlation matrix was created (see Figure 1, Appendix) to estimate linear (Pearson's) correlation values among the outcome and all features. Several features were highly correlated, particularly the feature pairs temperature/potential density, SiO_3/PO_4, and PO_4/potential density; as well as the outcome-feature pairs O_2/PO_4 and O_2/SiO_3. Since correlation amongst predictors can cause problems, I was careful to consider these highly-correlated predictors in both model selection and interpretation.

2.b. Nonlinearity.

Scatterplots were plotted for all non-datetime features against the outcome (see Figure 2, Appendix). To assess potential nonlinear associations between features and outcome, a cubic spline smoother was plotted

with each scatterplot for visualization purposes. Nonlinear relationships with the outcome were apparent for some features, especially water temperature, salinity, potential density, phosphate, and chlorophyll. Since a significant number of features displayed potential nonlinearity in their relationships with the outcome, I considered this during model selection.

2.c. Timewise trends.

Finally, timewise trends in the outcome were explored on seasonal and long-term scales (results not shown). To explore seasonality, I fit a cyclic cubic spline smoother to outcome levels plotted by month of year, irrespective of year. No clear seasonal trends in the outcome were apparent. I also fit both linear and cubic spline smoother to explore the long-term trend in outcome across years, but again, no clear patterns were apparent. (Note: the current restricted timeseries is only 9 years long, so we shouldn't necessarily expect to detect any long-term trends.) Code for these timewise trend analyses is included in an appendix at the bottom of the submitted RMarkdown file.

3. Models

3.a. Candidate models.

For each candidate model, the 14 predictors included were as follows: year, month, latitude and longitude (decimals), temperature (°C), salinity (PSS-78 scale), potential water density (kg/m^3), pressure (decibars), phaeophytin (µg/L), chlorophyll (µg/L), phosphate (µmol/L), silicate (µmol/L), nitrite (µmol/L), and ammonium (µmol/L).

I began by fitting several models using `caret::train()` using my training data, and comparing them using their respective cross-validation (CV) root mean square error (RMSE) rates. Figure 3 (Appendix) details the cross-validated RMSE distribution for each model.

I began with a linear model fit with least squares as a baseline model. I then fit two slightly more flexible linear methods, for which the data were centered and scaled: elastic net (EN) and principal components regression (PCR). After these, I fit two nonlinear regression models: a generalized additive model (GAM), and a multivariate adaptive regression spline (MARS) model.

3.b. Tuning.

I tuned each candidate model's tuning parameter(s) using 5 repeats of 5-fold cross validation. The elastic net model was tuned over a grid of 11 `alpha` values ranging from 0 to 1 and 100 `lambda` values ranging from *exp*(-12) to *exp*(-4). The PCR model was tuned over a grid of `ncomp` (number of principal components to include) ranging from 1 to $p = 14$. For the GAM model, thin-plate penalized regression splines were fit to predictors with 10+ unique values, and spline degrees of freedom and feature selection were determined via generalized cross validation. The MARS model was tuned over a grid of `degree` (of interaction) values from 1 to 3 and `nprune` (number of terms to retain) from 2 to 35.

Final candidate models were then compared using CV RMSE distributions, and the final model was selected using the model with the lowest median CV RMSE. More details can be found in the submitted RMarkdown file.

3.c. Final Model.

The final model chosen according to RMSE was the MARS model (see Figure 3). The final model selected 8 of the 14 predictors that were initially included in the candidate models, a decision made by cross validation (see Figure 4 for selected features).

MARS is an adaptive regression method using piecewise linear basis functions. It is implemented here via the `earth` package (within `caret::train()`). These models are nonparametric and thus make no assumptions about the underlying distributions of data. This proved useful in this case due to observed potential nonlinear relationships between exposures and outcome.

3.c.i. Test error rate. Upon evaluating the MARS model using my test data, the test RMSE rate was equal to 8.9550675.

3.c.ii. Important variables. The most important variable in the MARS model, determined by reduction in GCV error upon variable addition to model (Figure 4), was PO_4. Since PO_4 is highly negatively correlated with the outcome, oxygenation, we might be able to claim that as PO_4 increases in seawater, O_2 decreases on average. A list of coefficients (Table 1) also shows that the model fit interaction terms between PO_4 and salinity. This can be explored using partial dependence plots (Figure 5): both predictors have a negative association with oxygenation, but in the 3D dependence plot, we see that deoxygenation may be worst when both salinity and phosphates are high. (These plots are for average trends, however, and may not be true across values of other predictors.)

The phosphate-oxygen relationship has been noted in previous literature, in which increased phosphorus levels lead to increased ocean productivity, which increases oxygen demand and thus decreases oxygen levels (e.g. Watson et al., 2017). However, since this variable was highly correlated with other variables (see Figure 1), the importance of other variables could be obscured relative to PO_4's importance.

3.c.iii. Interpretation and limitations. While a limitation of MARS models is somewhat limited interpretability, especially for relationships between individual predictors and outcome, I believe a strength of this MARS model is its consideration of both nonlinear and interaction effects. From exposure scatterplots, we can be fairly confident that there were nonlinear relationships present, and I determined via GCV that including 3rd order interaction terms was best for the MARS predictive ability. Thus, I think that model complexity was sufficient to capture underlying truth.

4. Conclusions

4.a. Summary of findings.

In conclusion, I found that a MARS model with third-degree interaction terms and 8 selected features provided the best overall prediction of seawater oxygenation, given the 14 predictors initially included in candidate prediction models. Of these predictors, PO_4 levels were highly predictive of the outcome, which is logical given previous literature. Other important predictors may include salinity, water temperature, and chlorophyll, but interpretation is somewhat limited given highly correlated predictors. Similar models may be useful in several cases: (1) if future predictions of seawater oxygenation are desired given trends in other factors; and (2) if oxygen measurements are unavailable but other relevant predictor data are.

4.b. Limitations of findings.

These data were exclusively gathered in the Pacific Ocean off the coast of California; thus, these prediction models and any interpretations may not be generalizable to data outside of this geographic area. Additionally, I limited my analyses to data from 2008-2016, which may limit temporal interpretation.

References

Keeling RF, Kortzinger A, Gruber N (2010). Ocean Deoxygenation in a Warming World. Annu Rev Marine Sci, 2:463–93.

Watson AJ, Lenton TM, Mills BJW (2017). Ocean deoxygenation, the global phosphorus cycle and the possibility of human-caused large-scale ocean anoxia. Philos Trans A Math Phys Eng Sci, 375(2102).

Appendix: Figures & Tables

Figure 1. Correlation matrix of outcome and features

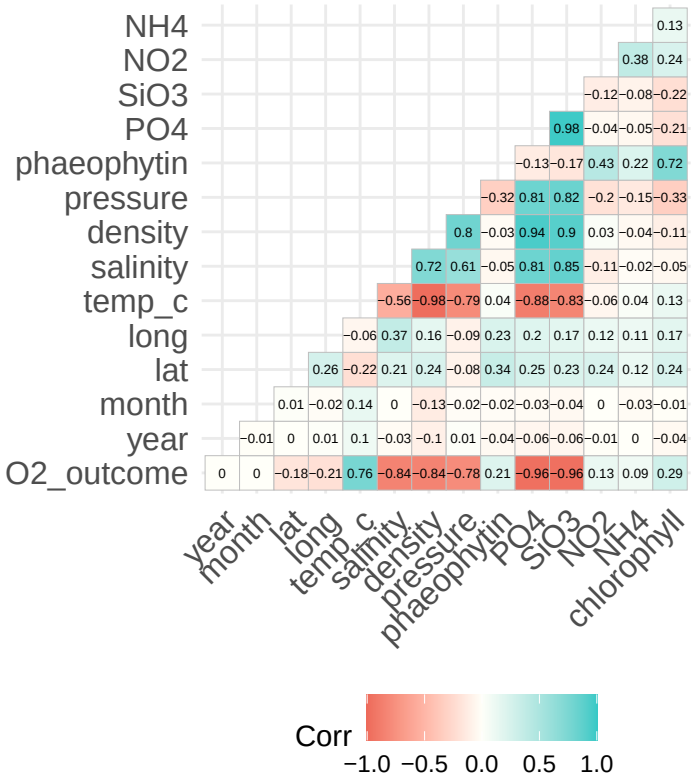


Figure 2. Ocean oxygenation level, by select predictors

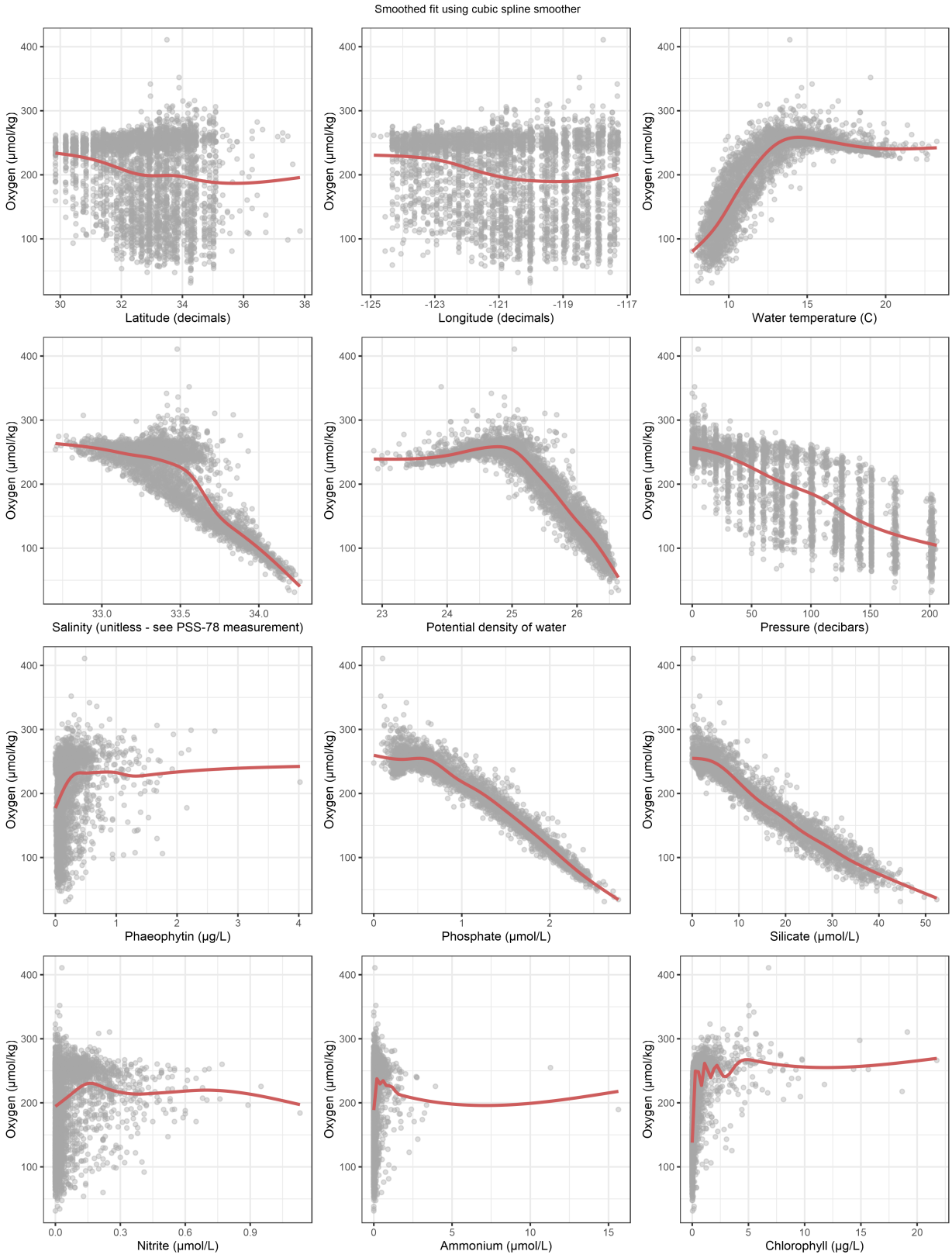


Figure 3. Cross−validated RMSE distribution, by model

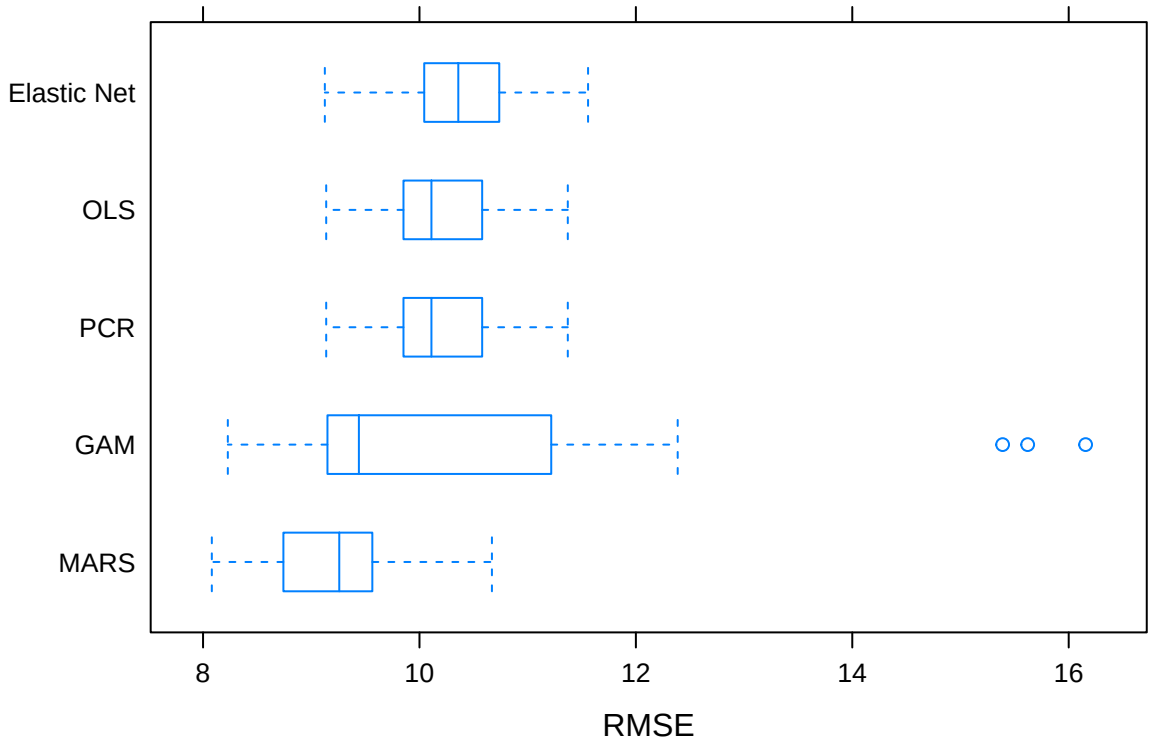


Figure 4. Final MARS model: variable importance

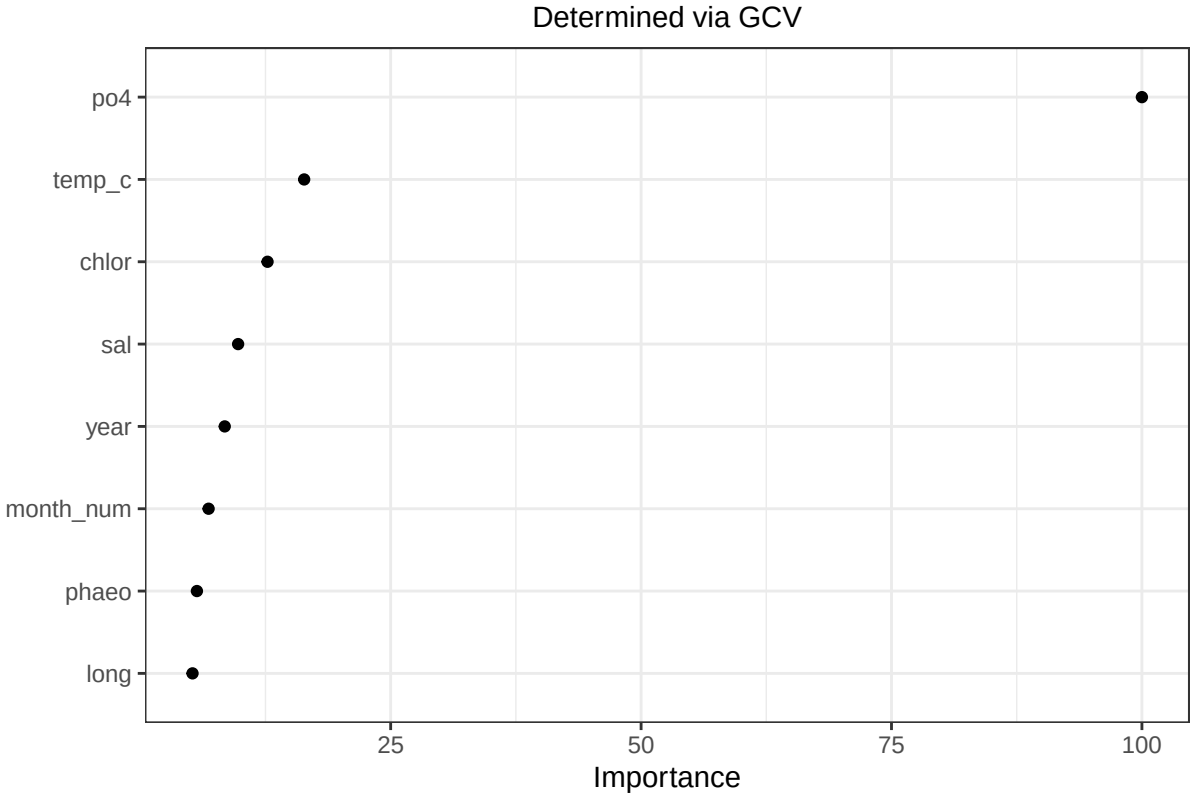
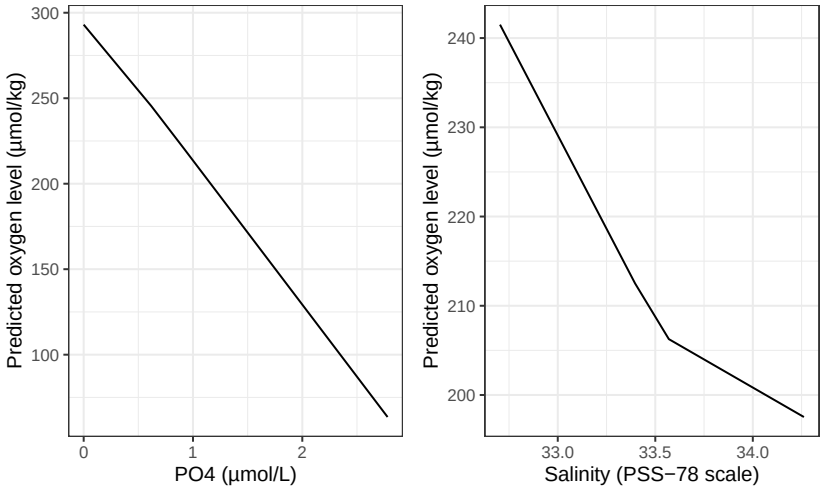


Figure 5. Partial dependence plots on outcome oxygenation
Predictors PO4 and salinity



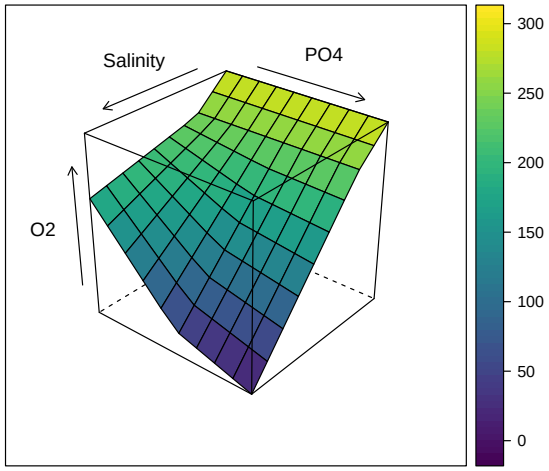


Table 1: Selected features and estimated coefficients, MARS model

Predictor	Estimate
(Intercept)	240.8830657
h(po4-0.66)	-94.4049168
h(0.66-po4)	76.9906609
h(temp_c-17.69)	-3.0008658
h(17.69-temp_c)	4.7598381
h(chlor-1.49)	1.9073204
h(1.49-chlor)	-17.7609290
h(sal-33.535) * h(po4-0.66)	-25.8497208
h(33.535-sal) * h(po4-0.66)	85.9116722
h(month_num-8)	-2.6758113
h(8-month_num)	-0.7728577
h(17.69-temp_c) * h(phaeo-0.09)	-2.0945926
h(17.69-temp_c) * h(0.09-phaeo)	11.8390133
h(year-2015)	-3.4954888
h(2015-year)	0.6580979
h(long- -120.143) * h(17.69-temp_c) * h(phaeo-0.09)	-0.8278474
h(-120.143-long) * h(17.69-temp_c) * h(phaeo-0.09)	1.9051867