# Using In-Situ Carbonate Chemistry Data to Validate Performance of Oceanographic Data Products and Investigate Biological Impacts of Ocean Acidfication

**Spring 2025 Interim Report**

## 1 Introduction

Since the start of the Industrial Revolution, atmospheric carbon dioxide levels have spiked upward due to practices like deforestation and the burning of fossil fuels, causing an ever increasing amount of carbon dioxide to dissolve into the ocean. Once dissolved, carbon dioxide undergoes a series of chemical equilibrium reactions that ultimately lowers the pH of the water in a process known as ocean acidification. This process has numerous ramifications on marine ecosystems, including but not limited to the impairment of specific pH dependent biological processes and the leaching of calcium carbonate required to build the shells and skeletons of many marine species.[1]

This project is sponsored by California Cooperative Oceanic Fisheries Investigations (CalCOFI), an organization founded in 1949 to study the ecological aspects of the Pacific sardine collapse off of the coast of California. CalCOFI is committed to studying California's coastal marine environment and collecting relevant oceanographic data in order to provide insight on important climate change related topics such as renewable energy, integrated ocean management, and marine spatial planning.[2]

We aim to extend the research done in "A 37 year record of ocean acidification in the Southern California current" by Wolfe et al.[3] on the yearly rate of change of certain carbonate chemistry variables by examining all available carbonate chemistry data collected across CalCOFI observation stations rather than only surface data collected at station 90.90, and study the impact of ocean acidification on zooplankton and krill biovolumes.

## 2 Problems of Interest

The goals of this project can be split into two core parts: the analysis of ocean carbonate chemistry variables, and the study of zooplankton and krill biovolumes off of the California coast.

For the carbonate chemistry portion, it is of interest to examine how important ocean carbon chemistry and oceanographic variables, namely total alkalinity ($TA$), total dissolved inorganic carbon ($DIC$), the Revelle Factor, $pH$, $pCO_2$, Omega aragonite ($\Omega_{aragonite}$), Omega calcite ($\Omega_{calcite}$), temperature, salinity, and $CO^{2-}$. Additionally, we wish to assess the performance of Empirical Seawater Property Estimation Routines (ESPER)[4] in predicting carbonate chemistry variables across different depths with easy to collect oceanographic variables such as temperature and salinity as inputs.

[1] Ocean acidification: https://www.nature.com/scitable/knowledge/library/ocean-acidification-25822734/
[2] CalCOFI report: https://calcofi.org/downloads/publications/CalCOFI3YrReports/CalCOFI_Review_2017_2021.pdf
[3] Wolfe Paper: https://www.nature.com/articles/s43247-023-01065-0
[4] ESPER: https://github.com/BRCScienceProducts/ESPER

Regarding zooplankton and krill biovolumes, we want to conduct a cross-comparison of the hydrographic and biological datasets to assess the amount of overlap spatially and temporally. Using co-located measurements, we wish to model the effects of ocean acidification, using the carbonate chemistry and oceanographic variables aforementioned, on zooplankton and krill abundance. In addition, we are interested in exploring how pH and related environmental factors affect the abundance of calcifying versus non-calcifying species.

# 3 Materials and Methods

## 3.1 Oceanographic and Carbonate Chemistry Data[5]

CalCOFI samples from a predetermined sampling grid off the coast of California on a quarterly basis. Typical stations are set 40 nautical miles apart. At each sampling point, identified by a station and line number, CalCOFI lowers a carousel of 24 bottles into the water, which collect seawater samples from around 20 different depths (typically ranging from 20 to 515 meters). Researchers on the ship then measure oceanographic values such as the temperature, salinity, macronutrient concentration and other properties of these samples. This results in the oceanographic dataset used in this study.

While oceanographic data is measured on every CalCOFI cruise, carbonate chemistry values such as TA and DIC are only occasionally measured from the collected water samples and are stored in their own dataset. We merged these two datasets through a left join on the carbonate chemistry dataset to create merged_bottle_data.csv.

## 3.2 Biological Data

There are three datasets that we are focusing on for the biological data: (1) CalCOFI NOAA Zooplankton Volume, (2) BTEDB (Krill) Abundances, and (3) PRPOOS Data (for Zooplankton Calcifiers/Non-Calcifiers Abundance). The zooplankton and krill biovolume data are obtained using net tows (Bongo and/or Pairovet) at each standard CalCOFI station. The PRPOOS data is also obtained using a net tow but rather than sampling at station, it is conducted during transits between stations. These three datasets have each been merged with the bottle data to create zoop_data/zooplankton_pH.csv, krill_data/CV_merged_krill.csv, and PRPOOS/prpoos_summary.csv, respectively.

### 3.2.1 CalCOFI NOAA Zooplankton Volume[6]:

The zooplankton biovolume data measures the amount of "plankton" (the small and microscopic organisms floating in the sea, consisting chiefly of diatoms, protozoans, small crustaceans, and the eggs and larval stages of larger animals) in the volume of sea water sampled. In particular, we are interested in the variables `total_plankton` and `small_plankton`.

### 3.2.2 BTEDB (Krill) Abundances[7]:

The krill dataset provides information on krill abundance from the Brinton and Townsend Euphausiid Database (BTEDB). The samples collected includes species such as *Euphausia pacifica*, *Nematoscelis difficilis*, and *Thysanoessa spinifera*, with individuals categorized by size and developmental phase (e.g., calyptopis, furcilia, juvenile, adult).

[5] Bottle Database: https://calcofi.org/data/oceanographic-data/bottle-database/
[6] CalCOFI Zooplankton Volume Database: https://oceanview.pfeg.noaa.gov/erddap/tabledap/erdCalCOFIzoovol.html
[7] BTEDB (Krill Volume) Data: https://portal.edirepository.org/nis/mapbrowse?packageid=knb-lter-cce.313.1

#### 3.2.3 PRPOOS (Calcifiers/Non-Calcifiers)[8]:

The PRPOOS (Planktonic Rate Processes in Oligotrophic Ocean Systems) dataset contains abundance and estimated biomass values for various zooplankton taxa, which can be categorized into calcifying and non-calcifying groups. The calcifying taxa are defined as *byrozoan larvae*, *pteropoda heteropoda*, *ostracods*, and *rhizaria*; the remaining taxa are considered non-calcifying.

### 3.3 Carbon Chemistry Methods

#### 3.3.1 ESPER Model Validation

A mixture of qualitative and quantitative analyses were applied to assess ESPER model performance. To determine the best performing model, we compared the RMSE and relative errors across models. To identify areas of weakness in ESPER's predictive ability, we generated plots of predictions and/or residuals against the model's inputs. Finally, we performed t-tests of the residuals to ascertain whether a statistically significant bias in the model predictions exists. Specifically, given that $\mu_\epsilon$ is the mean residual value for a particular model's predictions, we tested the following hypotheses:

$$H_0 : \mu_\epsilon = 0 \quad \text{and} \quad H_a : \mu_\epsilon \neq 0 \tag{1}$$

To account for alpha inflation from multiple testing, we applied the Benjamini-Hochberg procedure to adjust our obtained p-values.

#### 3.3.2 Ocean Acidification Trend Analysis

CO2SYS, a program used to mechanistically calculate other carbonate chemistry values given oceanographic values and at least two carbonate chemistry values as input[9], was used on the merged bottle dataset to obtain values of $pH$, the Revelle Factor, $pCO_2$, $CO_3^{2-}$, Omega aragonite, and Omega calcite for each observation. These values were then merged with the bottle data set to create merged_bottle_co2sys.csv. The values of $pH$, the Revelle Factor, $pCO_2$, $TA$, $DIC$, temperature, salinity, Omega aragonite, Omega calcite, and $CO_3^{2-}$ were then seasonally detrended using the procedure recommended in "Advancing best practices for assessing trends of ocean acidification time series" by Sutton et al[10]. Since the relationship between depth and different carbonate chemistry variables can be hard to model, and since we would expect that there is an interaction between time trends and depth, we have decided to limit our study to observations with depths of 20 meters or less for the time being. Observations from stations with less than 20 observations were then filtered out in order to ensure more accurate model fits. Linear regression models were then fit for each variable of interest against time at each station. The Benjamini-Hochberg procedure was used to adjust p-values to account for multiple testing. Additionally, a mixed effects model was fit to each variable of interest regressed against time, depth, and with a random intercept for station. Maximum likelihood estimation was used to estimate the parameters for every model.

### 3.4 Biological Systems Methods

The main methods we have implemented so far are a combination of linear models and spatial models. Our initial model consisted of the carbonate chemistry variables of interest as predictors (TA, DIC, temperature, and salinity) with the appropriate abundance variables as our response variables. To build on these models, we included more variables from the CO2SYS output, such as pH, and performed LASSO regression to identify any other potentially significant predictors. For spatial modelling, we used a spatial mesh model as well as GAM with spatial splines to incorporate the effects of latitude and longitude on the generalized linear models.

[8] PRPOOS (Zooplankton Calcifiers/Non-Calcifiers Volume) Zooscan Database: https://oceaninformatics.ucsd.edu/zooscandb/
[9] CO2SYS: https://www.ncei.noaa.gov/access/ocean-carbon-acidification-data-system/oceans/CO2SYS/co2rprt.html
[10] Seasonally Detrend: https://www.frontiersin.org/journals/marine-science/articles/10.3389/fmars.2022.1045667/full

# 4 Preliminary Findings

## 4.1 Carbonate Chemistry Analyses

### 4.1.1 ESPER Model Validation

All three ESPER models show comparable performance when predicting in-situ values for TA and DIC (Table 1). Model performance is improved (as expected) when the full set of input variables is used, as opposed to predicting based on temperature and salinity. The metrics also show superior performance of the models when predicting TA over DIC values.

ESPER Error Metrics

| | | RMSE[1] | | Median Error[1] | | Mean Error[1] | | Error SD[1] | |
|---|---|---|---|---|---|---|---|---|---|
| Model | Input[2] | TA[3] | DIC[3] | TA | DIC | TA | DIC | TA | DIC |
| LIR | lim | 10.39 | 28.01 | −0.07% | −0.25% | −0.02% | −0.36% | 0.44% | 1.32% |
| | all | 10.18 | 15.44 | −0.01% | −0.08% | 0.03% | −0.16% | 0.43% | 0.73% |
| Mixed | lim | 10.31 | 23.25 | −0.09% | −0.15% | −0.05% | −0.27% | 0.44% | 1.10% |
| | all | 9.89 | 14.01 | −0.06% | −0.06% | −0.02% | −0.13% | 0.42% | 0.67% |
| NN | lim | 10.44 | 22.47 | −0.13% | −0.17% | −0.09% | −0.27% | 0.44% | 1.06% |
| | all | 10.08 | 13.64 | −0.12% | −0.09% | −0.08% | −0.14% | 0.42% | 0.64% |

[1] Relative error, i.e. (Predicted - Observed)/Observed

[2] lim refers to ESPER calculations performed using only temperature and salinity as predictors; all refers to calculations using all six input variables

[3] TA SD = 33.95128; DIC SD = 117.10278

Table 1: RMSE, median and mean percent error for the three ESPER models. Performance is comparable between models and all models show a slight bias in their predictions.

Overall, all ESPER models have a small but statistically significant bias when predicting TA and DIC values using the CalCOFI bottle data, as shown in Table 2, which displays the p-values for the hypothesis test described above in Equation 1.

ESPER Residuals Hypothesis Testing Results (p-values)

| | TA | | DIC | |
|---|---|---|---|---|
| Model | lim | all | lim | all |
| Mixed | <1e-04 | 0.0093 | <1e-04 | <1e-04 |
| LIR | 0.0096 | <1e-04 | <1e-04 | <1e-04 |
| NN | <1e-04 | <1e-04 | <1e-04 | <1e-04 |

Table 2: Hypothesis testing results for the mean of the ESPER residuals (the bias, as described in Equation 1). The displayed p-values have been adjusted for alpha inflation using the Benjamini-Hochberg procedure.

Figure 1 displays the ESPER NN model predictions against observations for TA and DIC, with observations colored by their depth (in meters). Generally, the accuracy of ESPER predictions decreases at both the shallowest and deepest depths and extreme values of TA and DIC.

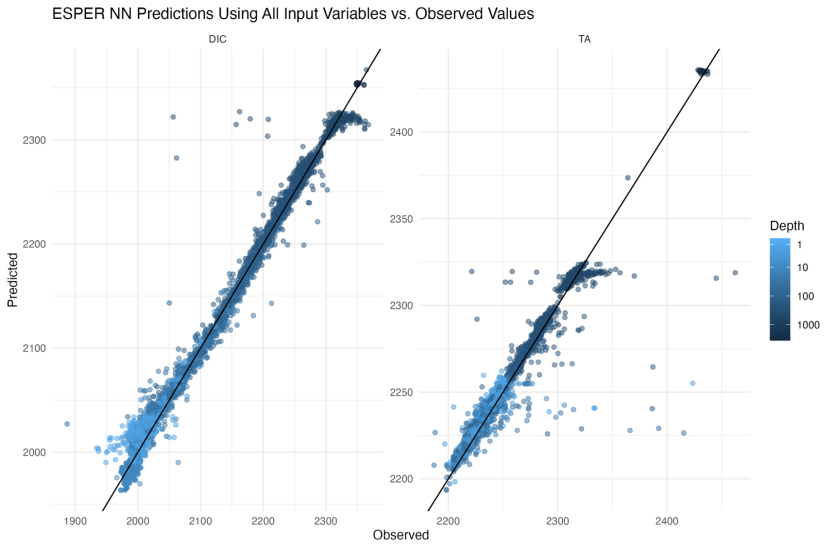


Figure 1: ESPER NN predictions plotted against observed values of TA and DIC. Perfect predictions are represented by the solid diagonal line in black (the further a point is from this line, the less accurate the prediction).

Figure 2 displays the ESPER NN model residuals against the six input variables to the model. Generally, the accuracy of ESPER predictions decreases for extreme values of oxygen concentration, phosphate, and silicate.

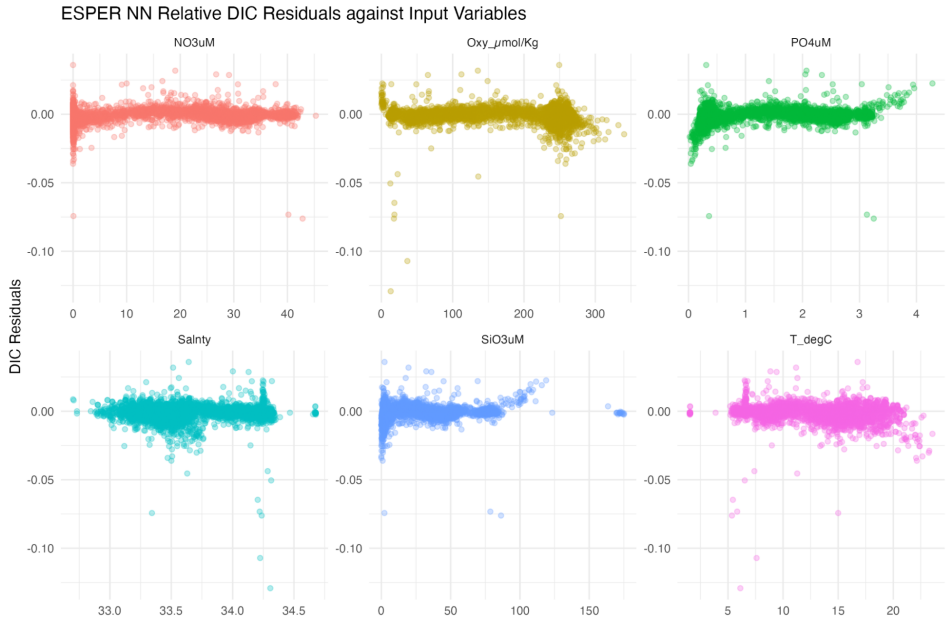


Figure 2: Plot of ESPER NN residuals against the six input variables to the model.

### 4.1.2 Ocean Acidification Trends

The regression results of the mixed effects model can be seen in Table 3 below. The slope, standard error, confidence intervals, and p-values refer to the time parameter estimated in the respective regression model. All carbonate chemistry and ocean acidification variables had a statistically significant time trend at the $\alpha = 0.05$ level, while both hydrography variables, temperature and salinity, were not found to have a significant temporal trend.

Surface Level Mixed Effect Regression Statistics for CalCOFI Stations with Depth Correction

| Parameter | Slope | Std. Error | 95% CI | Units | p-value | n | AIC | $r^2$ |
|---|---|---|---|---|---|---|---|---|
| Hydrography | | | | | | | | |
| Temperature | 0.0019 | 0.0075 | (-0.01287, 0.01667) | °C $yr^{-1}$ | 0.8006 | 860 | 2,925.2626 | 0.4353 |
| Salinity | 0.0012 | 0.0008 | (-0.00038, 0.00270) | $yr^{-1}$ | 0.1407 | 862 | −756.7924 | 0.3344 |
| Ocean acidification indicators | | | | | | | | |
| pH | −0.0013 | 0.0002 | (-0.00174, -0.00089) | $yr^{-1}$ | <1e-04 | 723 | −2,632.6786 | 0.2135 |
| $CO_3^{2-}$ | −0.4153 | 0.0873 | (-0.58664, -0.24350) | µmol $kg^{-1}$ $yr^{-1}$ | <1e-04 | 755 | 6,145.1207 | 0.2852 |
| $\Omega_{calcite}$ | −0.0100 | 0.0021 | (-0.01415, -0.00585) | $yr^{-1}$ | <1e-04 | 755 | 524.6576 | 0.2871 |
| $\Omega_{aragonite}$ | −0.0064 | 0.0014 | (-0.00909, -0.00365) | $yr^{-1}$ | <1e-04 | 755 | −114.1472 | 0.2940 |
| Seawater carbonate chemistry | | | | | | | | |
| $A_T$ | 0.1667 | 0.0677 | (0.03359, 0.29963) | µmol $kg^{-1}$ $yr^{-1}$ | 0.0141 | 810 | 6,414.0058 | 0.2145 |
| $C_T$ | 0.8242 | 0.1102 | (0.60737, 1.04070) | µmol $kg^{-1}$ $yr^{-1}$ | <1e-04 | 822 | 7,211.9176 | 0.3276 |
| $pCO2$ | 1.6448 | 0.2482 | (1.15769, 2.13296) | µatm $yr^{-1}$ | <1e-04 | 755 | 7,769.1251 | 0.2301 |
| Revelle Factor | 0.0194 | 0.0042 | (0.01121, 0.02756) | $yr^{-1}$ | <1e-04 | 723 | 1,426.1417 | 0.2966 |

Table 3: Mixed effects regression results for hydrography and carbonate chemistry variables regressed against time and depth. Shown values refer to the estimated parameter for time.

Figure 3 shows model results aggregated by station. Models with less than 10 observations used in model fitting were removed from this aggregation. Station 90.90 has the highest proportion of variables of interest with a significant temporal trend, and stations further from the coast tend to have a larger proportion of variables of interest with a significant temporal trend.

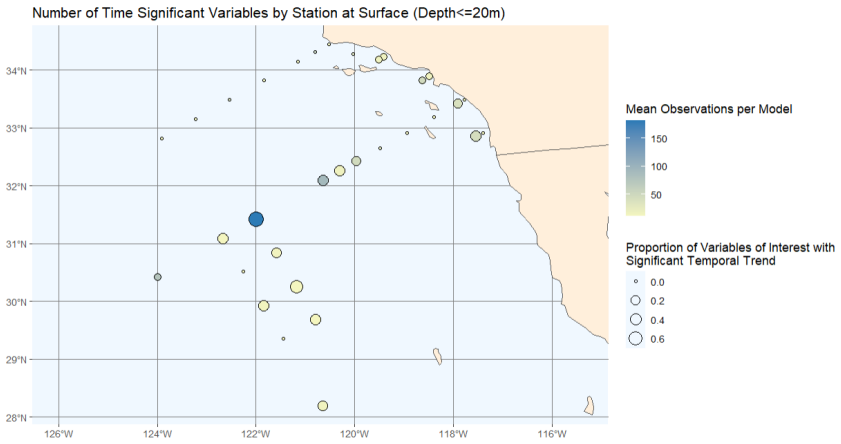


Figure 3: Map of aggregated model results by station.

## 4.2 Biological Data Analyses

### 4.2.1 Zooplankton Models: Spatial GAM

We created two spatial GAM models for log-transformed total plankton abundance and log-transformed small plankton abundance, respectively, where warmer colors indicate regions associated with lower abundance, while cooler yellow tones represent areas with higher predicted abundance. Both response variables use the smooth parameters pH, TA, DIC, salinity, temperature, year, month, and station ID and tensor product smooths between latitude and longitude.

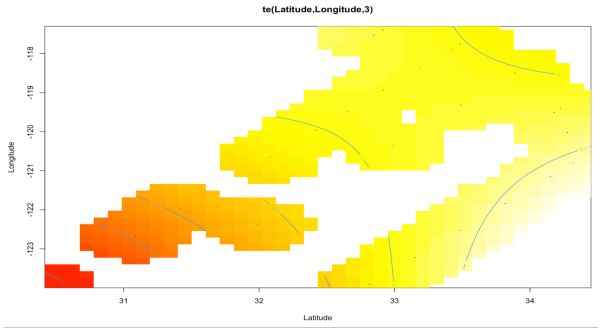


Figure 4: Spatial smooth from the GAM model showing the partial effect of latitude and longitude on log-transformed total plankton abundance.

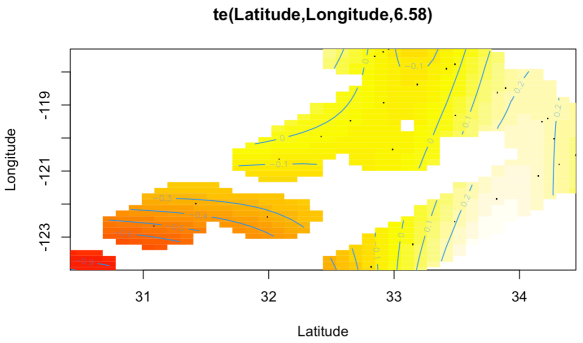


Figure 6: Spatial smooth from the GAM model showing the partial effect of latitude and longitude on log-transformed small plankton abundance.

For total plankton abundance, the GAM model reveals that spatial location, seasonality (month), and long-term temporal trends (year) are the strongest predictors, explaining 28% of the variability. Figure 4 shows higher

abundances are predicted along the eastern and southeastern coastal regions, while lower abundances appear offshore, especially in the southwestern region. Based on the model results, pH showed a slight positive relationship with abundance while other variables like TA, DIC, salinity, and temperature did not have significant effects in this model (see Figure 5 in Appendix).

For small plankton abundance, the GAM model explains approximately 35% of the variance, with significant effects from spatial location, pH, TA, temperature, year, and month. Figure 6 shows higher abundance predicted along the southern and eastern coastal regions, and lower abundance offshore to the southwest. The model shows that plankton abundance tends to increase with higher pH, suggesting potential sensitivity of small plankton to acidification. Higher TA is associated with decreased abundance, while temperature has a nonlinear effect. From the model summary, other variables such as DIC, salinity, and station ID showed no significant influence on abundance (See Figure 7 in Appendix).

#### 4.2.2 Krill Models

To investigate the spatial and environmental drivers of Thysanoessa gregaria occurrence, we applied two modeling strategies: a Random Forest classification model for presence/absence prediction, and a PCA + K-means clustering analysis to explore abundance patterns.

##### 4.2.2.1 Random Forest (Presence/Absence Classification)

```
Call:
 randomForest(formula = present ~ pHin + CTDTEMP_ITS90 + Salinity_PSS78 +      TA + DIC +
Depth + Month_UTC + Station_ID, data = df_model,      importance = TRUE, ntree = 500)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 2

        OOB estimate of  error rate: 15.38%
Confusion matrix:
    no yes class.error
no   3   8   0.7272727
yes  0  41   0.0000000
```

Figure 8: Random Forest model confusion matrix and call output for presence/absence classification of Thysanoessa gregaria.

In Figure 8, a random forest model was trained using a set of environmental predictors including pH, temperature (CTDTEMP_ITS90), salinity, total alkalinity (TA), dissolved inorganic carbon (DIC), depth, month, and Station_ID. The inclusion of Station_ID significantly enhanced model performance, reducing the out-of-bag (OOB) error rate to 15.38%, compared to a substantially higher error rate in the model without it.

Notably, the model achieved 100% accuracy for predicting presence ('yes'), correctly classifying all 41 presence cases. The confusion matrix indicates most misclassifications occurred in the absence category. Feature importance analysis (see Figure 9 in Appendix) revealed that Station_ID was the most influential variable by a wide margin, suggesting strong spatial dependence. This implies the model effectively "memorized" geographic patterns in T. gregaria distribution, highlighting that occurrence is highly dependent on location.

##### 4.2.2.2 PCA and K-means Clustering

To further understand abundance variation, we performed PCA on environmental variables followed by K-means clustering.


Figure 11: PCA biplot overlaid with K-means clustering results (k = 3).

The elbow plot (see Figure 10 in Appendix) shows a sharp decline is observed up to k = 3, suggesting three clusters as the optimal choice for grouping ecological sample points. In Figure 11, each cluster corresponds to a distinct grouping in environmental space, indicating different habitat types for T. gregaria. Cluster 1 has the highest mean abundance, while Cluster 3 contains only zero values, suggesting unsuitable habitat or timing. Each cluster reflected distinct habitat characteristics: Cluster 1 has a high mean log-abundance (3.10) which likely represents ideal environmental conditions, Cluster 2 has an intermediate abundance (2.58) which may correspond to marginal habitats or transitional seasonal conditions, and Cluster 3 has zero abundance which is likely unsuitable sites or times for T. gregaria presence (see Table 4 in Appendix).

These results confirm that T. gregaria abundance can be meaningfully grouped by environmental conditions and that site-specific factors strongly influence distribution.

### 4.2.3 PRPOOS Model: Spatial GAM

When modeling the difference between calcifiers and non-calcifers, we explored LASSO regression and spatial GAMS. Unfortunately, the performance of the LASSO regression was not optimal (see Tables 5-6 in Appendix) so we proceeded with spatial GAMs.

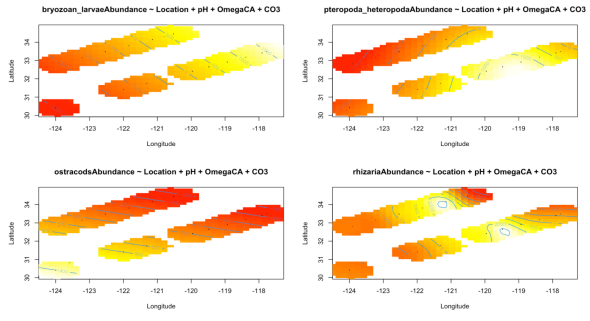


Figure 12: GAM model with spatial splines for each calcifying species.

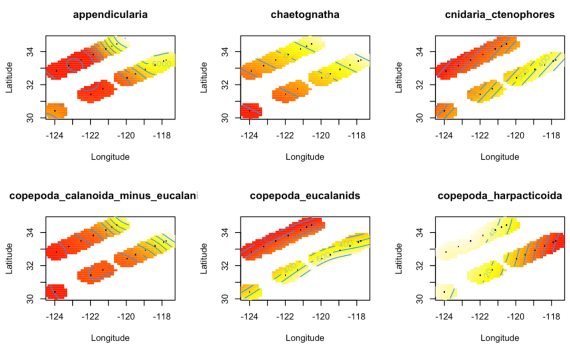


Figure 13: GAM model with spatial splines for other species (first 6 species are shown).

From Figures 12-13, we observe that the spatial patterns differ from species to species, but most models show a positive effect of location on abundance nearshore and a negative effect away from the coast. No obvious differences are found between calcifying and non-calcifying species. The models typically explain around 15% of the variance in the data.

# 5 Discussion

## 5.1 Carbonate Chemistry

The regression results from Table 3 generally agrees with the results from Wolfe et al., however the temporal effect sizes from our regression tended to be larger, and we additionally found that TA had a significant temporal trend, likely because we had much more data since our study is not just limited to station 90.90. Additionally, the map in Figure 3 potentially supports the idea that coastal stations with stronger upwelling may experience slower carbon uptake as stations further from the coast tended to have a higher proportion of variables of interest with a significant temporal trend. However, a formal hypothesis test is needed to more rigorously support this. The structure of the data has been a consistent struggle for this part of the project; there is a large gap from 2002 to 2007 in which no data was collected, and data collected in the late 1900's and early 2000's only collected measurements from surface water. Additionally, since the effect of depth on many carbonate chemistry variables and its potential interaction with the effect of temporal trends is hard to properly model, we have limited ourselves to examining surface level carbonate chemistry trends which has significantly cut down our total sample size.

## 5.2 Biological Data

A big takeaway is that there is spatial autocorrelation detected in the biological data, where abundance generally increases near coastal areas and decreases in offshore regions. We also observed a positive relationship with pH and abundance, which is consistent with expectations, as lower pH levels associated with ocean acidification can negatively impact species growth and survival. However, one of our biggest challenges that we have encountered is that the process of merging biological data leads to many observations being lost. As a result, the data we are working with tends to be sparse or biased towards certain stations and dates. For example, the krill dataset contains numerous missing values and zero abundance entries across taxa, making it difficult to identify patterns among different species and life stages. To address this, we aggregated the data into total abundance, however, this approach limits our ability to draw specific conclusions and instead supports broader generalizations. This makes it difficult to perform an effective regression on our variables of interest and evaluate the effect of these predictors. Merging the PRPOOS data with the carbonate chemistry data poses another challenge because the PRPOOS dataset does not contain a 'depth' column. We considered averaging observations over depth to account for this, but this approach masks the effect of depth on variables like temperature, salinity, and pH.

# 6 Future Work

## 6.1 Carbonate Chemistry

In the future, we aim to conduct a rigorous hypothesis test for coastal stations experiencing slower carbon uptake. Additionally, we want to find ways to get more insight from the results of the station by station regression results. Given sufficient time, we would like to explore ways to properly model depth with our variables of interest, and in particular examine whether depth and time have a statistically significant interaction for each variable of interest.

## 6.2 Biological Data

To improve model performance, we plan to incorporate additional nutrient-based and ocean chemistry variables, such as aragonite, pCO2, and chlorophyll, and metrics such as the depth of the saturation horizon. We aim to increase the proportion of explained variability to at least 50% of the variability so we can make more definitive predictions and clearer interpretations of how specific chemical factors influence biovolumes. Furthermore, we intend to use the depth of the saturation horizon to create vertical profiles of aragonite saturation values and examine how changes in water column chemistry impact calcifying/non-calcifying species, initially focusing on specific taxa and stations.

# 7 Appendix

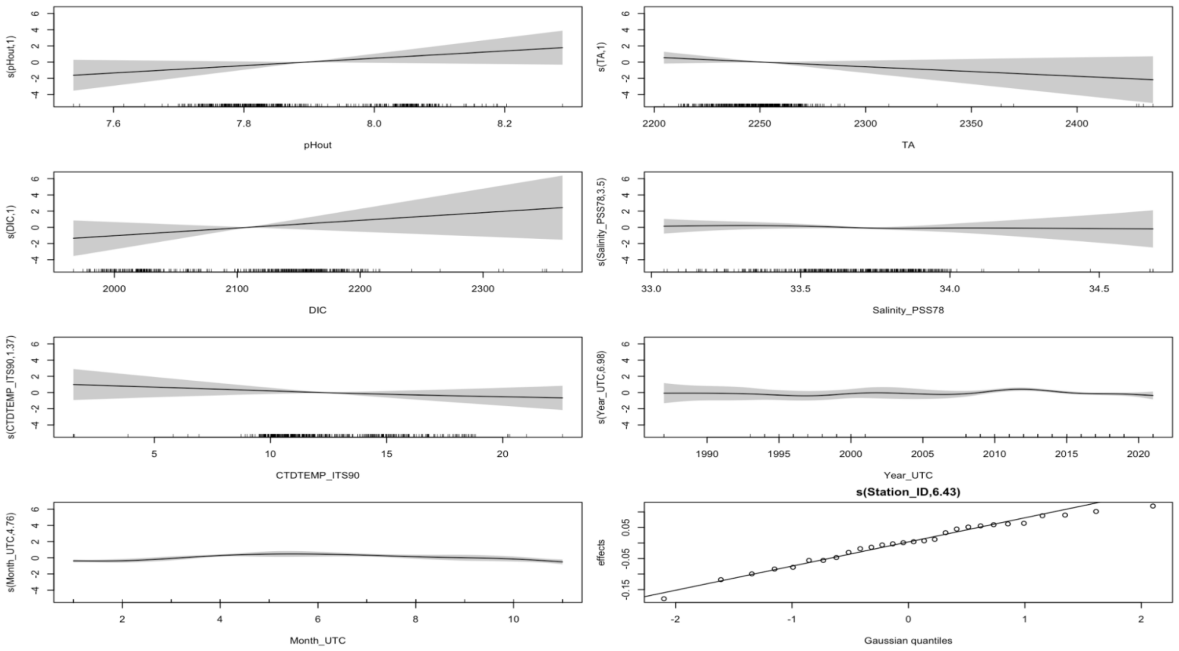


Figure 5: Partial effect plots for non-spatial smooth terms in the GAM model for log-transformed total plankton abundance. Shaded regions represent confidence intervals for each smooth.

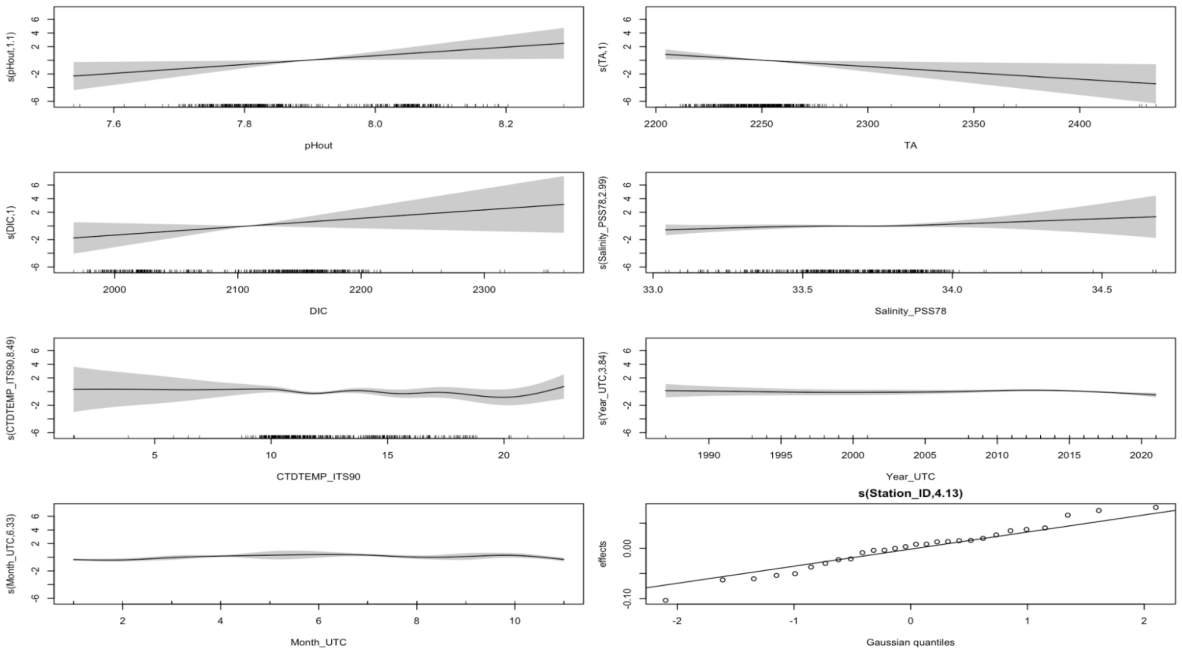


Figure 7: Partial effect plots for remaining GAM smooth terms on small plankton abundance.

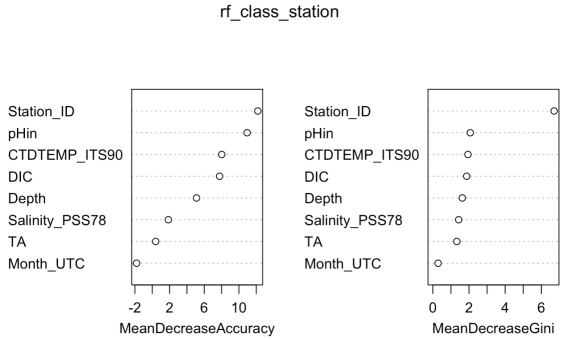


Figure 9: Variable importance plots based on Mean Decrease Accuracy and Gini Index from the Random Forest model.

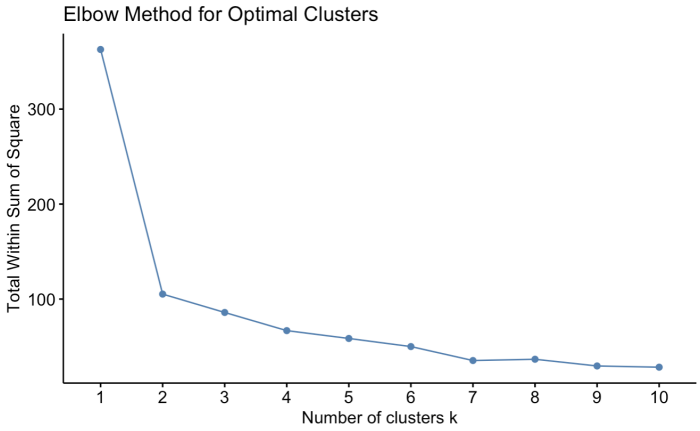


Figure 10: Elbow plot showing the total within-cluster sum of squares for k = 1 to 10.

A tibble: 3 × 2

| cluster <fctr> | mean_log_abundance <dbl> |
|---|---|
| 1 | 3.103321 |
| 2 | 2.581341 |
| 3 | 0.000000 |

3 rows

Table 4: Mean log-transformed abundance of Thysanoessa gregaria for each of the three clusters.

#### 7.0.0.1 PRPOOS Models: LASSO Regression Attempt

| Cacifiers Model Summary Table | | | | |
|---|---|---|---|---|
| Taxa | Term | Estimate | P-Value | Adjusted $R^2$ |
| rhizariaAbundance | (Intercept) | −105,049.609 | 0.353 | 0.039 |
| rhizariaAbundance | CO3_mean | −803.037 | 0.014 | 0.039 |
| rhizariaAbundance | OmegaCA_mean | 28,768.901 | 0.041 | 0.039 |
| rhizariaAbundance | pH_mean | 15,673.783 | 0.291 | 0.039 |
| ostracodsAbundance | (Intercept) | 50,452.448 | 0.168 | 0.034 |
| ostracodsAbundance | OmegaCA_mean | −192.416 | 0.761 | 0.034 |
| ostracodsAbundance | pH_mean | −5,884.844 | 0.222 | 0.034 |
| bryozoan_larvaeAbundance | (Intercept) | 258,053.587 | 0.001 | 0.029 |
| bryozoan_larvaeAbundance | CO3_mean | −183.926 | 0.398 | 0.029 |
| bryozoan_larvaeAbundance | OmegaCA_mean | 10,686.272 | 0.255 | 0.029 |
| bryozoan_larvaeAbundance | pH_mean | −33,162.070 | 0.001 | 0.029 |
| pteropoda_heteropodaAbundance | (Intercept) | 46,407.845 | 0.502 | 0.010 |
| pteropoda_heteropodaAbundance | CO3_mean | −19.975 | 0.920 | 0.010 |
| pteropoda_heteropodaAbundance | OmegaCA_mean | 384.316 | 0.964 | 0.010 |
| pteropoda_heteropodaAbundance | pH_mean | −5,234.678 | 0.563 | 0.010 |

Table 5: OLS model results of the abundance of calcifying species using predictors selected by LASSO regression.

| Other Species Model Summary Table | | | | |
|---|---|---|---|---|
| Taxa | Term | Estimate | P-Value | Adjusted $R^2$ |
| copepoda_oithona_likeAbundance | (Intercept) | 2,059,456.531 | 0.000 | 0.218 |
| copepoda_oithona_likeAbundance | CO3_mean | −4,556.554 | 0.002 | 0.218 |
| copepoda_oithona_likeAbundance | OmegaCA_mean | 186,375.489 | 0.003 | 0.218 |
| copepoda_oithona_likeAbundance | pH_mean | −252,030.957 | 0.000 | 0.218 |
| naupliiAbundance | (Intercept) | 606,276.046 | 0.088 | 0.124 |
| naupliiAbundance | CO3_mean | −4,962.985 | 0.000 | 0.124 |
| naupliiAbundance | OmegaCA_mean | 196,929.347 | 0.000 | 0.124 |
| naupliiAbundance | pH_mean | −70,394.335 | 0.131 | 0.124 |
| cnidaria_ctenophoresAbundance | (Intercept) | 128,808.835 | 0.000 | 0.115 |
| cnidaria_ctenophoresAbundance | pH_mean | −15,808.571 | 0.000 | 0.115 |

Table 6: OLS model results of the abundance of other species using predictors selected by LASSO regression.

Response variables (the abundance of different taxa) are shown in the taxa column in Tables 5-6. Predictors are shown in the Term column, and the corresponding coefficient estimate, p-value, and adjusted-$R^2$ values of the models are shown in the next 3 columns, respectively. P-values less than 0.05 indicates that the predictor is statistically significant. Adjusted-$R^2$ values reflect the proportion of variance of the data explained by the model. The best model explains only around 50% of the variance in the data so we cannot draw any conclusion from the models at this point.