Nathan Greenslit Data Analysis Project

Nathan Greenslit

4/4/23

# 1. Questions for Dr. Handel:

1. Does my overall workflow in code/analysis\_code/stats/3\_uni\_variate.qmd make sense? Is there anything that I am missing, or be a better option for analysis?
   * Script Flow for Dust: Plot Dust and Copies –> Create Model –> Assess Performance (AIC/RMSE/R2) –> Plot Residuals.
   * If this looks good, I plan to do the same workflow for all of the environmental parameters as well. For now, I just have plots to see what may be worth diving into.
2. Looking at code/analysis\_code/stats/4\_multi\_variate.qmd , I am a little stuck on how to best create a multivariate model.
   * Example: We can see this Salinity at Blind Oso drop at the same time that dust is coming in. This will most likely have an impact of Vibrio growth, but Salinity is probably non-linear.
   * Taking a look at the section titled Blind Oso: Growth ~Salinity + Dust, does my overall workflow look correct?
3. Looking at my statistics/models, is there anything that you think is strikingly missing?

# 2. Summary/Abstract

Each year, plumes of Saharan dust travel across the Atlantic to be deposited in the surface waters of the Caribbean and Gulf of Mexico. Dust aerosols serve as a significant source of nutrients that can elicit a fertilization effect on marine coastal waters, leading to rapid and potentially harmful blooms (e.g. red tides). Previous work has characterized the response of the marine bacteria *Vibrio* to dust input under the oligotrophic settings of the Florida Keys, but less is known regarding settings with higher baseline nutrient levels. This project aims to quantify *Vibrio* population dynamics in response to dust input in coastal sites with higher ambient nutrient levels. Daily sampling took place in Corpus Christi, TX to capture before, during, and after a Saharan dust event and quantitative PCR (qPCR) was used to estimate total counts of *Vibrio.* This study will provide an increased understanding of the conditions that can elicit potentially harmful blooms, highlighting the need for further research to understand the effects of dust deposition in non-oligotrophic waters.

# 3. Introduction

## 3.1 General Background Information

The Sahara Desert is a significant source of atmospheric dust, eliciting an estimated one billion metric tons each year ([d’Almeida, 1986](https://journals.ametsoc.org/view/journals/apme/25/7/1520-0450_1986_025_0903_amfsdt_2_0_co_2.xml)). Annually, plumes of this dust travel across the Atlantic via easterly trade winds to be deposited in the surface waters of the Atlantic, Caribbean, and Gulf of Mexico (Fig. 1). These dust events are highly episodic, occurring 3- 4 times a year typically in the summer months and lasting 3-5 days each. Dust aerosols can harbor a wide diversity of bacteria ([Kellogg et al., 2004](https://link.springer.com/content/pdf/10.1023/B:AERO.0000032947.88335.bb.pdf)), fungi ([Ramírez Camejo et al., 2022](https://link.springer.com/article/10.1007/s10453-022-09752-9)), virus-like particles ([Griffin et al., 2001](https://link.springer.com/article/10.1023/A:1011868218901)), and minerals ([Formenti et al., 2003](https://agupubs.onlinelibrary.wiley.com/doi/full/10.1029/2002JD002648)). These aerosols also serve as a significant source of macro and micronutrients such as phosphate, nitrate, and iron ([Graham & Duce, 1967](https://www.sciencedirect.com/science/article/pii/0004698182901986), [Mills et al., 2004](https://pubmed.ncbi.nlm.nih.gov/15152251/), [Savoie et al., 1989](https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/JD094iD04p05069?casa_token=PMRaet9QSYsAAAAA:nb8yBorxdKaEWwGfI5aNILJ0D4h6drNu2YY1y1y8ExbQ03z59EUTQTOHoxinzsxRtR4jzsfSA73_Hig)). This addition of otherwise limiting resources can lead to rapid and potentially harmful blooms of certain microbial groups ([Lenes et al., 2008](https://www.sciencedirect.com/science/article/pii/S0278434308000575), [Westrich et al., 2016](https://www.pnas.org/doi/abs/10.1073/pnas.1518080113), [Westrich et al., 2018](https://www.frontiersin.org/articles/10.3389/fmars.2018.00012/full)).

Previous work conducted in the West Florida Shelf has shown that increased iron availability can stimulate the growth of the nitrogen-fixing cyanobacterium *Trichodesmium*. The high amount of nitrogen produced via fixation can be adequate enough to stimulate toxic red tides, caused by the dinoflagellate *Karenia brevis* ([Lenes et al., 2008](https://www.sciencedirect.com/science/article/pii/S0278434308000575)).

While it is documented that these dust events can stimulate growth of harmful phytoplankton, emerging evidence shows that heterotrophic bacteria are responding to the episodic influx of limiting resources and substrate, resulting in blooms of bacteria that are associated with disease, which we refer to as harmful bacterial blooms. Among those that respond are bacterium belonging to the genus *Vibrio*, a group of ubiquitous marine opportunistic heterotrophs. While considered conditionally rare (comprising of <1% of the microbial community), under the proper conditions, this genus can rapidly bloom to make up a large percentage of the community over a short period of time ([Westrich et al., 2016](https://www.pnas.org/doi/abs/10.1073/pnas.1518080113), [Westrich et al., 2018](https://www.frontiersin.org/articles/10.3389/fmars.2018.00012/full)). *Vibrio* species are found in brackish to marine waters with salinities ranging from 0-35 and will grow in waters with temperatures ranging from 5°C-40°C ([Sampaio et al., 2022](https://www.mdpi.com/1424-2818/14/2/97)). Each year, plumes of Saharan dust travel across the Atlantic via easterly trade winds. These plumes are then deposited in the surface waters of the mid-Atlantic, Caribbean, and Gulf of Mexico. These dust aerosols harbor a wide range of fungi, bacteria, virus-like-particles, minerals, and nutrients (NO3, Fe, PO4). Addition of these nutrients to otherwise oligotrophic settings can result in large and rapid blooms of potentially harmful microbes, presenting a danger to both marine and human health.

*Vibrio* species are excellent opportunists, having multiple advantages that allow them to quickly respond to any newly introduced substrate or nutrients. These advantages include multiple copies of rRNA genes, a rapid doubling time, efficient chemotactic motility, large genomic repertoire (consisting of two circular chromosomes), and multiple Fe-siderophore complexes that allow for rapid uptake of iron into the cell ([Jensen et al., 2009](https://academic.oup.com/femsle/article/294/2/207/512983), [Ringgaard et al., 2018](https://journals.asm.org/doi/10.1128/JB.00793-17), [Eagon, 1962](https://journals.asm.org/doi/abs/10.1128/jb.83.4.736-737.1962?casa_token=WdtRNnQWyhkAAAAA:Zre66xvwgEu3wXQykenhnSoy3OA08h1u-4Zl31k9cxxxfV6jq46dzL18Y02OmwPHdY7p7tgH0wVN-4I), [Okada et al., 2005](https://journals.asm.org/doi/10.1128/JB.187.2.752-757.2005), [Payne et al., 2015](https://journals.asm.org/doi/full/10.1128/MMBR.00046-15)).

Within the *Vibrionaceae* family, there are at least 12 species that are commonly known as human pathogens. Illness can be induced through either a foodborne route via the consumption of raw or uncooked seafood and contaminated drinking water, or through nonfoodborne routes such as wound exposure while swimming. The most common illnesses caused by *Vibrio* include self-limiting diarrhea and cholera (*V. cholerae*), shellfish-induced gastroenteritis (*V. parahaemolyticus, V. vulnificus*), extreme cases of necrotizing fasciitis and septicemia (*V. vulnificus*), and wound and ear infections  (*V. alginolyticus*, *V. vulnificus*) ([Baker-Austin et al., 2018](https://www.nature.com/articles/s41572-018-0005-8)). Some species also pose a threat to marine health, inducing mortality in oyster and clam larvae (*V. tubiashii, V. coralliilyticus*), bacterial bleaching and rapid tissue loss in corals (*V. mediterranei, V. coralliilyticus*) ([Richards et al., 2015](https://pubmed.ncbi.nlm.nih.gov/25344234/), [Rosenberg & Falkovitz, 2004](https://www.annualreviews.org/doi/10.1146/annurev.micro.58.030603.123610), [Ben-Haim et al., 2003](https://pubmed.ncbi.nlm.nih.gov/12656189/)), and mortality in important aquaculture species such as penaeid shrimp (*V. parahaemolyticus*) and sea breams (*V. harveyi)* ([Tran et al., 2013](https://www.int-res.com/abstracts/dao/v105/n1/p45-55/), [Haldar et al., 2010](https://www.sciencedirect.com/science/article/pii/S0944501309001177)), among others.

Previous work has characterized *Vibrio* blooms in response to Saharan dust input in the oligotrophic setting of the Caribbean and subtropical Atlantic (Barbados and Florida Keys, respectively), with surface water concentrations of total *Vibrio* increasing by five to thirty times that found during non-dust conditions and returning to baseline levels within 24-48 hours (Fig. 2). *Vibrio* composition within the larger microbial community also shifted following dust deposition, with initial levels of <1.4% to a peak of 19.8% of the bacterial community ([Westrich et al., 2016](https://www.pnas.org/doi/abs/10.1073/pnas.1518080113)). A similar phenomenon was also observed in the surface waters of the tropical and open ocean mid-Atlantic with *Vibrio* populations increasing 1.5-fold in the mid-Atlantic following deposition ([Westrich et al., 2018](https://www.frontiersin.org/articles/10.3389/fmars.2018.00012/full)). In a follow up study in the Florida Keys, episodic dust events during a daily time series promoted a succession of bacterial responses, with declines in *Prochlorococcus* coinciding with initial increases in bacteria belonging to the order *Vibrionales*, followed by subsequent shifts in response of different bacterial groups ([Borchardt et al., 2019](https://aslopubs.onlinelibrary.wiley.com/doi/full/10.1002/lno.11291)). To date, our understanding of *Vibrio* and other microbial response to dust input has focused on oligotrophic settings, where dust input is considered to be a critical source of limiting nutrients. However, the composition of desert dust can be complex and may have the potential to deliver critical resources that could be exploited by opportunistic microbes (like *Vibrio*) and elicit a growth response despite higher baseline nutrient levels. Given that human exposure is likely in these coastal waters, it is important to elucidate the relationship between dust input and microbial blooms in these higher nutrient environments.

This study aims to (1) quantify *Vibrio* population dynamics and composition, and (2) characterize potential microbial blooms and community shifts in response to dust input in coastal sites with higher ambient nutrient levels. The findings from this study will provide an increased understanding of the conditions that can elicit potentially harmful blooms (i.e., *Vibrio*), highlighting the need for further research to understand the effects of dust deposition in non-oligotrophic waters. We predict that (1) *Vibrio* will be among the first to respond to the influx of dust constituents and (2) that the fertilization effects of dust deposition (and possible successional changes) will be dampened at sites with higher baseline nutrient levels, as dust derived nutrients become less critical for growth.

With a heightened risk of *Vibrio* exposure in coastal waters, it is vital to understand microbial dynamics in nearshore environments. If dust is proven to elicit a microbial response, these events can be used to serve as an “early warning” of increased exposure risk to better protect public health.

## 3.2 Questions/Hypotheses to be addressed

How do Saharan dust events influence Vibrio populations in high nutrient coastal waters?

*What I am looking for* I expect to see a more dramatic growth response in the Gulf (low nutrient), whereas the higher nutrient sites may exhibit a dampened growth response since they already have high baseline nutrient levels. If we still see a growth response despite already having supportive background nutrient levels at these sites, this may suggest that there are other constituents in the dust that can elicit a growth response. Nutrients, dust input, salinity, and temperature will be the primary factors of interest.

*How I will analyze it* - Look at Vibrio growth over time series (estimated from qPCR) - Examine relationships between dust input and Vibrio growth - Examine influence of site-specific environmental parameters (Temperature, Salinity, Nutrients) on growth response - Run linear models on factors like: dust x growth and nutrients x growth - NMDS to see which parameters have the strongest influence on growth

# 4. Methods

## 4.1 Description of Study Area

Sampling took place at three locations in Corpus Christi, TX: Blind Oso Bay, a residential canal system on Padre Island, and the Gulf. Sites were chosen based on differing baseline nutrient levels. Blind Oso Bay is a shallow tributary that is popular amongst wadefisherman and kayakers. The Bay is often impacted by fluctuating salinities and high levels of nutrients and chlorophyl due to its proximity to a wastewater treatment plant ([Wetz, 2014](https://www.cbbep.org/manager/wp-content/uploads/1312-Oso-Bay-Nutrient-Assessment-Final_Report-6-24-14.pdf)). Additionally, the Bay has persistent issues with high levels of fecal indicator bacteria, placing it on the U.S. impaired waters list ([Nicolau & Hill, 2014](https://www.tceq.texas.gov/downloads/water-quality/tmdl/oso-bay-oyster-harvesting-assessment-103/103-oso-bay-monitoring-report-2013.pdf), [Texas Commission on Environmental Quality, 2022](https://www.tceq.texas.gov/downloads/water-quality/assessment/integrated-report-2022/2022-303d.pdf)). The residential canal system is also impacted by high nutrients and chlorophyl from storm water runoff, and salinity is primarily driven by precipitation. The Gulf site is characterized by constant salinities (~36) and low nutrient levels.

## 4.2 Sample Collection and Processing

#### 4.2.0.1 Collection

A high resolution (daily) time series was conducted starting on July 7th and ending on July 19th, 2022 capturing before, during, and after a Saharan dust event.

Vertical profiles of salinity, temperature, pH, and dissolved oxygen were obtained using a YSI ProPlus sonde. Samples were collected for chlorophyl *a* analysis, inorganic nutrients, dissolved and particulate organic matter, and microbial analysis.

#### 4.2.0.2 Vibrio Enumeration

Total *Vibrio* were estimated using genus-specific quantitative polymerase chain reaction (qPCR) ([Westrich et al., 2018](https://www.frontiersin.org/articles/10.3389/fmars.2018.00012/full)).

## 4.3 Data import and cleaning

This project works with three primary data sets. (1) Enumeration of Vibrio bacteria as copies/mL from quantitative PCR (copies\_master.csv), (2) Dust concentrations as aerosol optical thickness (AOT) derived from the Naval Research Lab (nrl\_conc.csv), and (3) Temperature and Salinity across the daily time series (ysi\_2.csv). Below is a brief summary of how each data set was cleaned. More details can be found in the supplementary files, containing the code and comments describing what each line does. Please refer to the project\_README.md file located in the NathanGreenslit-MADA-Project folder for details on the repository contents and instructions on reproducibility

#### 4.3.0.1 Cleaning of Total Vibrio Enumeration dataset

Quantitative PCR provides enumeration results as “cycle threshold or Cq” values. This depicts the cycle number in which enough of the DNA target was present to be amplified and thus detected. Lower Cq values correspond to higher target concentrations (as it takes less cycles to amplify), and vice verse for higher Cq values. The cleaning of this data set consisted of converting these Cq values into something that can be used to quantify Vibrio (copies of target per mL of seawater). To do so, I needed to take into account the amount of water sample that was concentrated on a filter, the total amount of DNA that was eluted following a DNA extraction, and the amount of DNA template added to each mix for qPCR (to name a few). Taking these into account, we are able to calculate our way from Cq values to copies of Vibrio per mL of seawater. qPCR was conducted in triplicate, so the last step was to take the average of the three replicates to have a final value per sample. A date column was also added for each sample.

#### 4.3.0.2 Cleaning of Dust Concentration (AOT) Data set

Dust concentration (AOT) was collected at time points 0 hour, 6 hour, 12 hour, and 18 hour. In this script, I made different data sets based on time points as well as a data set containing the summed dust concentration (of all time points) per day. For downstream statistical analysis, it will be helpful to have these different data sets to compare to time of sample collection.

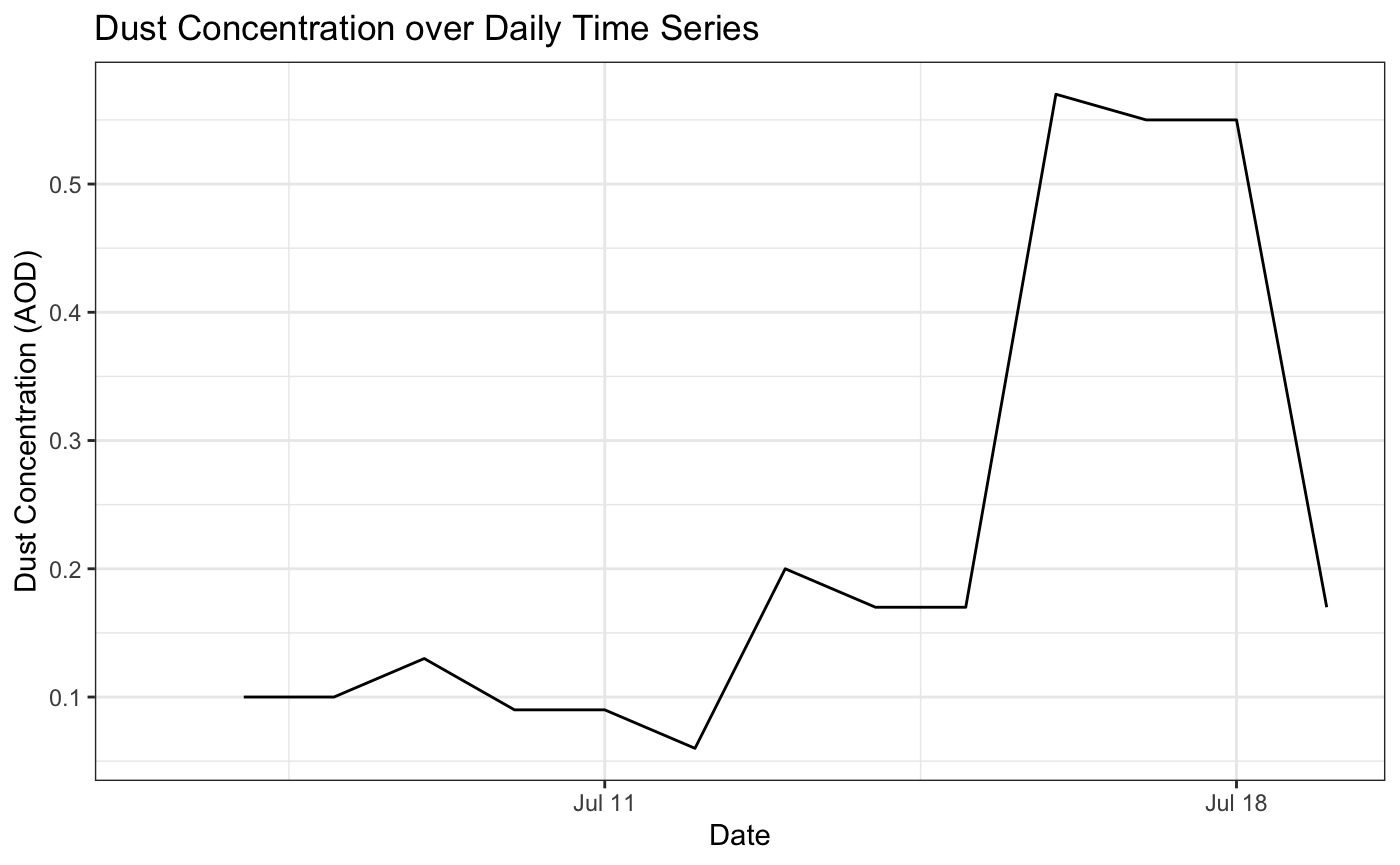
#### 4.3.0.3 Cleaning of Environment data set

This data set did not require any cleaning.

# 5. Results

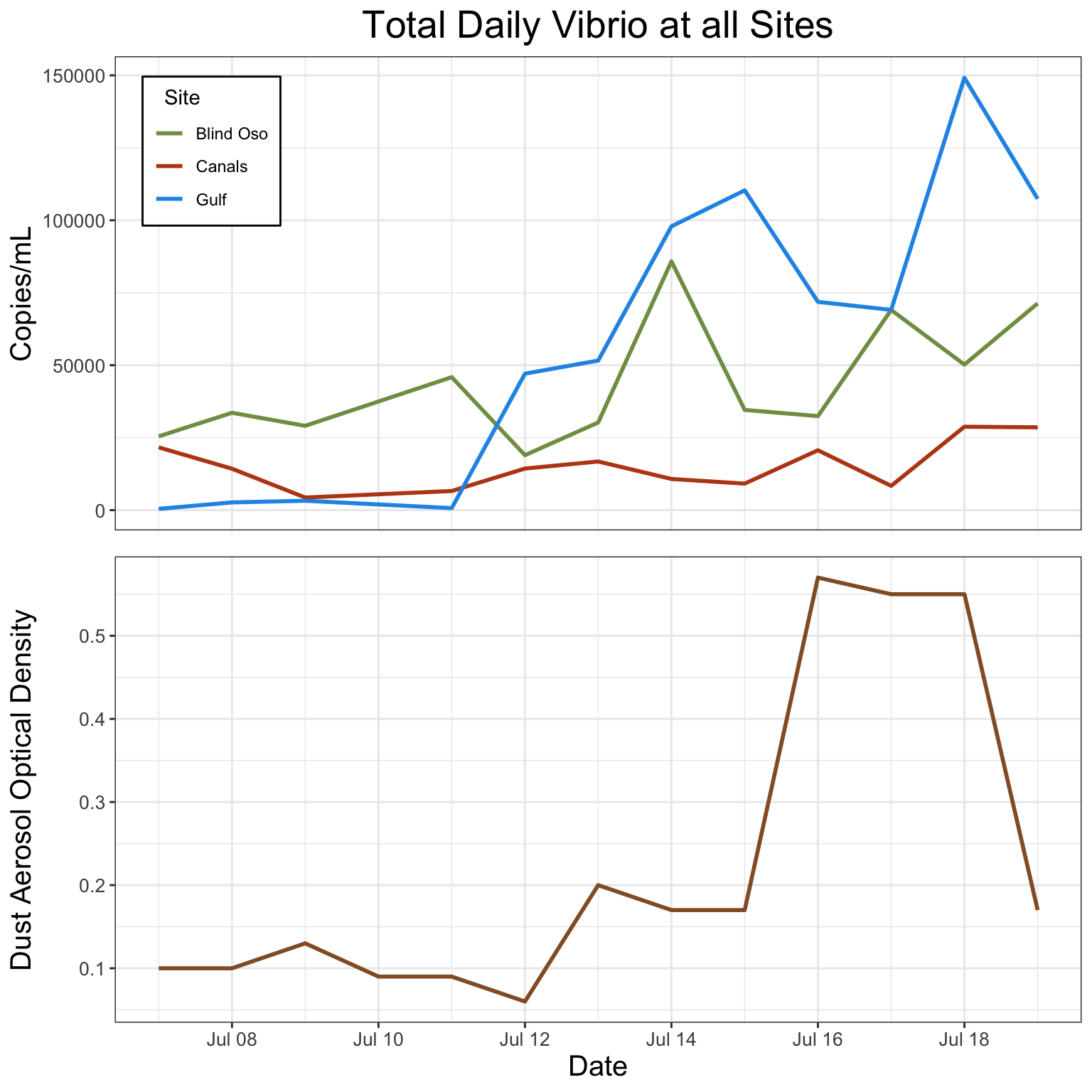
## 5.1 Exploratory/Descriptive analysis

#### 5.1.0.1 Dust



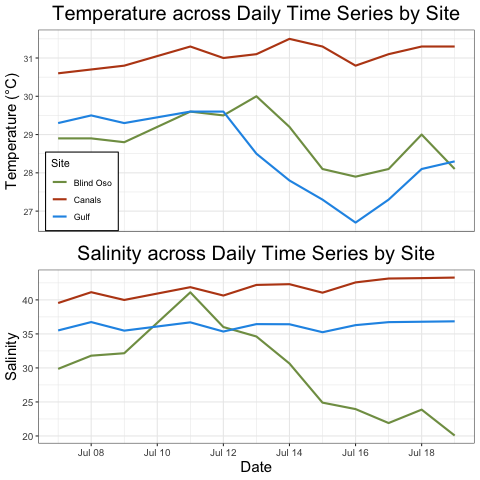
This figure depicts the summed dust concentration 24hr prior to sample collection (time points 13hr–> 7hr the next day) across the daily time series. During the time series, two small periods of higher dust AOT occurred on the 9th and 13th respectively. A much larger spike in AOT was observed on the 16th.

#### 5.1.0.2 Daily Vibrio



The top portion of the figure depicts total enumerated Vibrio as copies per mL, with color by site. There is a noticeable shift in copies in the Gulf site, a less noticeable but still present shift in Blind Oso Bay, and practically no changes in the Canal site. The bottom portion of the figure depicts the dust concentration (AOT) across the daily time series. Initial shift in copies per mL at the Gulf and Blind Oso occur 24-48hr following initial dust introduction on the 13th, and experience another shift around 24-48hr of higher dust input on the 16th.

#### 5.1.0.3 Temperature & Salinity



Temperature and salinity at the Canal site is relatively high and consistent as compared to the other sites. Blind Oso Bay experiences some shifts in temperatures and drastic fluctuations in salinity. The Gulf site had a drop in temperature (possibly due to upwelling) and relatively consistent ocean salinities

#### 5.1.0.4 Other Environmental Parameters

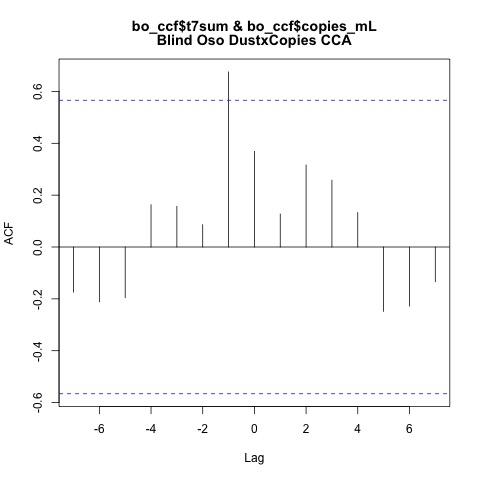
## 5.2 Basic statistical analysis

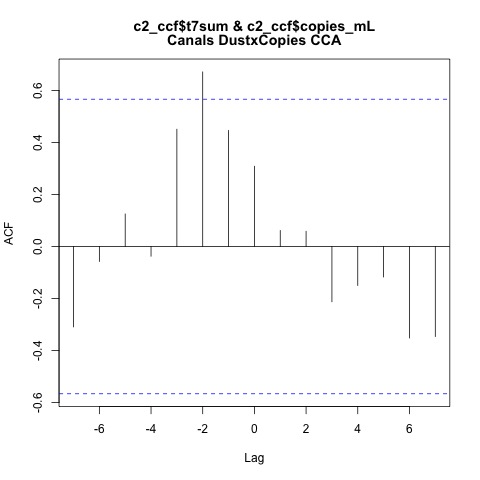
### 5.2.1 Check Distributions:

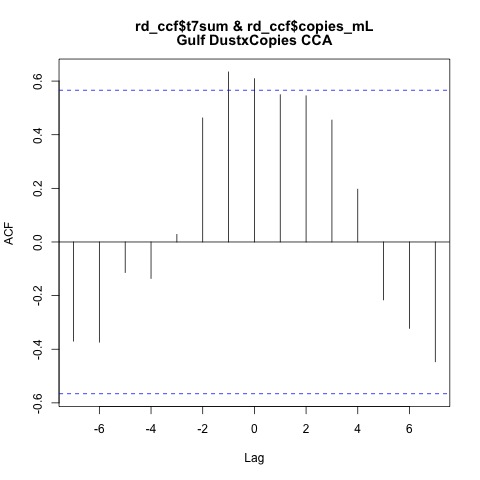
Shapiro-Wilk’s tests were conducted to assess distribution of the data. Any data that did not have a normal distribution was log-transformed and the test was run again.

### 5.2.2 Cross-correlation Analysis:

Previous literature has shown that Vibrio respond 12-24 hours following dust deposition events. Based on this, we need to look at lags between the two times series for dust data and growth data. To do this, we can set up a cross-correlation analysis, which allows us to examine the lag/lead relationship between the two variables.



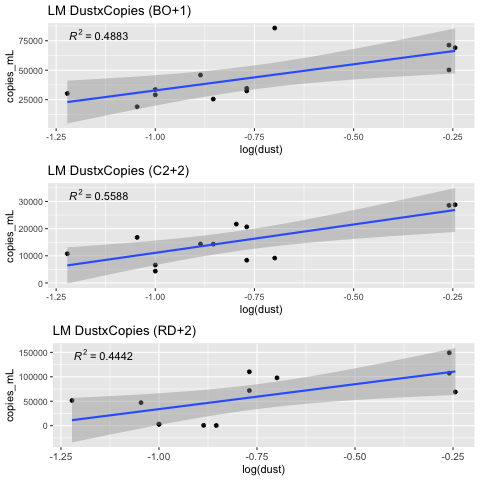




According to the ccf plots, BO has a significant lag relationship a value of -1. In other words, dust occurs, and one day later, we see a response in copies\_mL. C2 has a lag at -2, and RD at -1.

**One important thing to note**: Our data set contains dust data for all of 2022. Our copies\_mL data only contains values from 7/7/22 –> 7/19/22. Therefore, these CCA were made with only this time series. This limits our analysis, as we have dust data prior that can explain copies\_mL on the earlier dates 7/7 or 7/8. But we cannot incorporate this into the CCA as we have NA for copies\_mL for those early days. Below we will run linear models that DO contain these prior dust days. And since we have identified a lag, we can shift the dust data to line up with the copies\_mL that it corresponds with directly - and we no longer have NAs. But this ultiamtely means that the lag identified above. may not be the best fit for our model below, so some additional work is needed that tests other lags in the code.

### 5.2.3 Linear Regression Models



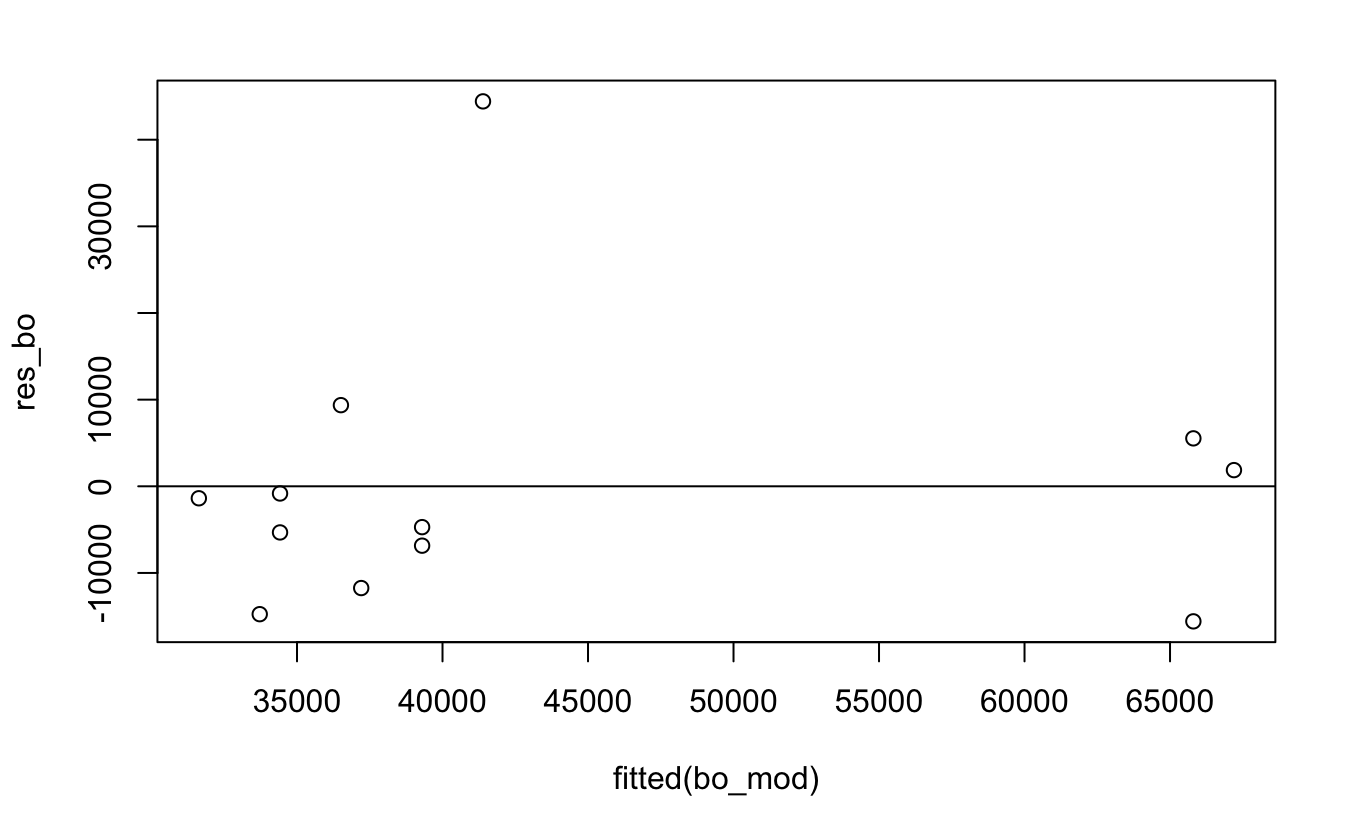
Above are linear regressions that examine the relationship between dust concentration and copies\_mL. As mentioned above, the lags identified may not correspond to the most significant relationship between the two variables. While a lag of -1 for Blind Oso and -2 for the Canals explains the data well, a lag of -2 for the Gulf site was more fitting. This makes sense, as taking a look at the copies\_mL plot above, we do see a ~2 day lag in response, where we have a spike in dust, and 2 days later, a spike in copies\_mL.

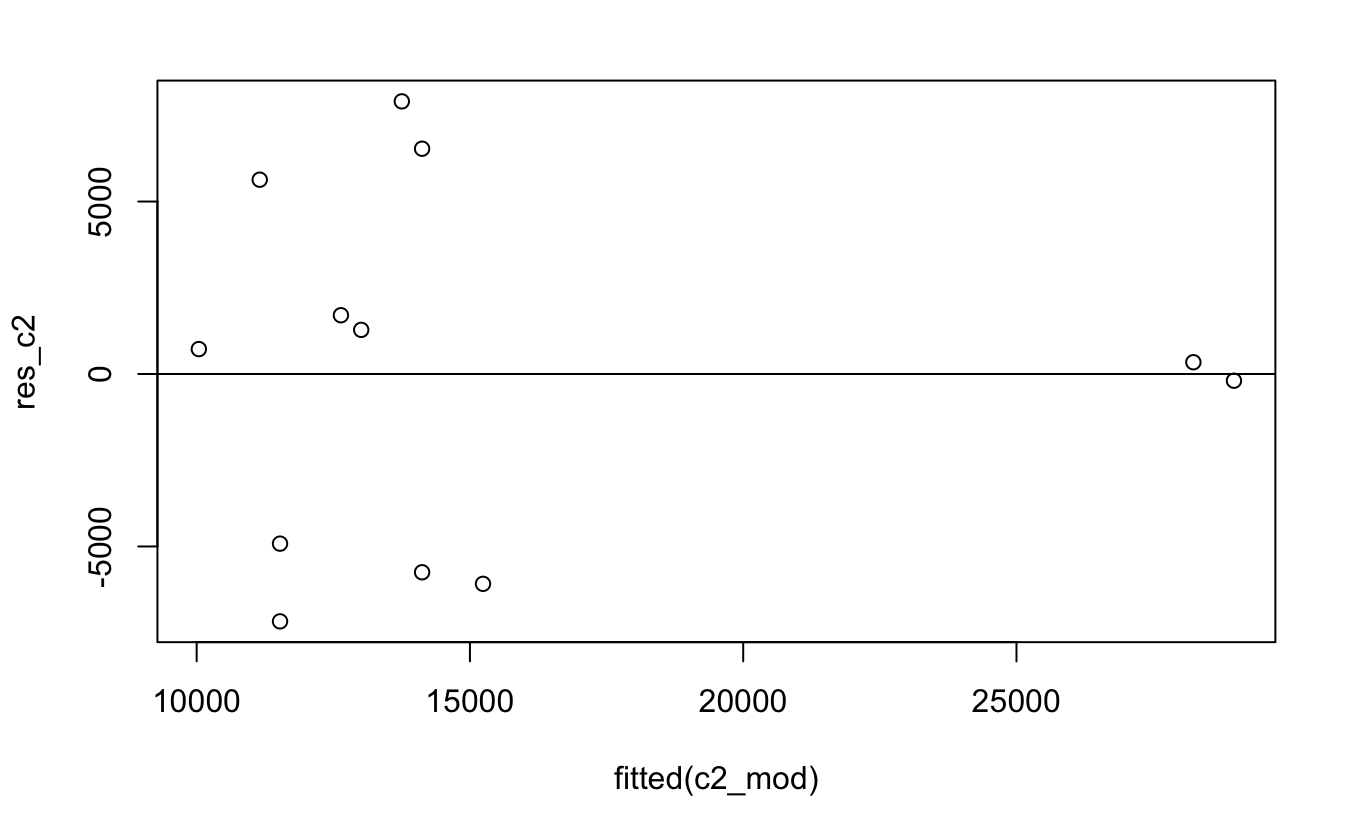
A few things to note:

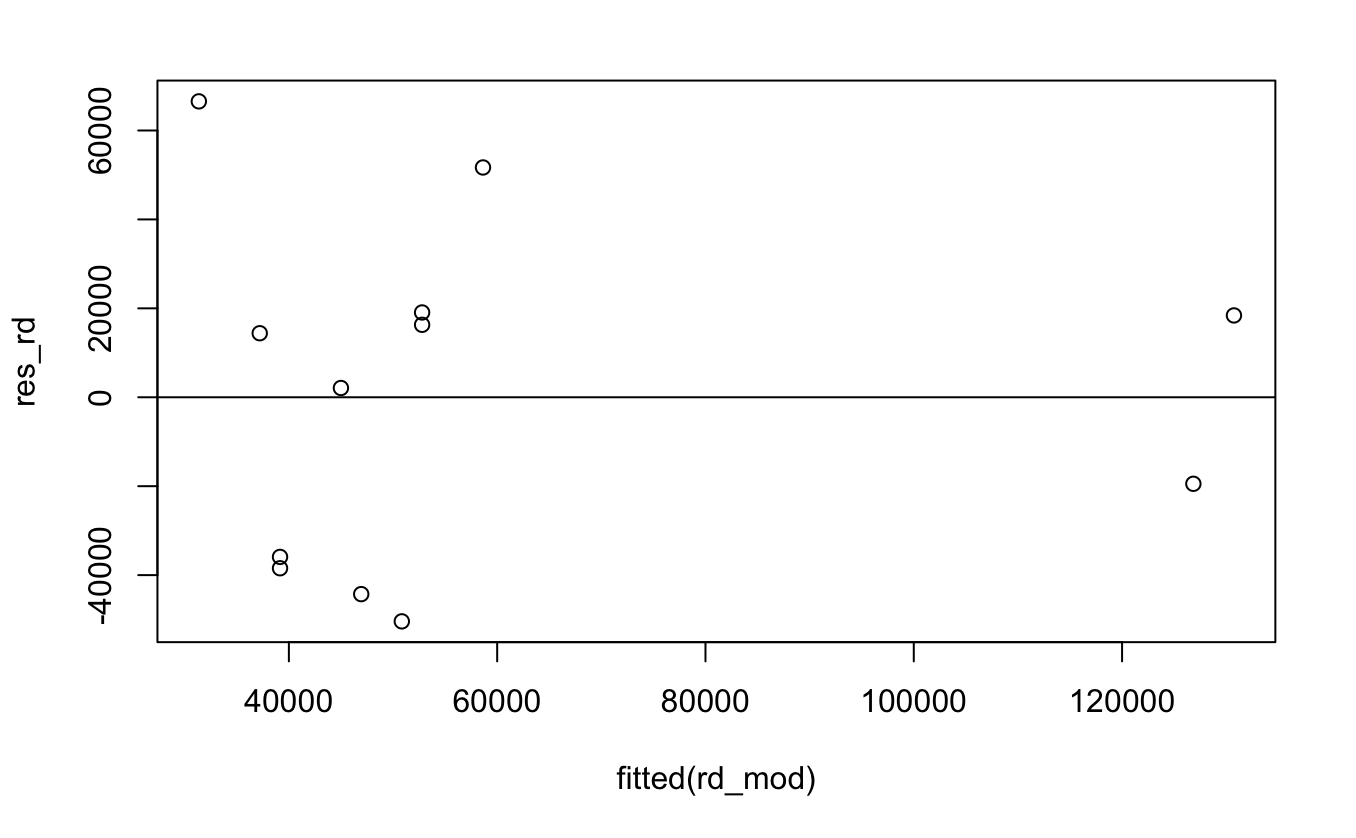
* My stats code contains some chunks that create predicted values based off of the current data set. Since my data set is small, I am not able to portion of ~25% to use to test my model as it could impact the strength of my model. Therefore, plotting the predicted vs. actual will give an R2 of 1.00. I am continuing to find a good validation method.
  + After reviewing module 9, it looks like Information Criteria may be a proper alternative (REF: MOD9/Model Performance and Overfitting).
    - Will need to interpret findings as exploratory and hypothesis generating, and will need to be careful about trying to draw generalizable conclusions.
    - AIC:
      * Blind Oso = 271.25
      * Canals = 243.97
      * Gulf = 292.13
* I am also planning on running a multi-variate model to look at multiple predictors of interests (temp, salinity, nutrients), but am currently running into an issue with my model output, where rather than getting values, I receive NaN. This code can be found in my dust\_copies\_stats.qmd code.

### 5.2.4 Residuals:

Instead of (or in addition to) plotting observed versus predicted for continuous outcomes, you can plot the difference between the two. These differences are called the residuals. What you are looking for is a cloud of points with no discernible pattern. If there is a pattern (e.g., an overall skew, or more points above the 0 y-axes than below), it again suggests that there is still some pattern/signal in the data that the model didn’t capture.







## 5.3 Full analysis

# 6. Discussion

## 6.1 Summary and Interpretation

## 6.2 Strengths and Limitations

## 6.3 Conclusions

# 7. References