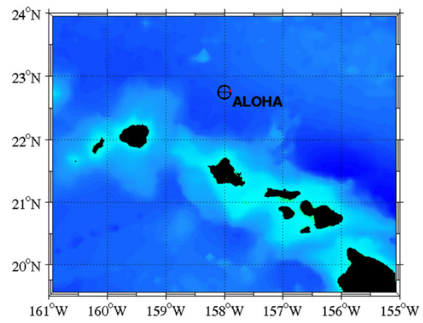
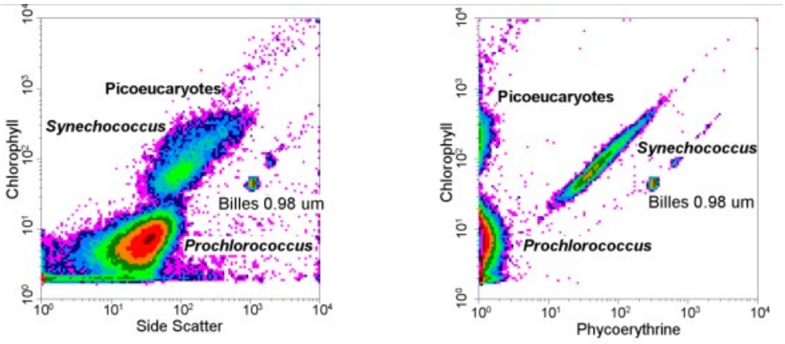
Homework 11 – Microbial niches in the Hawai‘i Ocean Time series





Beginning in 1988, the Hawai‘i Ocean Time series program has taken near-monthly samples of ocean biogeochemistry and physics at Station ALOHA, which is 100 km north of O‘ahu. Here we will use some of these data to learn about generalized additive models. The data are retrieved from the HOT-DOGS website. We will focus on flow cytometry counts of microorganisms from 2006-2022. The researchers distinguish four groups of microbes using flow cytometry: the cyanobacterium *Prochlorococcus* (pro), the cyanobacterium *Synechococcus* (syn), a diverse group of heterotrophic bacteria (hbact), and a diverse group of small eukaryotes collectively referred to as picoeukaryotic phytoplankton (picoeuk). We will focus only on pro, hbact, and picoeuk, and explore how the niches of these three groups differ from one another.

The attached dataset, HOT_cyto_counts_edit.csv, includes the following columns: "botid" (bottle ID), "date" (date, in mmddyy format), "press" (pressure in decibars) "chl" (fluorometric chlorophyll a concentration, in micrograms L^{-1}), "hbact" (heterotrophic bacteria concentration, in 10^{5} cells mL^{-1}), "pro" (Prochlorococcus concentration, in 10^{5} cells mL^{-1}), "syn" (Synechococcus concentration, in 10^{5} cells mL^{-1}), "picoeuk" (photosynthetic picoeukaryote concentration, in 10^{5} cells mL^{-1}), and "cruise" (cruise ID). Note that pressure in decibars is approximately equal to depth in meters (i.e., depth of the sampling device is measured using a pressure sensor).

1. For each of the three groups of microbes, fit a 2D smoother that characterizes how abundance changes with depth and with day of the year (i.e., from day 1 to day 365 or 366). To create a day of the year predictor you will need to first convert the ‘date’ column to date format, and then make a new column that extracts the day of the year from the date column. There are helpful functions in the package ‘lubridate’ that will do these steps for you.

   When fitting the 2D smoother for each type of microbe, consider how the response variable should be modeled (transformed or not, normal or non-normal). You can see the probability distributions available for the gam() function in package mgcv by looking at the help file titled ‘family.mgcv’. Consider whether the basis dimension needs to be increased beyond the default value. Plot the fitted smoother in a way that is visually appealing. Finally, figure out how to test whether the relationship between abundance and depth changes over time or not. What are your interpretations of the results so far?

2. Now let’s compare the niches of the three groups to each other. Use a GAM including all groups simultaneously to simultaneously test three questions:

(a) Do the different kinds of microbes have different mean abundances?

(b) Do the different kinds of microbes have different *average* depth distributions (i.e., averaging over time)?

(c) Do the different kinds of microbes have different *average* seasonal dynamics (i.e., averaging over depths)?

To fit GAM(s) including all three groups simultaneously you will need to convert the data to 'long' format, where there is a column that contains all the concentrations of all three types of microbes, and a second column that codes which microbe was counted in that row, as well as additional columns for the other model predictors. You can convert to long format by hand using a spreadsheet, or you can use a helpful function called pivot_longer() in the package 'tidyr'.

Now that you have fit models to test questions (a)-(c), make appropriate plots and perform appropriate hypothesis tests. How do you intepret the results?

3. Finally, let's investigate how abundances of the three groups at shallower depths correlate with mixed layer depth (an index of stratification) and chlorophyll a concentration. The second attached file, hot_mlds.csv, contains the average mixed layer depth for each HOT cruise. You'll need to merge the information in this file with the dataset you have been analyzing.

   Using only data from the top 45 meters, how does the concentration of each group of microbes vary with (a) mixed layer depth and (b) Chl *a* concentration? Test whether the three types of microbes exhibit different relationships with the two predictors. Use appropriate GAM(s), hypothesis tests, and smoother plots to assess these questions.