Homework 10 – Predicting *Vibrio vulnificus* in the Ala Wai Canal

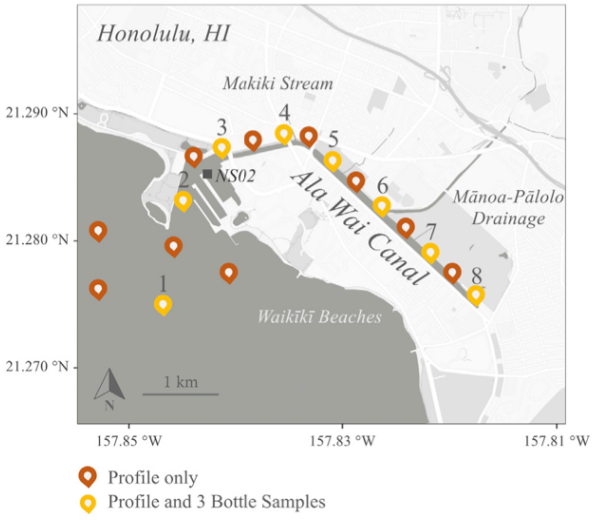

Fig. 1. Map of the Ala Wai Canal in Honolulu, Hawai'i depicting monthly survey sampling locations from October 2018–September 2019. Full depth profiles for salinity, temperature, dissolved oxygen, turbidity, and chlorophyll were collected from all sites (n = 18) marked with a pin. Numbered sites (1–8) are locations where discrete bottle samples were collected at 3 depths (surface, pycnocline, and bottom water) for additional nutrient, organic matter, and microbial measurements. The PacIOOS nearshore sensor NS02 at the Ala Wai Harbor is marked with a grey box.

Vibrio vulnificus is a bacterial pathogen that occurs in warm estuarine environments. Non-choleric *Vibrio* causes 20-40 infections per year in Hawai'i, and *V. vulnificus* is known to occur in the Ala Wai Canal and Harbor, but the environmental drivers of *V. vulnificus* abundance are not well understood. Bullington et al. (2022) investigated spatial and temporal variability of *V. vulnificus* in the Ala Wai Canal, with the goal of providing better predictive capacity for *V. vulnificus* risk. We will use their dataset as a test case for learning about model selection with complex survey datasets.

The attached data file includes 213 samples from 9 sampling dates (Oct 2018-Sept 2019), 8 sites along the canal and harbor (see figure above), and 3 depths (surface, pycnocline, and bottom). There are a large number of columns, reflecting the large number of variables measured during these sampling events. We will focus on the following columns: vvhA (*Vibrio vulnificus* concentration estimated by qPCR – gene copies per mL), Season (dry vs. rainy), Site, SampleDepth, Rainfall_5Day (5 day average rainfall), AirTemp (air temperature), O2Conc (oxygen concentration), Salinity, WaterTemp (water temperature), Turbidity (from optical backscatter) , Chlorophyll (Chl *a* in vivo fluorescence), NOx (nitrate +nitrite), Silicate, POC (particulate organic carbon), TotalP (total phosphorus), Tyrosine.like (tyrosine like dissolved organic matter), HIX (dissolved organic matter humification index), and VisibleHumic.like (visible humic-like DOM).

Although this is a fairly exploratory study, there is some context for thinking about potential drivers and how they are represented by the measured variables. *V. vulnificus* is a heterotrophic bacterium, which means it feeds on dissolved organic matter. Temperatures above 18ºC and salinities between 15 and 25 are thought to favor *V. vulnificus* based on prior work. Rain events cause salinity and water temperature to decline, while delivering dissolved nutrients, organic matter, and particulates from runoff. The sampling dates in this study included both the dry season and the rainy season, although during this year the rainy season did not have greater average precipitation.

1. Start by analyzing general spatial and temporal patterns of *V. vulnificus* abundance. Plot how abundance varies between sites, seasons, and depths. Construct an appropriate model that tests whether there are effects of these predictors, and interactions between them. Figure out how the response variable (gene copies per mL) should be modeled (i.e., whether to transform or not, whether to use a non-normal distribution). What are the magnitudes of the modeled effects? What are your interpretations of the results so far?

2. The authors measured many potential predictors of *V. vulnificus*, but only ended up using this set when performing model selection: Rainfall_5Day, AirTemp, O2Conc, Salinity, WaterTemp, Turbidity, Chlorophyll, NOx, Silicate, POC, TotalP, Tyrosine.like, HIX, and VisibleHumic.like. The reason is that there is substantial intercorrelation among the full set of predictors, and this set of variables are either relatively easy to measure or hypothesized to be particularly important. Recreating the authors' analysis of predictor correlation is outside the scope of this assignment, but you may want to explore it if you are interested.

   Use scatterplots of *V. vulnificus* concentration vs. the predictors to consider which should be transformed for use in a linear model / generalized linear model. Recall that transformation of predictors is not about normality (predictors are not assumed to follow any distribution). Rather, transformation is useful for achieving linear relationships (when using linear models), and for avoiding extreme values that have a large leverage on the modeled relationship.

   Now let's compare different approaches to inference about the importance of these predictors.

   Start by making single-predictor models for each of the 14 predictors listed above. Which predictors, on their own, best explain *V. vulnificus* concentration? If you make an AIC table of the 14 models, what do the Akaike weights look like? What does this mean? Note that in order to compare models by AIC each model needs to use the same set of samples (rows) – this means you should remove any rows that include NAs for any of the 14 predictors, or for the response variable, before fitting any of the models. What is the downside of using only single-predictor models in this context?

   Now make one big model that contains all 14 predictors. Do marginal null hypothesis tests on the predictors. Which seem important for explaining *V. vulnificus*? Collectively, how much variation in *V. vulnificus* concentration can be explained? What is a potential downside of this approach to inference (i.e., one big model with marginal tests)?

   Now construct all possible models containing these 14 predictors (you can ignore interactions…and recall that there is an R function that automates this process). Which predictors consistently occur in the most-supported models? What are the sums of Akaike weights for each predictor? Plot the fitted relationships from the best model. How do these results compare to the other approaches?

   Finally, attempt an interpretation of the results. What mechanisms could underlie the strong correlations identified by the analysis?