Testing Logistic Regression Models (Supplement to: Environmental DNA Metabarcoding Reveals Winners and Losers of Global Change in Coastal Waters)

Gallego et al.

We aimed to develop a quantitative model relating the probability of a taxon's presence at a site as a function of the environmental conditions at that site. Creating and tuning separate models for each of more than 200 taxa would have been analytically and practially unwieldy, and moreover would have missed the benefits of partially pooling information across taxa. Consequently, we developed a hierarchial logistic regression model in which the effects (slopes) of environmental parameters were allowed to vary by taxon.

Given the available continuous explanatory variables (Temperature, pH, and Salinity) and the discrete location data (geographic Area: Hood Canal vs. San Juan Island), a threshold decision was whether Area should feature in the model. PERMANOVA analysis (see main text) indicated strongly different biological communities across the two Areas, and as a result, plausible models could include an intercept term that varied by Area and taxon. Conversely, it was implausible that the relationship between individual taxon presence and a given environmental variable (say, Temperature) would vary between Areas.

For all model fitting, we used *rstanarm*, precompiled functions from the Stan language for Bayesian modeling, which implement Hamiltonian Monte Carlo parameter searches.

Preliminary Model Testing and Model Selection

We conducted a preliminary model test on a suite of models meeting the above criteria using a random subset of taxa drawn from the overall dataset, further restricting this subset to those taxa occurring in $> 10\%$ and $< 90\%$ of samples, ensuring a degree of variability in taxon presence on which to train the model. Furthermore, we standardized environmental variables to have mean = 0 and sd = 1 in order to provide sensible priors and to improve computational efficiency.

```
##                          elpd_diff se_diff
## mod1.2vars                  0.0       0.0
## mod1.3vars                 -3.8       1.8
## mod1.1vars.pH              -7.2       4.1
## mod1.1vars.Temperature     -7.7       4.0
## mod2.2vars                -64.7      11.3
## mod2.3vars                -68.5      11.5
## mod2.1vars.Temperature    -70.3      11.8
## mod2.1vars.pH             -71.1      11.9
## mod3.2vars               -329.1      20.5
## mod3.1vars.Temperature   -329.8      20.6
## mod3.1vars.pH            -330.2      20.6
## mod3.3vars               -330.5      20.6
```

The best-fit model included effects of Temperature and pH that varied by taxon, and an intercept term that varied for each unique taxon-Area combination. Formally, each observation is modeled as a single (Bernoulli) draw from a binomial distribution, with probability p. This probability p depends, in turn, on a linear combination of the explanatory variables, using a logit link.

$$Presence \sim Bernoulli(p_i)$$

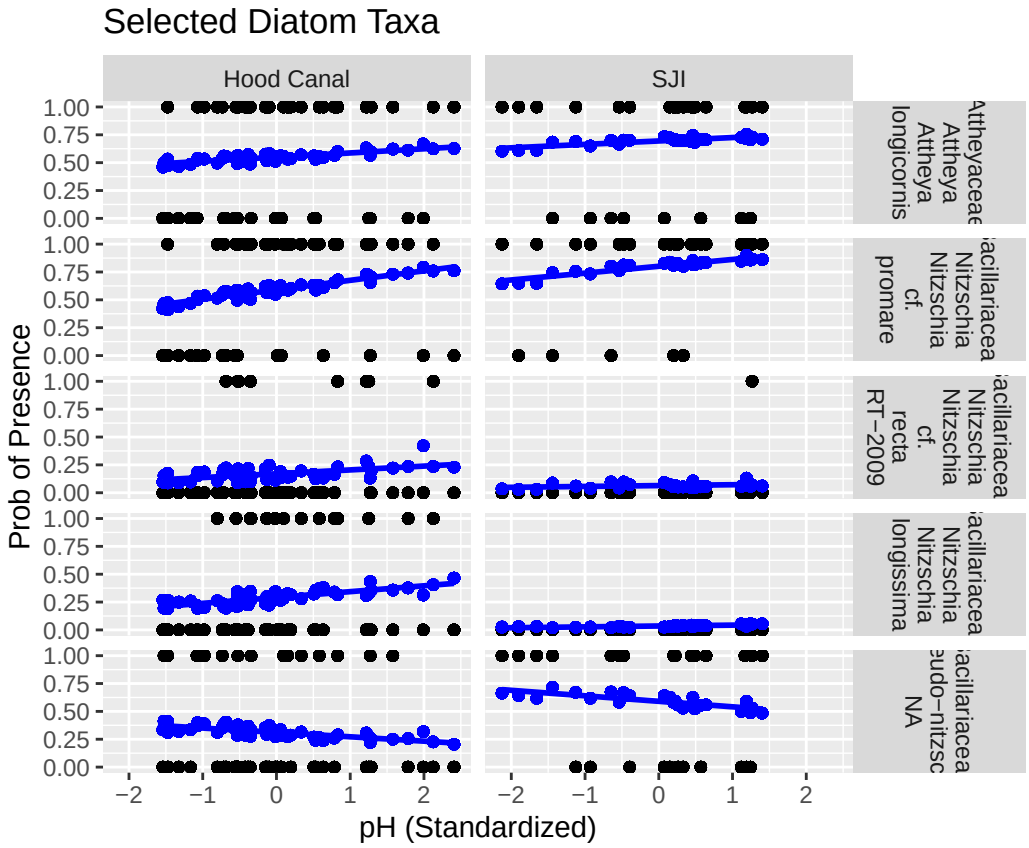
$$logit(p_i) = \alpha_j + Temperature_i * \beta_{1_k} + pH_i * \beta_{2_k}$$

Where i indexes the individual observations in the dataset, j indexes unique combinations of taxon and geographic Area, and k indexes taxon identity. The parameters α, β_1, and β_2 represent the intercept, effect of Temperature, and effect of pH, respectively. Priors and hyperpriors were drawn from a regularizing Student's t distribution (df = 7, location = 0, scale = 2.5).

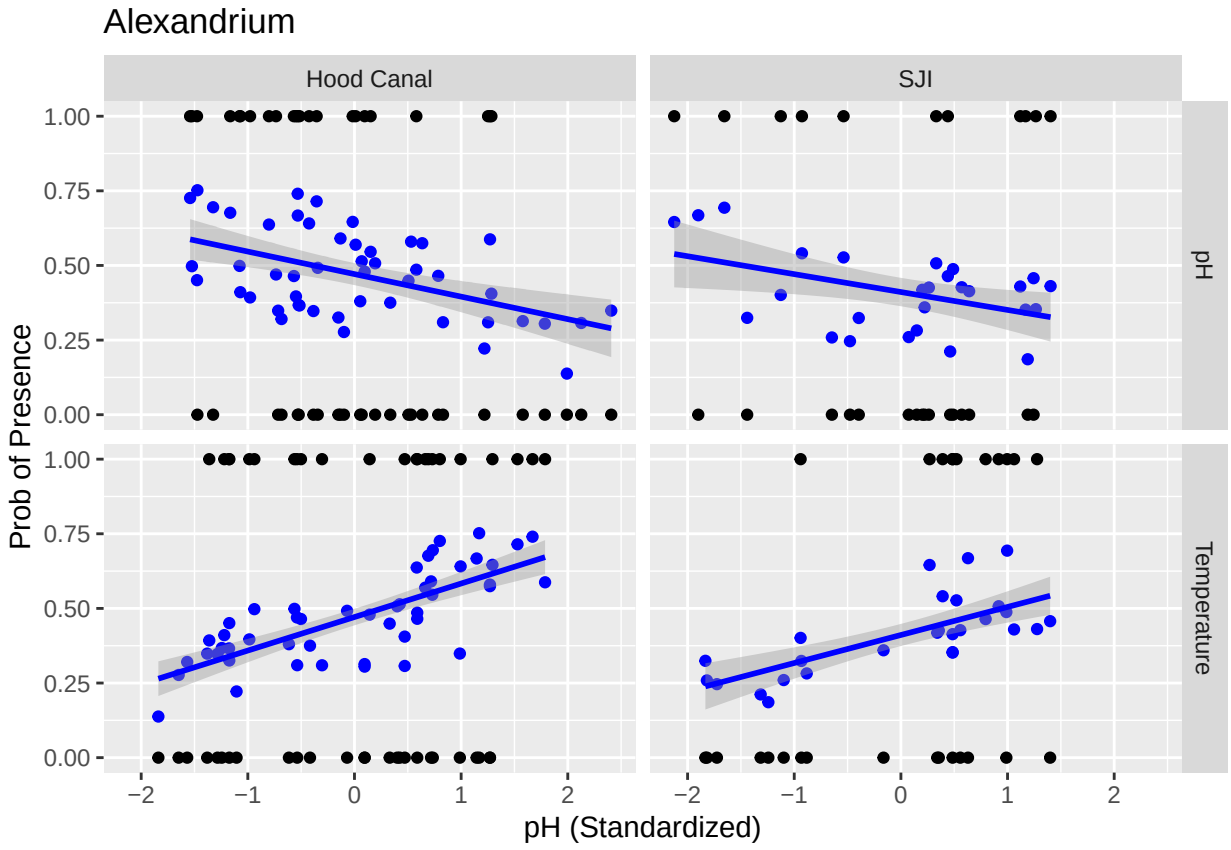
Final Model Fit and Illustration

Finally, we fit this model to the overall dataset – again, keeping only taxa with at least 10% variability in presence across samples.

The plot below illustrates the model fit, with observed data in black and modeled probabilities in blue.




```
## `geom_smooth()` using formula 'y ~ x'
```


Alexandrium
Hood Canal
SJI
pH
Temperature
Prob of Presence
1.00
0.75
0.50
0.25
0.00
−2
−1
0
1
2
pH (Standardized)