Comparison of model transferability by method

Table of contents

1 Background

The first objective of my NSF project is to compare among four different modeling methods trained on Gulf of Mexico green turtle data and determine which produces the most accurate predictions of habitat selection on independent data sets from Brazil and Qatar. Each model was fit using a resource selection function, typically expressed as $w(x) = exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ...)$. The modeling methods that were compared include a generalized linear model (GLM), generalized additive model (GAM), boosted regression tree (BRT), and Gaussian Process regression (GPR). To account for individual functional

responses to each of the selected model covariates (i.e., bathymetric depth, net primary productivity, sea surface temperature), each model was fit as a hierarchical model using a random intercept and random slopes per each of the 49 individuals tracked in the GoM, with the exception of the BRT model since there is not currently a way to implement such effects. These models are therefore abbreviated as HGLM, HGAM, and HGPR. To allow the HGLM model to more flexibly fit nonlinear relationships, quadratic terms were included for each covariate.

Below, I've shown rasters for each of the three covariates (i.e., depth, NPP, SST) in October 2011 for the Gulf of Mexico. Additionally, I've shown predicted surfaces of log(intensity) of use per model for the GoM.

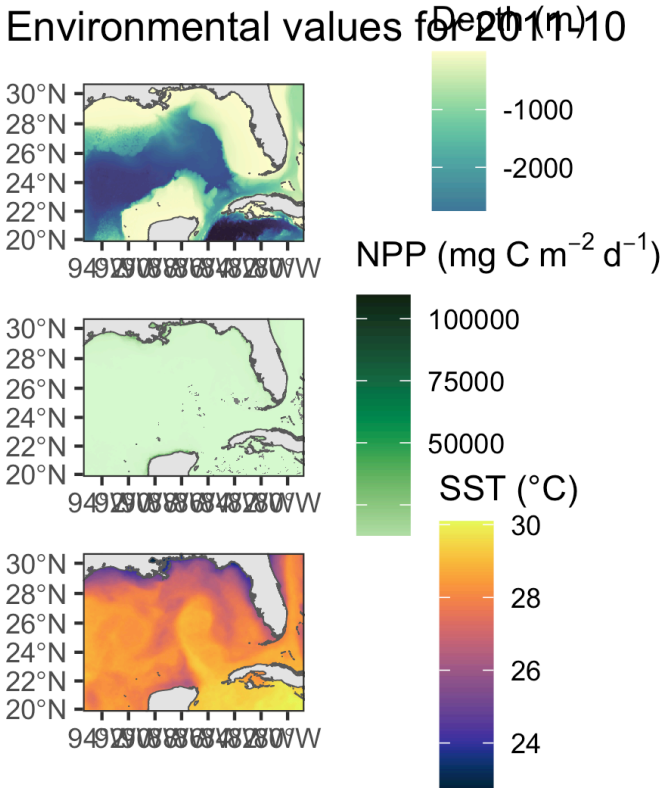

Figure 1: Example raster layers from the Gulf of Mexico for each of the selected covariates included in the model. Since NPP and SST are dynamic variables, these data were accessed on a monthly basis.

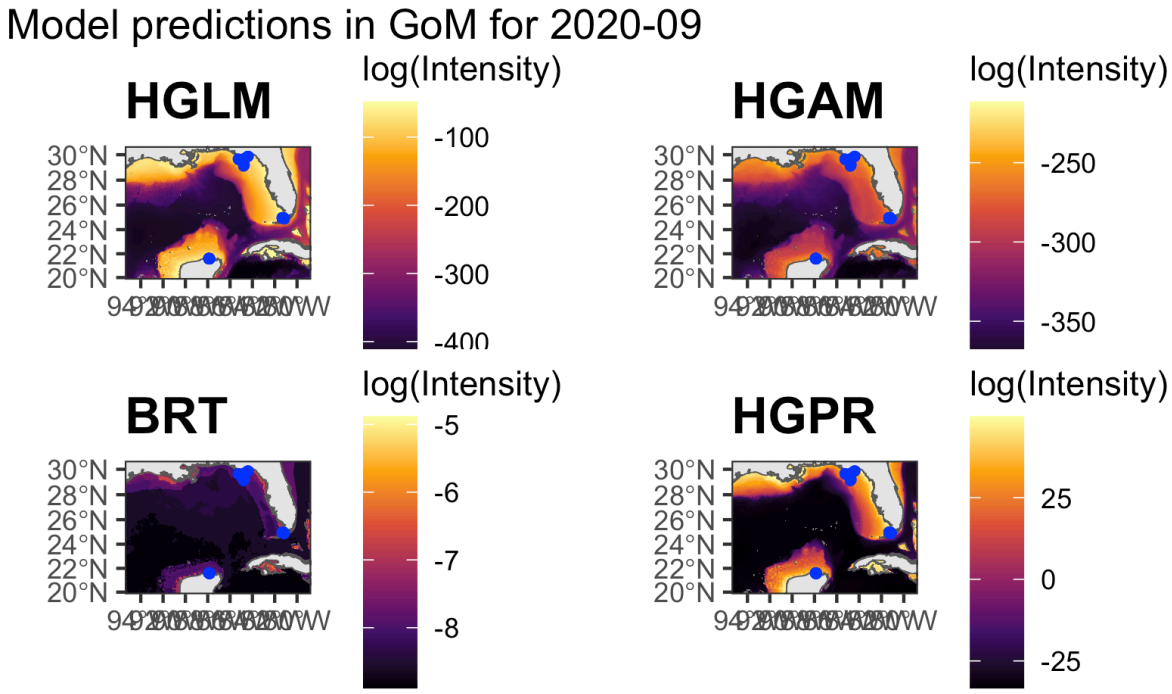

Figure 2: Predictions from each of the four models are mapped onto space for 2020-09 at the population-level. Note that each of the legends have different scales, but that the color scale (i.e., relative intensity) should be treated as comparable across results. Blue points in each plot indicate the points of 'resident' locations from three individuals that were tracked during this month-year.

2 Methods

I will not be presenting the details of how each method was parameterized or fit in this document, but I will show example predictive surfaces across methods in addition to a comparison of the metrics of predictive accuracy (i.e., transferability). For this project, I evaluated model transferability using the Boyce Index calculated per month-year, which essentially divides the predicted values from each model into 10 bins of equal width and then calculates the Spearman correlation of the predicted/expected ratio of points found within each bin. This index ranges from -1 to +1, where values close to -1 indicate

a model predictive of where the animal isn't, values close to 0 are no different from random, and values close to +1 indicate a perfectly predictive model. Additionally, I calculated another metric that determines the number of bins required to account for $\geq 90\%$ of all observations. This second method is included to potentially distinguish between methods with high correlations from the Boyce Index, but few observations in areas of greatest predicted suitability.

3 Results

3.1 Example of predictive surfaces

As was shown prior for the predictive surfaces in the Gulf of Mexico (on the training data), here I will show how each of the models predict log(intensity) over space for a single month-year in Brazil and Qatar separately.

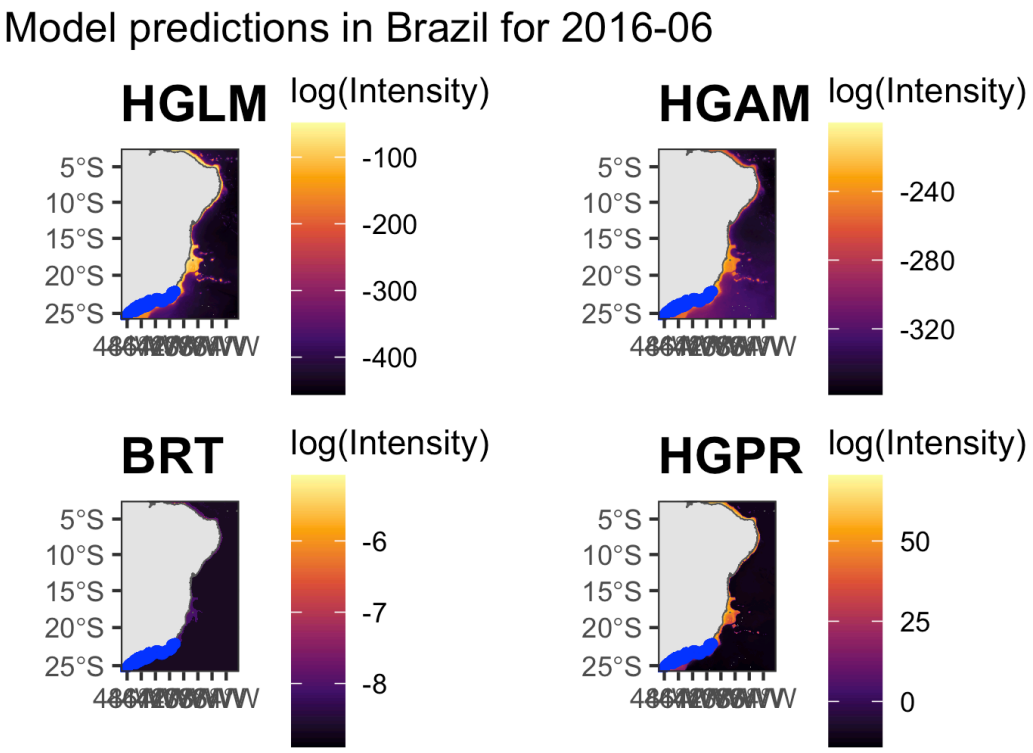

Figure 3: Predictions from each of the four models are mapped onto space at the population-level in Brazil and Qatar. Note that each of the legends have different scales, but that the color scale (i.e., relative intensity) should be treated as comparable across results. Blue points in each plot indicate the points of 'resident' locations from individuals that were tracked during the selected month-year.

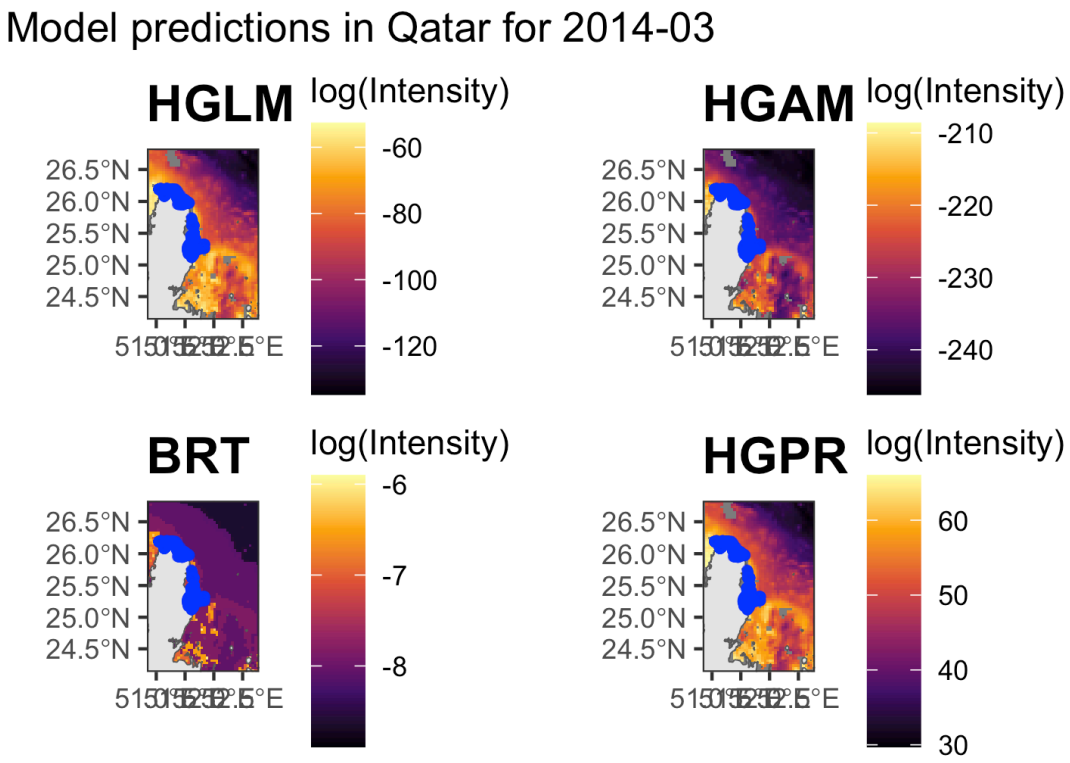

Figure 4: Predictions from each of the four models are mapped onto space at the population-level in Brazil and Qatar. Note that each of the legends have different scales, but that the color scale (i.e., relative intensity) should be treated as comparable across results. Blue points in each plot indicate the points of 'resident' locations from individuals that were tracked during the selected month-year.

3.2 Evaluating model transferability

The above predictions shown for the Brazil and Qatar data were performed for every month-year of the respective datasets, where time-matched observations were then used to extract these predicted estimates of log(intensity) per method. For each location, the range of all possible log(intensity) values were broken up into 10 bins of equal width per method (i.e, HGLM, HGAM, HGPR, BRT) and a ratio of predicted:expected observations was calculated per bin. A Spearman correlation was then calculated for the predicted:expected ratio over the binned values of log(intensity) to calculate the Boyce Index, which ranged from -1 to +1.

Inspection of some of these plots for the Boyce Index showed that correlations could be relatively high despite the observed points falling in the middle of the predicted log(intensity) bin range, rather than being the highest at the greatest predicted values. Therefore, this method was supplemented by calculating the average number of bins that accounted for $\geq 90\%$ of the observations per site (Brazil, Qatar) per method, starting from the upper end of the distribution since this is where the greatest number of observations should be found for a highly predictive model.

Upon exploratory inspection of some of the predicted intensity surfaces from the models, it appeared that the spatial resolution of 4.5 km could not always properly represent the environmental conditions at the small island of Fernando de Noronha for the Brazil dataset. Since 20 individuals spent some or all of their time at this island, each of the model transferability assessments were calculated using all Brazil data ('Brazil_full') or only individuals tracked along the mainland ('Brazil_sub') to account for potentially biased estimates of transferability.

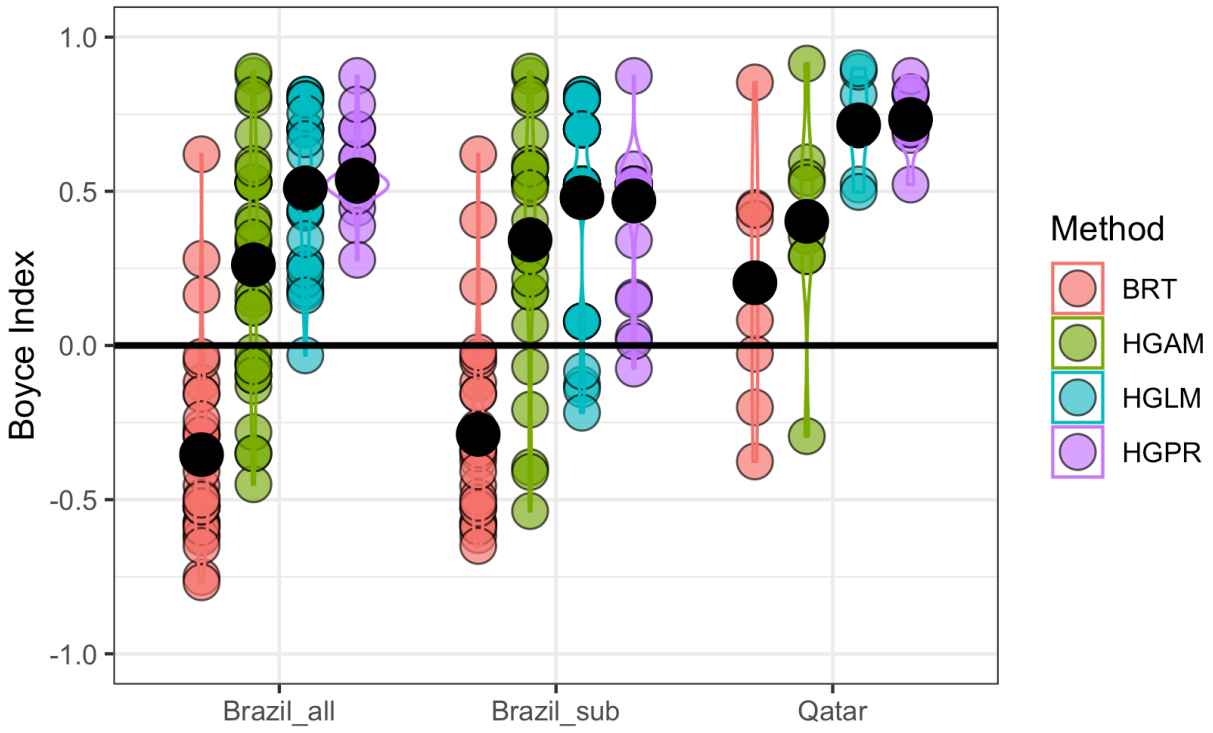

Figure 5: Violin plots of Boyce Index values per method and study region. The Brazil dataset has been split to evaluate model performance when including (Brazil_all) or excluding (Brazil_sub) observations at Fernando de Noronha. The black points indicate the mean values, whereas the colored points indicate the values per month-year of the dataset.

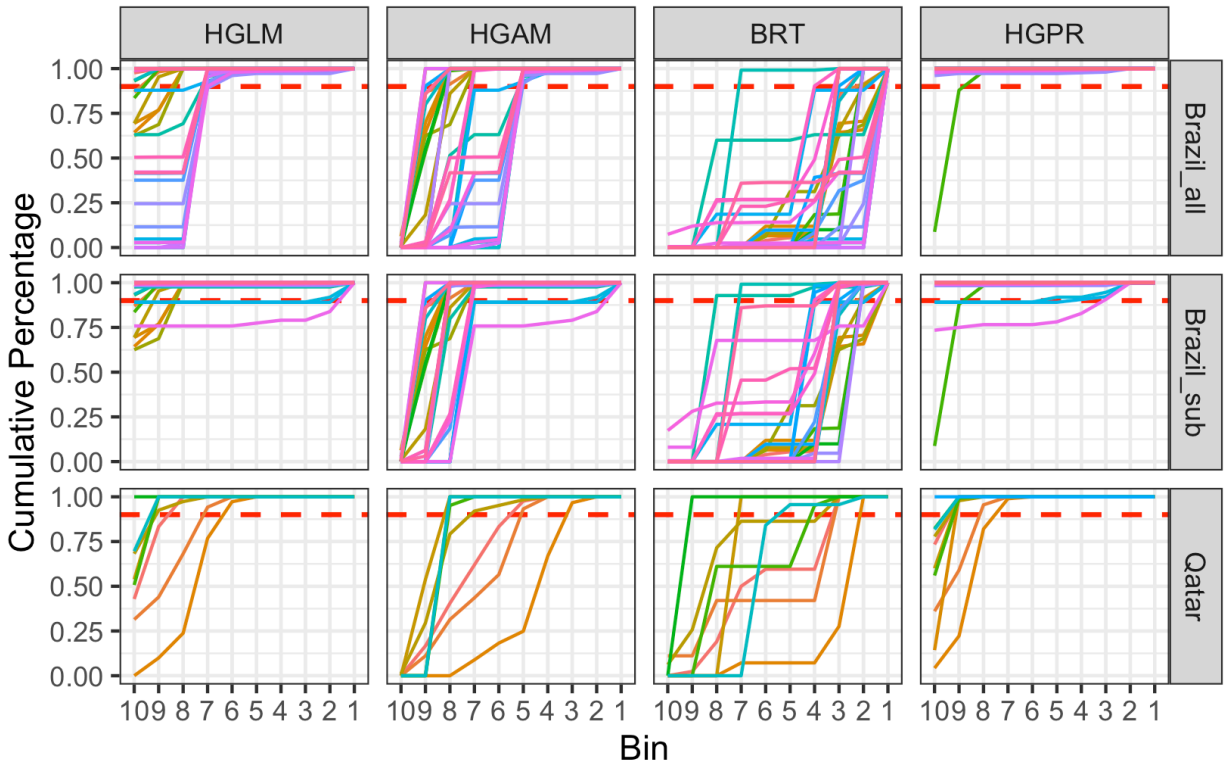

Figure 6: Line plots showing the cumulative percentage of observations found in each bin, working backwards from the greatest predicted intensity (bin 10) to the lowest (bin 1). Color of the lines denote each of the month-years of predictions per site. The horizontal red dashed line indicates 90%.

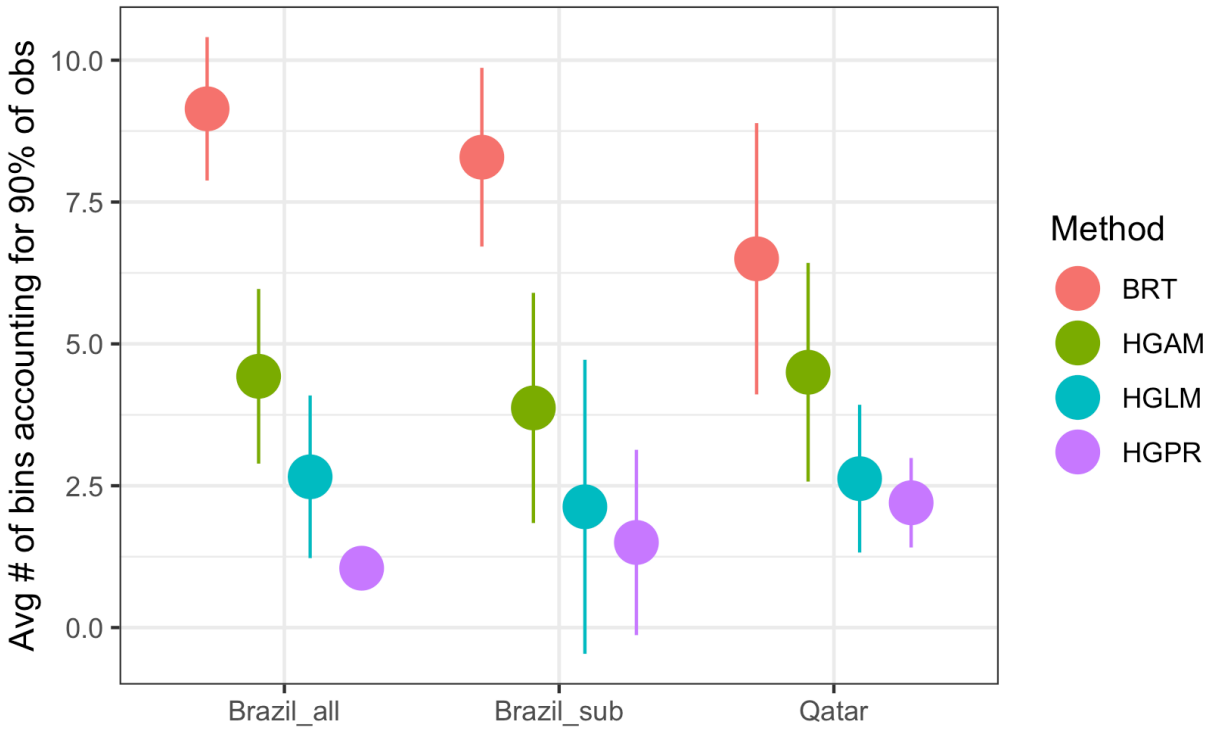

Figure 7: Plot showing the mean (+/- SD) number of bins that accounted for at least 90% of all observations. Fewer bins required to capture >90% of the data indicates a better performing model.

4 Conclusions

Based on these findings from comparing the model transferability based on the statistical model that was used, it appears that the hierarchical Gaussian Process regression performed best based on both the Boyce Index and the mean number of bins to capture $\geq$ 90% of the data. This result held up for both study locations (Brazil and Qatar), although differences were small between the HGPR and HGLM models.

For the remaining two objectives of my project, I'll be using the HGPR model to evaluate the effect of spatial resolution and accounting for life stage preferences on model tranferability.