Predicting Abundance of Key Algal Species in The Rocky Intertidal of Acadia National Park

Final Report

DSE6311 – SU2 2025

Nick Lagoni

8/27/2025

Executive Summary

The rocky intertidal, including the shores of Acadia National Park, are experiencing escalating stressors and change, including warming waters and introduction of invasive species. Some of these algal species, including rockweed (*Fucus vesiculosus*) and knotted wrack (*Ascophyllum nodosum*) are disproportionately important for the wellbeing of the ecosystem, as they provide habitat, shade, and food for the other organisms living in the rocky intertidal. The rocky intertidal ecosystem represents an important recreation and interpretation space for park-goers. The goal of this analysis is to understand what is driving change in these critical algal species and to build a model that can tell us how well or poor they will do in the future.

Using five years (2013 – 2018) of data collected by the Northeast Temperate Network, multiple predictive models were built to predict *F. vesiculosus* and *A. nodosum* abundance, and to determine which other algae and invertebrates contribute to that prediction most.

Of the models tested, eXtreme Gradient Boosting (XGBoost) provided the strongest predictions, predicting about 89% of the variation in *Ascophyllum* cover and about 78% for *Fucus*, with an error rate in predictions of about 10 – 15% (on a scale of 0 – 100%). The most important predictors were the previous year's abundance of each target species, followed by signs of competition with each other, and signs of competition with other algae such as *Chondrus* and *Mastocarpus*. Interestingly, invertebrate abundance, including snails which are known to graze on algae, only slightly influenced prediction.

For park management, the key takeaway is that abundance of these target species, at least within current environmental parameters, is relatively consistent, despite uncertainty for the ecosystem. Management should continue to monitor broad-strokes trends, either through the existing NETN protocol or an expanded protocol. Investigating environmental variables may shed additional light on the future of these algal ecosystem engineers, which may provide more avenues for management actions. The current model and monitoring protocol may eventually inform whether or not these target species are on the decline, but they may not be proactive enough or include enough information on why those changes are occurring to step in and address them.

1 Introduction

1.1 Background

Intertidal communities in the Gulf of Maine are experiencing rapid change, including changes in vertical zonation and introduction of invasive and range expanding species (Co(Trott 2022)hen et al. 1995, Yamada 2001, Epifanio 2013, Johnson 2015, Cheng et al. 2025). These changes may compound over time into significant changes to community structure. Perennial algal species including, *Fucus vesiculosus* and *Ascophyllum nodosum* (Linnaeus 1753), are subject to escalating stress and competition in the Gulf of Maine. As keystone species, *F. vesiculosus* and *A. nodosum* contribute disproportionately to the functioning of the ecosystem than their abundance would imply, as they create habitable space for motile and sessile invertebrates, epiphytes, and other macroalgae. (Dayton 1972, Råberg and Kautsky 2007, Parrot et al. 2019, Westerbom and Koivisto 2022). Because of this and the importance of these ecosystems ecologically and as a source of ecosystem services, including as interpretive spaces in National Parks, it is important that we understand what the future holds for *Fucus and Ascophyllum* in the rocky intertidal on the Gulf of Maine, and which factors contribute most to that future.

1.2 Research Question and Hypothesis

The primary research question posed by this analysis is: can we accurately predict how the community assemblage of the rocky intertidal in Acadia National Park will change in the near future, and what variables (including substrate composition, invertebrate abundance, and temperature) are the strongest predictors of that. While long-term monitoring has been ongoing through the Northeast Temperate Network (NETN) in the Inventory and Monitoring Division of the National Park Service (NPS) since 2013, no significant analysis has been done on this subject since that protocol began.

I hypothesize that the presence and abundance of motile invertebrates will significantly predict the substrate percent cover of *F. vesiculosus* and *A. nodosum* in the Acadia rocky intertidal over time.

I predict that an increase in the abundance of motile invertebrates *Littorina obtusata* will be associated with a significant reduction in the algal cover of *F. vesiculosus* and *A. nodosum*, as these periwinkles have been shown to feed on both of these species (Hadlock 1979, Watson and Norton 1987). I predict that *Littorina littorea* abundance will not significantly predict the algal cover of *F. vesiculosus* and *A. nodosum*, as this species has been shown to avoid feeding on these species (Watson and Norton 1985). I predict that increased temperatures will be correlated with a decrease in *F. vesiculosus and A. nodosum* in the higher plots, and an increase in both target species in the lower plots (i.e., the red algae plots); these species have been shown to expand their range as waters warm (Jueterbock et al. 2013, Marbà et al. 2017), which may extend to their local zonation range as well, expanding into deeper water.

2 Data Description

2.1 Data Source

NETN has a dataset spanning 8-years (2013 – 2021) of long-term ecological monitoring data across six sites in Acadia, as well as three sites in the Boston Harbor Islands National Park (Northeast Temperate Network 2021). At the time of this analysis, only a subset spanning five years from 2013 – 2018 was available. All nine sites are visited annually for data collection. The Acadia sites are the focus of this analysis, as the Boston Harbor Islands are a very different ecosystem, both naturally, with an uncommon mixed coarse substrate intertidal habitat, and artificially, in part due to Boston's position as a global shipping hub introducing a significant amount of invasive species (Putnam et al. 2024). The dataset is broken down into different sub-protocols, including motile invertebrate and substrate percent cover photoplots, which will be used for this analysis. See the data dictionary located in Appendix A for a detailed description of the dataset.

2.1.1 Motile Invertebrate and Substrate Percent Cover Photoplots

Each site is divided into subsites based on target species or group (*A. nodosum, F. vesiculosus,* red algae, barnacles, and mussels), with each of those comprising five photoplots each. This subprotocol includes count data of all motile invertebrate individuals collected from each photoplot, separated by species, as well as measurements of a random subsample of ten individuals for each species and photoplot, or however many were collected if there were less than ten in a plot. This data also includes percent cover data for each photoplot, determined using a proprietary model with visual validation by a human. After pivoting each dataset to wide form and joining them together, there are 40 parameters and 760 observations.

2.2 Variable Breakdown

The target variables in question are the percent cover of *F. vesiculosus* and *A. nodosum*, These are keystone species and ecosystem engineers, a focus of the NETN protocol and, from personal anecdotal experience working in the field on this protocol, have been undergoing dramatic changes in distribution along the intertidal in recent years. The predictor variables will include percent cover of each other measured algal species (or group when speciation is inappropriate or unfeasible) and sessile invertebrates like *Semibalanus balanoides* (Linné and Salvius 1758) and *Mytilus edulis* (Linné and Salvius 1758), as well as the abundance and mean measurement of each motile invertebrate species collected.

3 Data Exploration

3.1 Exploratory Data Analysis Methodology

A correlation matrix was used to give an overview of the trends in the data, and how different features may influence each other. To manage the high number of predictors and the overall weak multicollinearity, a threshold of r > 0.5 was used, resulting in a legible heatmap of moderately correlated predictors.

Scatterplots were used to examine the relationships between motile invertebrate abundances and each algal target variable (*Fucus vesiculosus* and *Ascophyllum nodosum*). These scatterplots included the target variable percent cover on the Y-axis, the specific motile invertebrate count on the X-axis, and a trendline, calculated via a simple linear model method.

Histograms were used to examine the frequency distribution for each target variable (Figure 4). These histograms illustrated that the target variables are heavily skewed and zero-inflated, suggesting that significant transformation would be necessary to fit the assumptions of many models and that non-parametric models may be a better fit.

Summary statistics (mean, median, standard deviation, minimum, maximum, skewness, and kurtosis) were calculated for all continuous variables and compiled into Table 1.

A Hellinger-transformed principal component analysis (PCA) was used to examine what variables were responsible for explaining most of the variance in the data and reveal a path towards reducing dimensionality. The Hellinger transformation is meant to improve the efficacy of PCA when analyzing ecological data. A scree plot (Figure 5) was used to visually examine the loadings of each principal component to get a sense of the structure of the data.

While the Hellinger transformation applied does work to address the zero inflated nature of the ecological survey data, studies have shown that even a Hellinger transformed PCA is a poor choice for reducing dimensionality in ecological data (Minchin and Rennie 2010). Non-parametric multidimensional scaling (NMDS) was used as a more appropriate dimensionality reduction technique for this data type (Clarke 1993, Minchin and Rennie 2010, Dexter et al. 2018). The results of the NMDS are visualized in Figure 6.

3.2 Correlation Heatmap

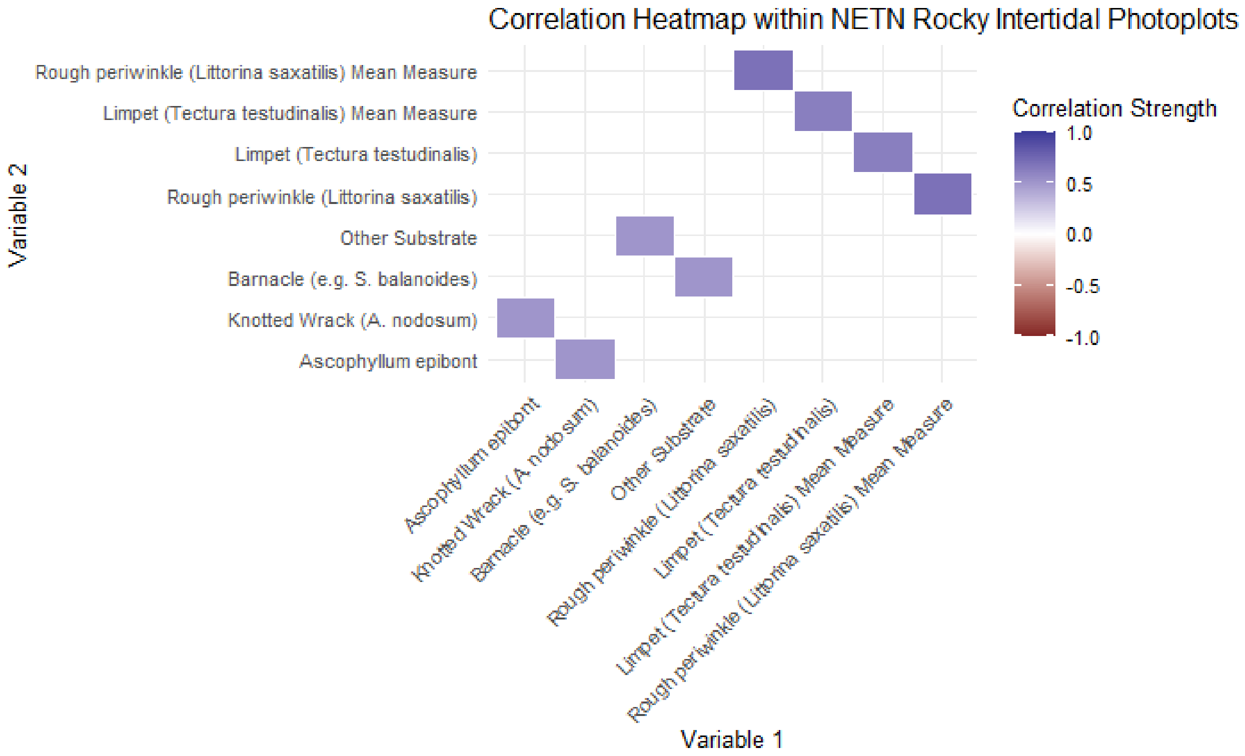


Figure 1 - Correlation Heatmap between moderately correlated variables in the NETN rocky intertidal long term ecological monitoring protocol photoplots. Variables include motile invertebrate count data, motile invertebrate mean length measurement data, macroalgae percent cover data, and substrate percent cover data.

The correlation heatmap showed that there is very little collinearity within this dataset. The strongest collinearity was seen between the motile invertebrate count data and their respective mean length measurement data. This conceptually makes sense as, in plots that are more hospitable to certain species (higher count abundance values) those same species would be more likely to thrive and grow larger. The other correlations were between *A. nodosum* and its epibiont, which requires *A. nodosum* as substrate to grow on, and barnacles and other substrate. Scanning the notes column in the raw data revealed that dead barnacles/leftover shells were frequently labelled as other substrate, which explains that correlation.

3.3 F. Vesiculosus *and* A. nodosum *relationship with Motile Invertebrates*

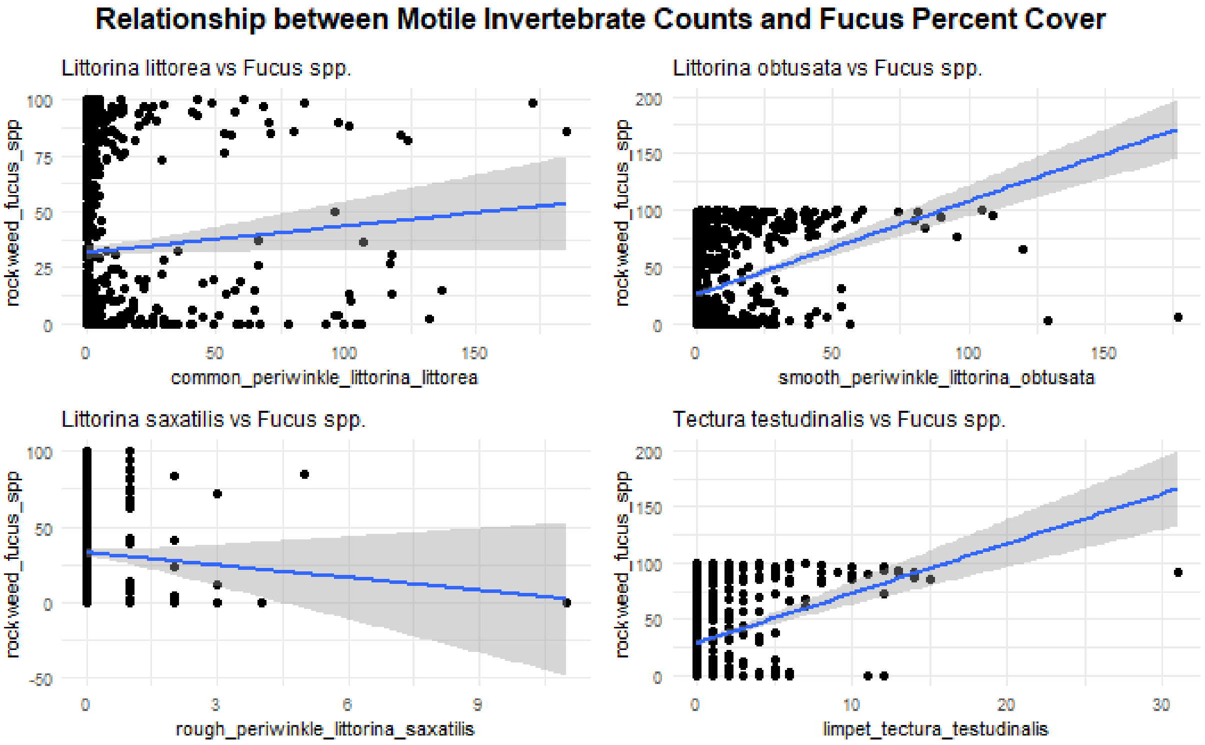


Figure 2 - Relationship between Fucus Vesiculosus percent cover and motile invertebrate abundances in NETN rocky intertidal photoplots.

Examining *F. vesiculosus* percent cover as a function of motile invertebrate abundances suggests that common periwinkles, smooth periwinkles, and limpets have a positive correlation, while rough periwinkles have a negative correlation. This may suggest that either the commons, smooths, and limpets promote *Fucus* in some way, or that they prefer that as habitat.

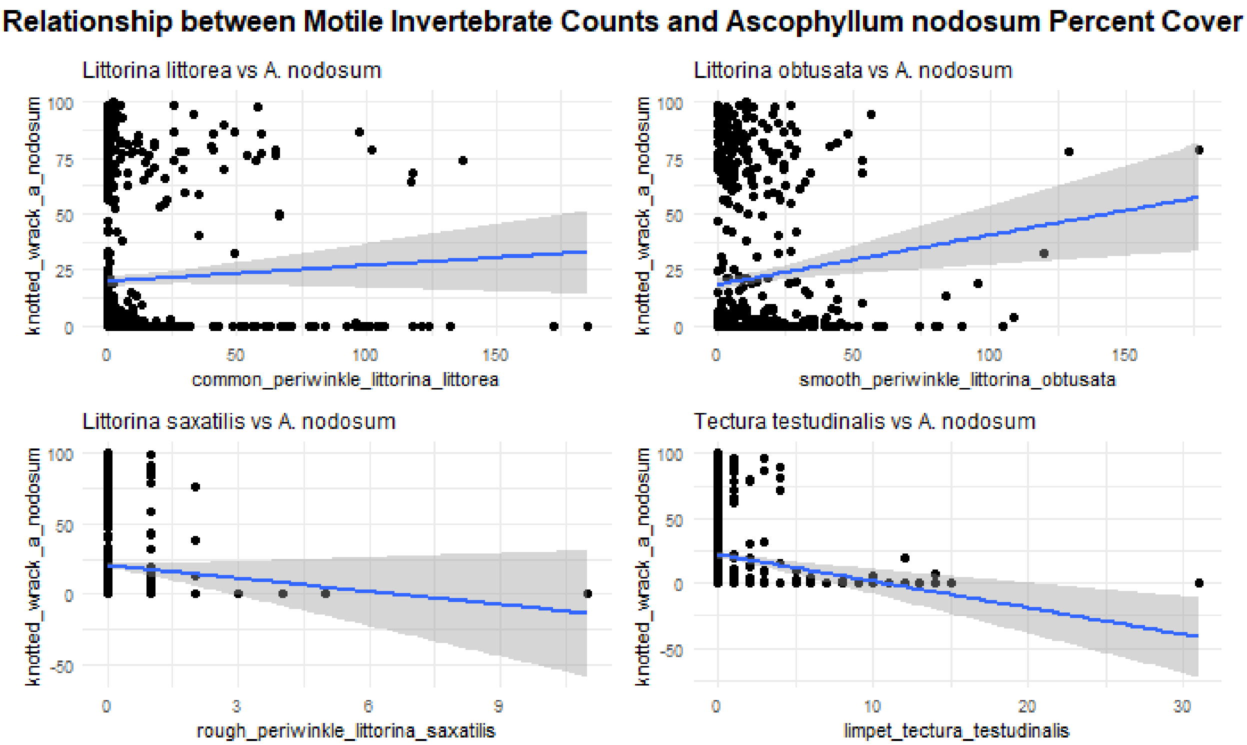


Figure 3 - Relationship between Ascophyllum nodosum percent cover and motile invertebrate abundances in NETN rocky intertidal photoplots.

The plots examining *A. nodosum* percent cover as a function of the various motile invertebrate abundances suggest that common periwinkle abundance has no meaningful relationship with *Ascophyllum* percent cover, while smooth periwinkle abundance is associated with improved percent cover, and both rough periwinkle and limpet abundances are associated with reduced percent cover. The common periwinkle trend is in line with the established, which predicted that there was no relationship between the abundance of the two species, as *L. littorea* do not preferentially feed on *A. nodosum*.

Both sets of plots suggest that there are some relationships between macroalgal percent cover and motile invertebrate abundance, which is a promising sign for the ability to predict macroalgal percent cover.

3.4 F. Vesiculosus *and* A. nodosum *Frequency Distribution*

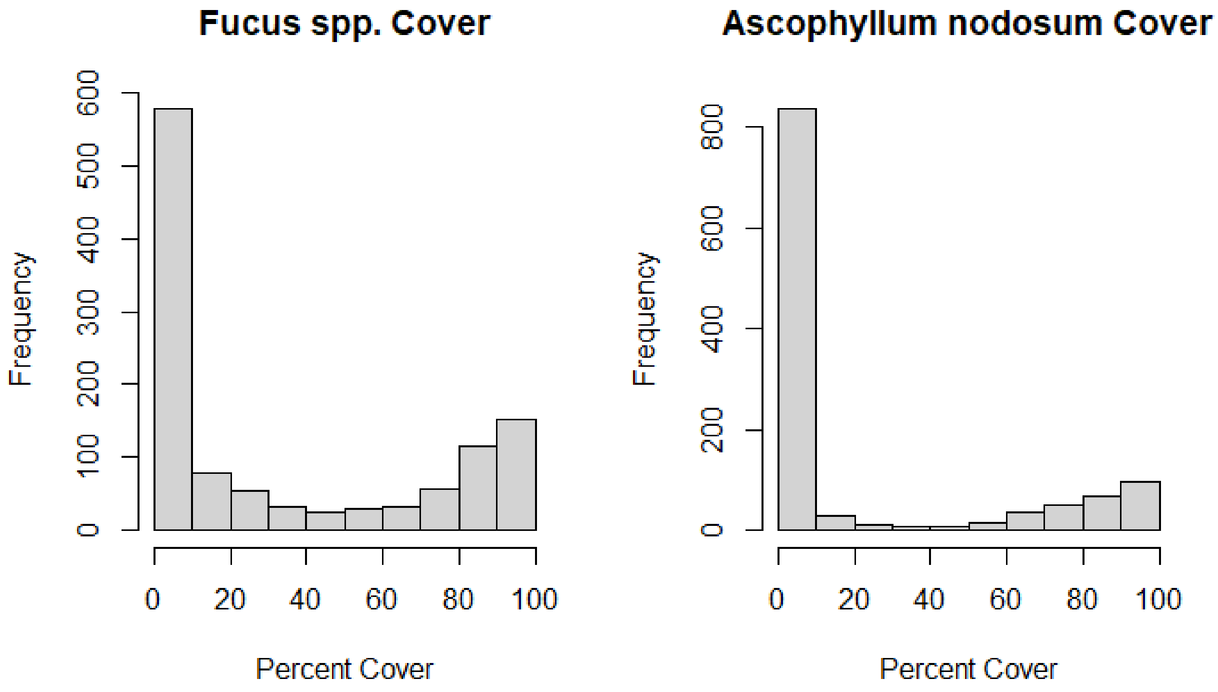


Figure 4 - Frequency distribution histograms for **Fucus** ***and*** **Ascophyllum** ***percent cover values in NETN rocky intertidal photoplots. Both target variables have heavily skewed distributions and are zero-inflated.***

Both *Fucus* and *Ascophyllum* have distributions weighted very heavily towards 0. This suggests strong spatial constraints on these species, as they are completely absent from most plots. Their next most common value is 100% cover, which suggests that, when they are present, they are able to do quite well in their niche. This zero-inflated distribution poses difficulties for modeling, as this will violate many common assumptions such as normality.

Table 1 - Summary statistics for numeric variables including percent cover data, invertebrate abundance data, and invertebrate mean measurement data. Summary statistics include mean, median, standard deviation, minimum, maximum, skew, and kurtosis.

Variable	Mean	Median	SD	Min	Max	Skew	Kurtosis
Other Algae - Brown	0.00	0.00	0.11	0	3.00	24.26	616.23
Other Algae - Green	0.55	0.00	2.95	0	47.00	8.62	93.74
Other Algae - Red	1.36	0.00	5.68	0	75.00	6.68	55.30
Articulated Corallines	1.07	0.00	5.27	0	54.00	6.72	49.50
Ascophyllum epibont	2.54	0.00	6.58	0	56.00	3.55	14.34
Knotted Wrack (A. nodosum)	20.66	0.00	35.12	0	100.00	1.31	-0.09
Barnacle (e.g. S. balanoides)	15.81	0.00	29.05	0	100.00	1.75	1.58
Irish Moss (Chondrus / Mastocarpus)	12.01	0.00	25.16	0	100.00	2.01	2.68
Crustose coraline	0.07	0.00	1.21	0	39.00	29.61	939.51
Fucus epibiont	1.09	0.00	4.66	0	96.00	13.81	253.43
Rockweed (Fucus) spp.	33.20	10.00	38.32	0	100.00	0.64	-1.33
Kelp (e.g.Laminaria/Alaria)	0.03	0.00	0.56	0	17.00	25.43	725.61
Mussel (e.g. Mytilus edulis)	1.49	0.00	6.84	0	60.00	5.92	37.64
Crustose non-coraline	1.46	0.00	5.14	0	55.00	5.56	36.61
Not Sampled	0.57	0.00	2.59	0	54.00	14.41	255.32
Other Invertebrate	0.17	0.00	0.59	0	6.00	4.54	25.29
Other Plant	0.00	0.00	0.00	0	0.00	NaN	NaN
Other Substrate	0.53	0.00	1.49	0	17.00	4.28	24.73
Dulse (Palmaria palmata)	1.33	0.00	5.32	0	51.00	5.56	35.40
Laver (Porphyra spp.)	0.61	0.00	3.11	0	46.33	8.96	99.18
Rock	3.44	1.00	7.50	0	58.00	3.82	17.44
Sand	0.01	0.00	0.26	0	9.00	33.93	1150.01
Tar	0.00	0.00	0.00	0	0.00	NaN	NaN
Grass kelp (Ulva intestinalis)	0.01	0.00	0.12	0	3.00	17.71	359.18
Sea Lettuce (Ulva lactuca)	1.90	0.00	5.57	0	56.00	4.54	25.46
Unidentified	0.06	0.00	0.34	0	5.00	7.66	76.79
Green crab (Carcinus maenas)	0.00	0.00	0.00	0	0.00	NaN	NaN
Asian shore crab (H. sanguineus)	0.00	0.00	0.00	0	0.00	NaN	NaN
Common periwinkle (Littorina littorea)	8.26	1.00	20.62	0	185.00	4.08	19.62
Smooth periwinkle (Littorina obtusata)	6.34	1.00	14.13	0	177.00	4.92	35.90
Rough periwinkle (Littorina saxatilis)	0.09	0.00	0.50	0	11.00	11.68	205.17
Dogwhelk (Nucella lapillus)	3.05	0.00	8.50	0	117.00	6.11	52.88
Limpet (Tectura testudinalis)	0.71	0.00	2.12	0	31.00	5.54	47.74
Smooth periwinkle (Littorina obtusata) Mean Measure	5.23	7.75	4.90	0	21.00	-0.03	-1.74
Common periwinkle (Littorina littorea) Mean Measure	8.76	9.50	8.74	0	29.00	0.19	-1.61
Dogwhelk (Nucella lapillus) Mean Measure	8.45	0.00	10.72	0	42.00	0.70	-1.11
Limpet (Tectura testudinalis) Mean Measure	2.16	0.00	4.27	0	17.00	1.58	0.82
Rough periwinkle (Littorina saxatilis) Mean Measure	0.45	0.00	1.93	0	16.00	4.36	18.83
Green crab (Carcinus maenas) Mean Measure	0.00	0.00	0.00	0	0.00	NaN	NaN

Some immediately noteworthy trends in the summary statistics table are the high frequency of zeroes (many variables have a median value of zero), which suggests a very zero heavy dataset. Some variables have no data at all, which is either an error, suppression, or suggests total absence from the plots. The target variables of *F. vesiculosus* (or *spp.*) and *A. nodosum* have moderately high average cover, but have high standard deviation, suggesting high variability in percent cover. They are both relatively present but also have spikes of high concentration in certain plots.

3.6 Hellinger-Transformed Principal Component Analysis Scree Plot

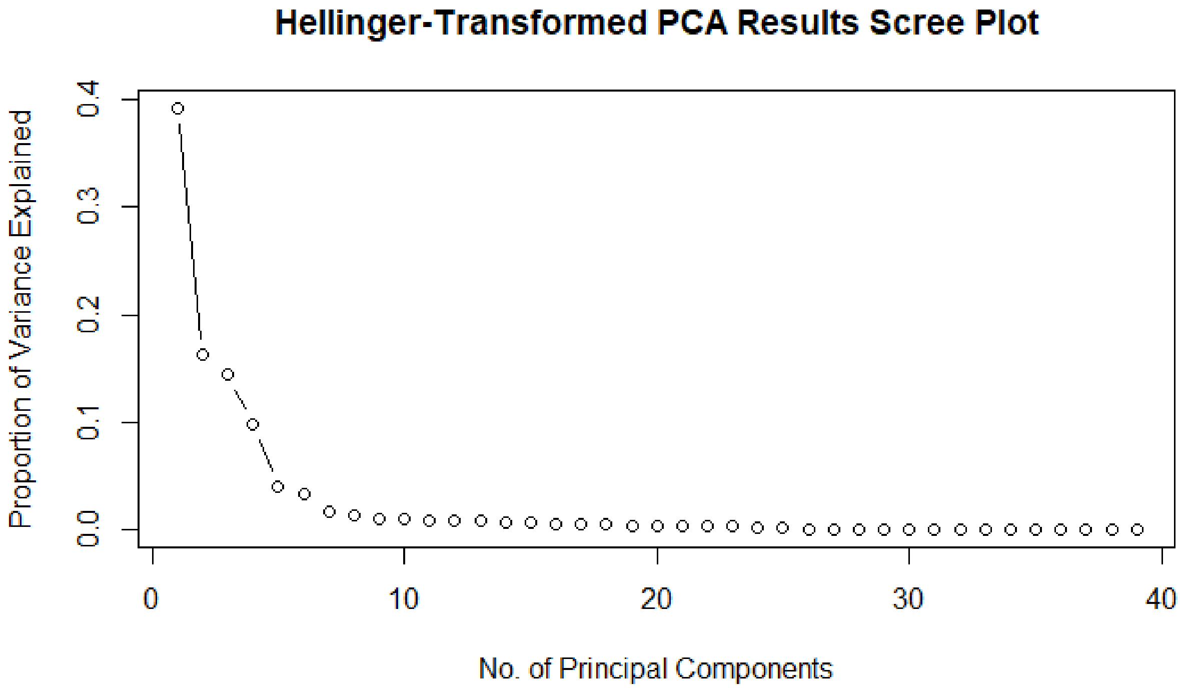


Figure 5 – Hellinger Transformed Principal Component Analysis proportion of variance explained by each principal component. Variance explained per PC is low, drops off precipitously between the first three PCs, and has a consistent but slow decline afterwards.

The scree plot shows that each individual principal component from the PCA is capturing very little of the variability in the data. PC1 accounts for around 39% of the variability, and it takes a large number of PCs to explain the majority of the variability, with a precipitous drop in variance explained after PC1. This suggests that the data is weakly structured and high dimensional.

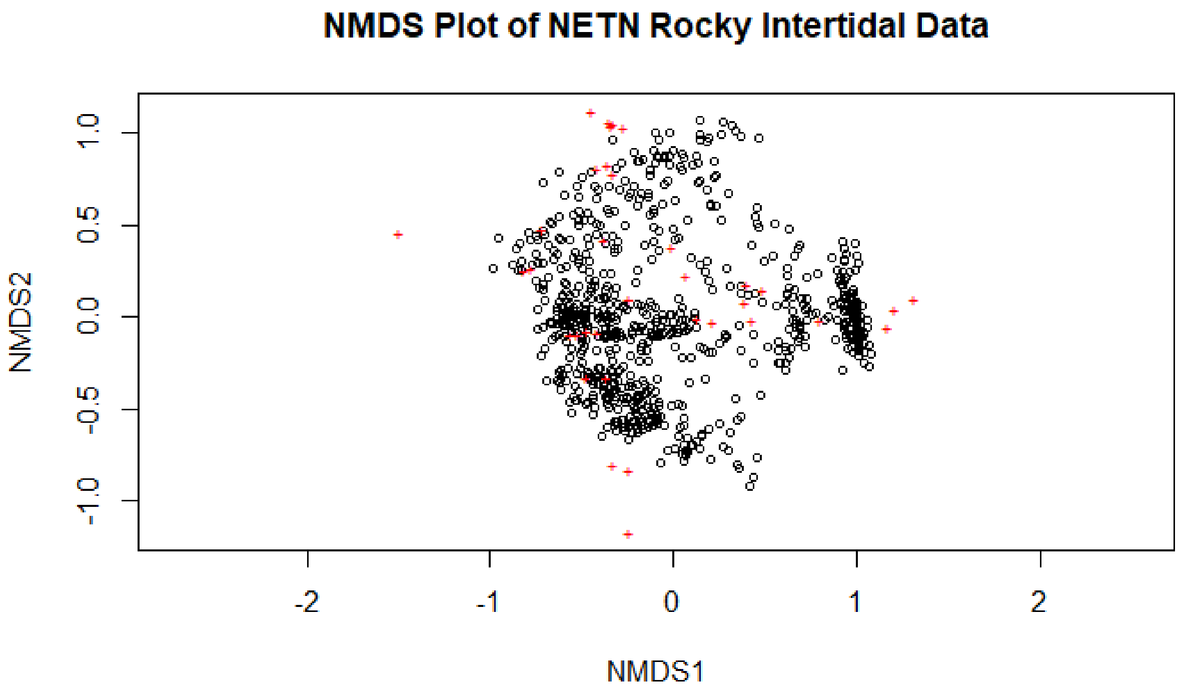


Figure 6 – Non-parametric multidimensional scaling (NMDS) clustering plot for intertidal community data in Acadia National Park. Black dots represent individual observations while red crosses represent features. There is little evidence of meaningful clustering within observations or features, suggesting no clear path for reducing dimensionality.

NMDS also failed to reveal any meaningful clustering, which suggests that there may not be an effective way to reduce dimensionality using unsupervised clustering methods. Non-parametric and decision tree models that do not fall victim to the curse of dimensionality and do not require normally distributed data may have the best performance when working with this dataset.

4 Modeling Methods

4.1 Data Preprocessing

Multiple transformations were tested to address the zero-inflated nature of the data, in order to open up the linear regression family of models, which generally assume normally distributed data.

4.1.1 Square Root Transformation

A square root transformation was used as a conservative starting point. The square root transformation did not result in a meaningful improvement in the distribution of the target variables, leaving both of them still skewed heavily towards zero.

4.1.2 Log (x + 1) Transformation

The Log (X + 1) transformation is intended to address right-skewed data that still contains zeroes. While this data is heavily zero-inflated and therefore left-skewed, the non-zero data is actually skewed to the right, giving this transformation possible merit. However, this transformation did not do much to ameliorate the zero skewed distribution.

4.1.3 Presence - Absence Transformation

As a drastic measure, all the count and percent cover variables were converted to presence-absence data. While this did improve the distribution of the data meaningfully, it dramatically reduces the efficacy of the data and the interpretability of it. In the end it was deemed too limiting of a transformation to continue with.

4.1.4 Logit Transformation

The logit transformation is generally used to transform probability data bounded between 0 and 1 into log odds, which is unbounded, ranging from positive to negative infinity. While this dataset does not contain probability data, it does contain percent cover data, which is similarly bounded between 0 and 100. A logit transformation was applied to only the percent cover variables, as the abundance data is not properly bounded. The logit transformation with an epsilon of 0.0001 was fairly successful in reducing the skew of the data, with the untransformed Ascophyllum and Fucus having skew values of 1.282 and 0.672, respectively, and the transformed *Ascophyllum* and *Fucus* having skew values of 0.767 and 0.067, respectively. This represents a more symmetrical distribution post-transformation which should improve model performance.

4.2 Feature Selection

4.2.1 Manual Feature Selection

Location name, target species, and plot name were manually removed. Location name was first filtered to only include observations from Acadia National Park, then was removed as a column, as this analysis is only focused on Acadia, not the Boston Harbor Islands. Target species, the value referring to the original target species of the plot (i.e., Ascophyllum plots, red algae plots, etc.) was removed, as separating the site into areas that are historically associated with a specific species is not a helpful lens for this analysis. Plot name was removed, as they are arbitrary labels for the different plots (i.e., R5 for the fifth red algae plot on any given site) and do not contain any meaningful information. There is no meaningful comparison between an R5 plot on one site and an R5 plot on another, aside from the fact that they are both in the red algae area.

4.2.2 Unsupervised Feature Selection

Two out-of-bag random forest models were run with *F. vesiculosus* and *A. nodosum* as target variables, with all other variables as predictors. Importance metrics, represented by percent increase in mean squared error (%incMSE) were pulled from both models. Any feature with a percent increase in mean squared error of 0 or below for either target species was removed from the dataset. Percent increase in mean squared error in this instance is a measure of how much the mean squared error increases when an individual variable is imputed randomly. This means that values of zero represent variables that are not significantly contributing to the model's performance, and, in the case of negative values, are actively adding noise.

5 Model Evaluations

Prior to any modeling, a time lagged version of the data was created, lagged by one year. This allows for the use of time lagged regression models to actually predict the percent cover of *A. nodosum* and *F. vesiculosus*, rather than just descriptive models for within years. To that end, a training-testing split was created by selecting 2018 as an entire year of held out data, as it is the most recent year in the available data. 2013 – 2017 serve as the training set for training the models, with the goal of accurately predicting the values contained within 2018.

5.1 *Random Forest* (*untransformed and logit-transformed*)

Random Forest models were tuned by adjusting two primary hyperparameters: the number of trees grown and the number of predictor variables randomly sampled at each split (*mtry*).

The number of trees was cross validated along a sequence from 1 to 1000, in increments of 50. This range provided a sufficiently broad range combined with a relatively granular step size for optimizing performance. Efficiency was not quantified, though noticeable increases in runtime with this hyperparameter matrix.

For *mtry*, values ranging from 2 up to the total number of predictors were tested in increments of 2. The lower bound of 2 was selected to avoid overly simplistic splits, while the inclusion of the full predictor set allowed the model to make full use of the data.

All combinations of hyperparameters were evaluated in a nested loop. Model performance was tracked using R^2, mean squared error (MSE), and root mean squared error (RMSE). The best-performing model was selected based on the lowest MSE/RMSE, ensuring optimization for unseen test data rather than training data. This prevented selection of an overfit model, despite the large hyperparameter ranges and number of parameters which may allow for overfitting through high complexity.

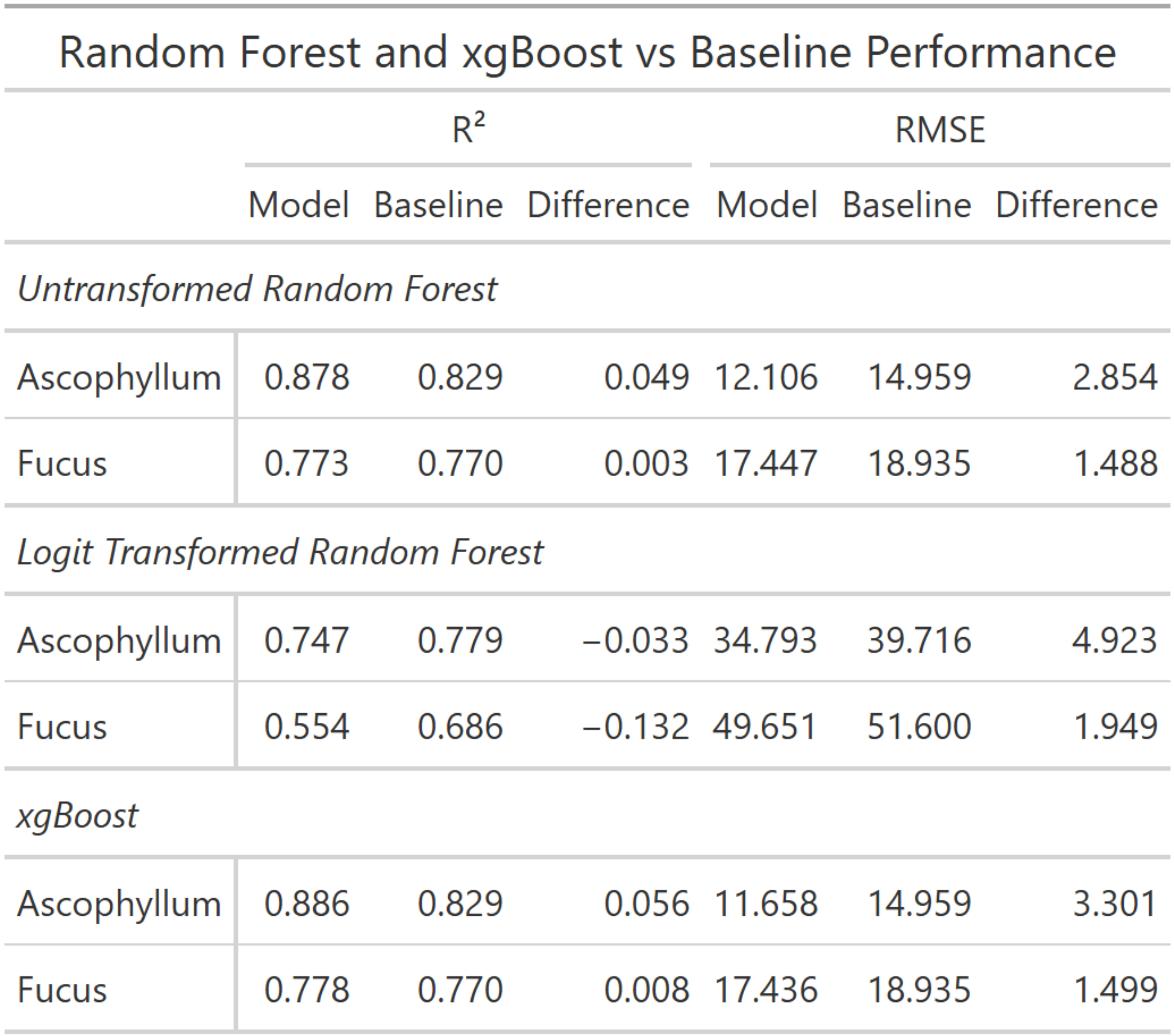
5.2 eXtreme Gradient Boosting

An eXtreme Gradient Boosting (XGBoost) model was trained on the untransformed data, using a similar nested loop structure for hyperparameter tuning. Since XGBoost has a larger number of relevant hyperparameters than Random Forest, the hyperparameter matrix was more restricted by computational efficiency. Learning rate (η) was tested across values of 0.05, 0.1, and 0.2. Moderate values were selected as a high learning rate resulted in the model fixing to the high number of zeroes in the data. Maximum tree depth was tested at 3, 5, and 7 to avoid overfitting to the sparse data and relatively small number of years in the sample size. Row subsampling was restricted to moderate values of 0.7 and 0.8, as well as just sampling the entire set with a value of 1. These moderate values were meant to introduce some stochasticity into the data to help counteract the static nature of sampling the same plots at regular intervals. Random feature sampling per tree was also kept at moderate values of 0.7, 0.8, and, as a baseline, 1. The values here introduce more stochasticity and also allow less impactful features to be examined, when the more dominant features are randomly selected out of the model.

5.3 Final Model Selection

All three models were evaluated based on their R^2, MSE, and RMSE values, and were compared to the baseline of predicting purely using the lagged variables. The results are summarized in Table 2, which shows that both XGBoost and the untransformed Random Forest perform well, with high R^2 values and relatively low MSE values, while the logit-transformed Random Forest performed comparatively poorly. Both the untransformed Random Forest and the XGBoost models were more accurately able to predict the percent cover value of *A. nodosum* than *F. vesiculosus*. Both models represent only moderate improvement over the baseline, with between a ~10% and ~20% reduction in MSE magnitude. XGBoost narrowly outperforms Random Forest for predicting both target variables and is computationally more efficient while still maintaining interpretability via importance metrics, so it is the standout model here.

Table 2 - Summary table comparing Random Forest, Logit-transformed Random Forest, and XGBoost model performance over baseline predictions generated directly from the 1-year lag terms. Untransformed Random Forest and XGBoost show moderate improvements in performance over the baseline, which is already a strong predictor for the following year. Logit-transformed Random Forest (with back-transformed MSE and RMSE values) shows a proportionally larger increase in performance compared to baseline, though the overall performance remains lower than the other two models. XGBoost performs the best for predicting both Ascophyllum and Fucus percent cover in future years / held out test data.



Random Forest and xgBoost vs Baseline Performance

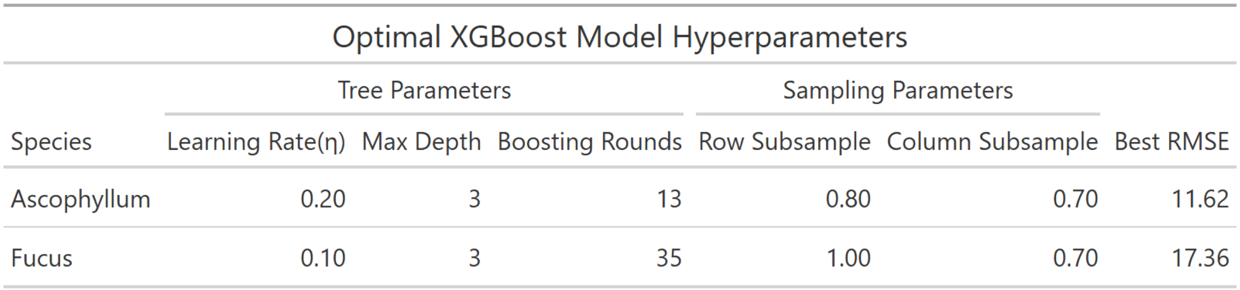
	R^2			RMSE		
	Model	Baseline	Difference	Model	Baseline	Difference
Untransformed Random Forest						
Ascophyllum	0.878	0.829	0.049	12.106	14.959	2.854
Fucus	0.773	0.770	0.003	17.447	18.935	1.488
Logit Transformed Random Forest						
Ascophyllum	0.747	0.779	−0.033	34.793	39.716	4.923
Fucus	0.554	0.686	−0.132	49.651	51.600	1.949
xgBoost						
Ascophyllum	0.886	0.829	0.056	11.658	14.959	3.301
Fucus	0.778	0.770	0.008	17.436	18.935	1.499

5.4 eXtreme Gradient Boosting Model Interpretation

The optimal hyperparameters for the XGBoost model (detailed in Table 3) illustrate a difference in the learning patterns of the model when predicting *A. nodosum* compared to *F. vesiculosus*. When predicting *A. nodosum*, the model prefers an aggressive approach with a higher learning rate (η) and a smaller number of boosting rounds. The model also prefers moderate subsampling of both features and observations. This suggests that, when predicting *Ascophyllum*, the model is able to relatively quickly fix to the data and generate its optimal predictions.

When predicting *Fucus*, the model prefers a more conservative approach, with a moderately low learning rate (η) and a high number of boosting rounds. This model still prefers a moderate subsampling of features to introduce diversity into trees but does best when incorporating the entire sample of observations. This suggests that the relationships present that contribute to *Fucus* percent cover may be more complex and difficult to sort out for the model, requiring a more methodical, thorough, and computationally slow approach.

Table 3 - Optimal hyperparameters for XGBoost model. When predicting Ascophyllum, the model prefers a more aggressive approach with a high learning rate and a lower number of boosting rounds. When predicting Fucus, the model prefers a more conservative approach with a moderate learning rate, a larger number of boosting rounds, and inclusion of all observations.



Optimal XGBoost Model Hyperparameters

	Tree Parameters			Sampling Parameters		
Species	Learning Rate(η)	Max Depth	Boosting Rounds	Row Subsample	Column Subsample	Best RMSE
Ascophyllum	0.20	3	13	0.80	0.70	11.62
Fucus	0.10	3	35	1.00	0.70	17.36

5.5 SHapley Additive exPlanations

SHapley Additive exPlanations (SHAP) was used to explore the importance of each predictor in contributing to model predictions of *Ascophyllum* and *Fucus*. Figures 6 and 7 examine SHAP plots for prediction of *A. nodosum* and *F. vesiculosus*, respectively. High values for a predictor indicate a strong influence on model performance; high SHAP values indicate a positive correlation with the target variable (an increase in percent cover) while negative values indicate a negative correlation with the target variable (a decrease in percent cover). Model values are included to indicate how high and low values influence the target variables.

Figure 6 shows that the strongest predictors of *A. nodosum* percent cover are the previous year's percent cover of *A. nodosum* (SHAP = 16.238), followed by *Ascophyllum* epibiont (SHAP = 5.248), *F. vesiculosus* percent cover (SHAP = 3.667), and *Chondrus/Mastocarpus spp.* percent cover (SHAP = 2.271). The previous year's percent cover of *A. nodosum* and *A. nodosum* epibiont coverage have strong positive correlations with the following year's percent cover, with high values increasing predicted percent cover. High values of *F. vesiculosus* and *Chondrus/Mastocarpus spp.* percent cover are associated with a reduction in *A. nodosum* percent cover.

Figure 7 illustrates that the strongest predictors of *F. vesiculosus* percent cover are the previous year's percent cover of *F. vesiculosus* (SHAP = 20.382), *A. nodosum* percent cover (SHAP = 4.789), *Chondrus/Mastocarpus spp.* percent cover (SHAP = 4.292), and *Littorina obtusata* count (SHAP = 3.840). The previous year's percent cover of *F. vesiculosus* has a strong positive influence on the following year's percent cover, with high values increasing percent cover. *A. nodosum* and *Chondrus/Mastocarpus spp.* percent cover is negatively correlated, with high values reducing *F. vesiculosus* percent cover. *L. obtusata* count shows some positive association with *F. vesiculosus* percent cover, with high values improving percent cover.

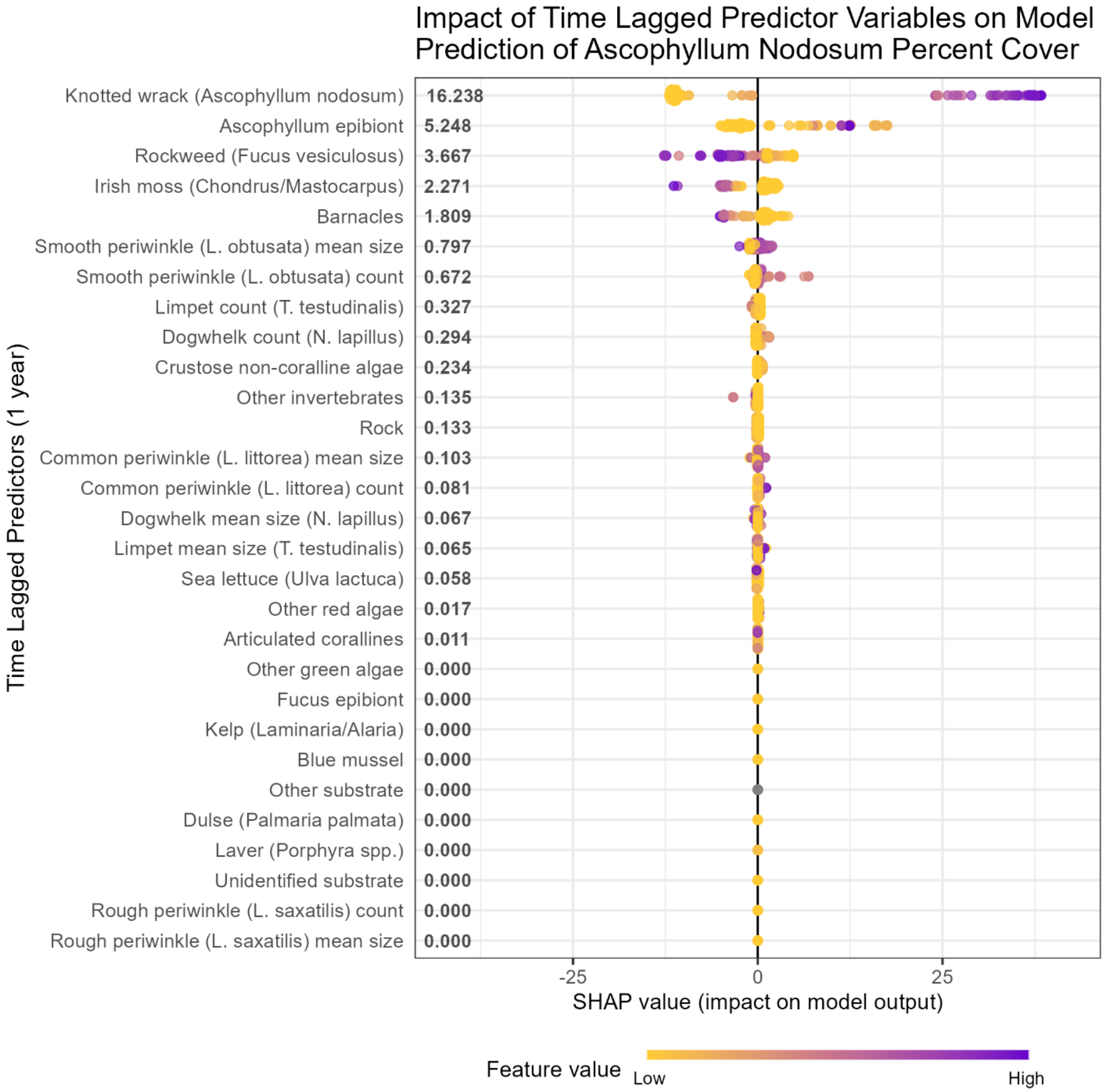


Figure 6 - SHAP plot showing the impact of time lagged predictor variables on model prediction of Ascophyllum nodosum percent cover in the following year. The previous year's Ascophyllum percent cover and Ascophyllum epibiont values are the most dominant predictors, suggesting that the pure time lagged percent cover is a good indicator of the following year's percent cover. High values of Fucus percent cover are also associated with reduced Ascophyllum percent cover, which may indicate competition or a direct harmful effect (i.e., allelopathy).

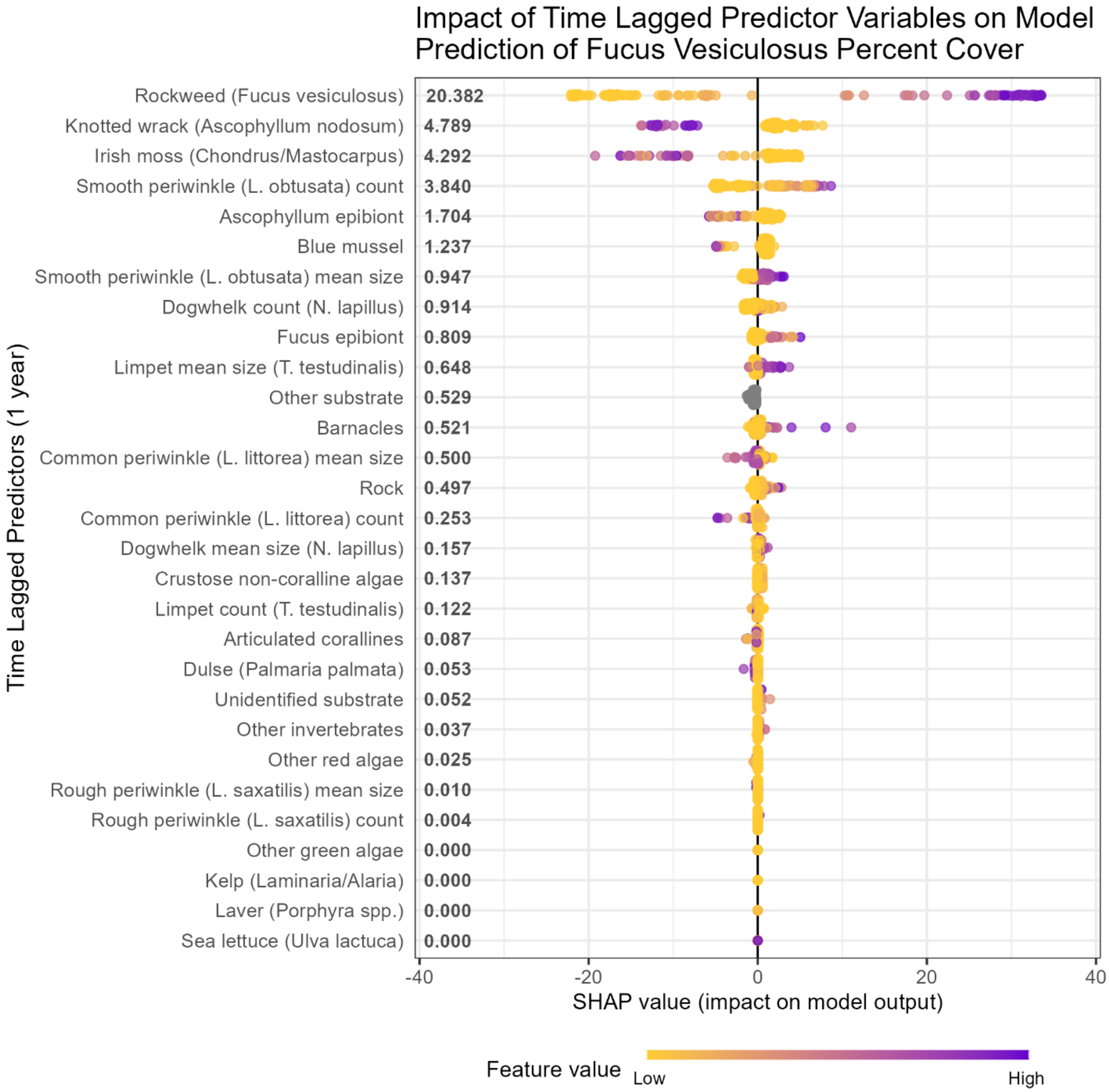


Figure 7 - SHAP plot showing the impact of time lagged predictor variables on model prediction of Fucus vesiculosus percent cover in the following year. F. vesiculosus percent cover from the previous year shows strong influence on model performance and prediction magnitude, with high and low values associated with high and low values the following year, respectively. A. nodosum and Chondus/Mastocarpus percent cover are inversely associated with F. vesiculosus percent cover, both contributing moderately to model performance and reducing F. vesiculosus percent cover. L. obtusata count is slightly positively associated with F. vesiculosus percent cover.

6 Discussion

6.1 Model Performance and Conclusions

XGBoost shows good performance overall with a moderately high R^2 value of 0.889 for predicting *A. nodosum* and 0.780 for predicting *F. vesiculosus*, and a moderate RMSE value of 11.624 and 17.361 for predicting each, respectively. XGBoost also shows a moderate improvement on the baseline model, with a reduction in RMSE of 22.29% when predicting *A. nodosum* and 8.31% when predicting *F. vesiculosus*. This suggests that, while the model is performing well overall and is improving upon the baseline moderately well for *A. nodosum*, it shows only moderate success and slight improvement compared to the baseline when predicting *F. vesiculosus*. The relationships that determine *F. vesiculosus* percent cover may be more complex or less well captured by the study design. There may also be room for improvement in fitting the model for predicting *F. vesiculosus* specifically. In it's current state, the model may prove adequate for improving park management, depending on the specific needs of the stakeholders. It is at least a step in the right direction, if not sufficiently accurate for management planning.

6.2 Model Importance Metrics

The XGBoost SHAP results suggest that the previous year's *A. nodosum* percent cover is a strong indicator of the following year's, with both high and low values contributing strongly positively and negatively, respectively. These results also indicate that *F. vesiculosus* and *Chondrus/Mastocarpus spp.* may be competing with *A. nodosum*, as they are associated with a reduction in predicted percent cover. This is somewhat supported by the literature, as these macroalgal species have been shown to compete with each other for space and light, though the interaction is complex and, under specific environmental conditions (i.e., desiccation) the competition skews towards mutualism (Choi and Norton 2005). While these associations may be evidence of competition, they may also be an indicator of complex, unseen, environmental factors that contribute to the distribution of these macroalgal species. For example, *A. nodosum* has been shown to shade other macroalgal species with its long fronds, outcompeting them for light (Cervin et al. 2004) and *F. vesiculosus* has been shown to exhibit allelopathy (Budzałek and Śliwińska-Wilczewska 2021), but *A. nodosum* and *F. vesiculosus* have also been shown to have different tolerances for salinity (Bäck et al. 1992, Perry et al. 2020). It is therefore impossible to state what exact effects are influencing *A. nodosum* and *F. vesiculosus* distribution, only that certain features are associated with them. Being able to single out associated features, their importance, and their effect magnitude and direction is, however, sufficient for the research question posed. Interestingly, very few motile invertebrates were strongly associated with the percent cover of the target variables. Only the smooth periwinkle, *L. obtusata*, was moderately associated with *F. vesiculosus*.

6.3 Future Analysis

The XGBoost model described in this analysis is a good starting point for accurately predicting key macroalgal species future abundance. The model in its current state may be underfit, as performance on both training and testing data could be improved. Future analysis with the full dataset (2008 – 2024, rather than 2013 – 2018) may allow for improved performance via increased lag times and the inclusion of temperature as a key covariate. Normalizing the epibiont percent cover values (and other collinear values that may appear with the full dataset) by their respective host's percent cover will allow that to serve as an indication of how much of that target species has been colonized, rather than basically a proxy for lagged percent cover, which is already present. Inclusion of *Carcinas maenas*, the European Green Crab, a critically important invasive species in the rocky intertidal, may provide additional predictive power and valuable interpretation, depending on its contribution to the model.

7 Appendix A: GitHub

This project is available in full from https://github.com/nicklagoni/DSE6311/.

8 Appendix B: Data Dictionary

8.1 NETN Rocky Intertidal Long Term Monitoring Protocol Data Dictionary

VARIABLE NAME	DESCRIPTION	UNITS / SCALE	TRANSFORMATION	TYPE
LOC_NAME	Site name within Acadia National Park where survey was conducted (6 total sites)	Categorical (string)	None	Identifier
START_DATE	Date of annual survey for that site/year	Date (MM/DD/YYYY)	None	Identifier
ROCKWEED_FUCUS_SPP	Percent cover of rockweed (*Fucus* spp.)	Percent cover (%)	None/Logit transformed	Numeric
KNOTTED_WRACK_A_NODOSUM	Percent cover of knotted wrack (*Ascophyllum nodosum*)	Percent cover (%)	None/Logit transformed	Numeric
BARNACLE_E_G_S_BALANOIDES	Percent cover of barnacles (*e.g.*, *Semibalanus balanoides*)	Percent cover (%)	None/Logit transformed	Numeric
MUSSEL_E_G_MYTILUS_EDULIS	Percent cover of mussels (*e.g.*, *Mytilus edulis*)	Percent cover (%)	None/Logit transformed	Numeric
IRISH_MOSS_CHONDRUS_MASTOCARPUS	Percent cover of Irish moss (*Chondrus crispus*, *Mastocarpus stellatus*)	Percent cover (%)	None/Logit transformed	Numeric
KELP_E_G_LAMINARIA_ALARIA	Percent cover of kelp (*e.g.*, *Laminaria* spp., *Alaria* spp.)	Percent cover (%)	None/Logit transformed	Numeric
DULSE_PALMARIA_PALMATA	Percent cover of dulse (*Palmaria palmata*)	Percent cover (%)	None/Logit transformed	Numeric
LAVER_PORPHYRA_SPP	Percent cover of laver (*Porphyra* spp.)	Percent cover (%)	None/Logit transformed	Numeric
SEA_LETTUCE_ULVA_LACTUCA	Percent cover of sea lettuce (*Ulva lactuca*)	Percent cover (%)	None/Logit transformed	Numeric
ARTICULATED_CORALLINES	Percent cover of articulated coralline algae	Percent cover (%)	None/Logit transformed	Numeric
CRUSTOSE_NON_CORALINE	Percent cover of crustose non-coralline algae	Percent cover (%)	None/Logit transformed	Numeric
OTHER_ALGAE_GREEN	Percent cover of other green algae species	Percent cover (%)	None/Logit transformed	Numeric
OTHER_ALGAE_RED	Percent cover of other red algae species	Percent cover (%)	None/Logit transformed	Numeric
FUCUS_EPIBIONT	Percent cover of epibionts on *Fucus* spp.	Percent cover (%)	None/Logit transformed	Numeric
ASCOPHYLLUM_EPIBONT	Percent cover of epibionts on *Ascophyllum nodosum*	Percent cover (%)	None/Logit transformed	Numeric
OTHER_INVERTEBRATE	Percent cover of other sessile invertebrate species (non-listed)	Percent cover (%)	None/Logit transformed	Numeric
ROCK	Percent cover of bare rock substrate	Percent cover (%)	None/Logit transformed	Numeric

OTHER_SUBSTRATE	Percent cover of substrate types not listed elsewhere	Percent cover (%)	None/Logit transformed	Numeric
NOT_SAMPLED	Percent cover of quadrat area not surveyed or unable to be sampled	Percent cover (%)	None/Logit transformed	Numeric
COMMON_PERIWINKLE_LITTORINA_LITTOREA	Abundance of common periwinkle (*Littorina littorea*) per plot	Count	None	Numeric
SMOOTH_PERIWINKLE_LITTORINA_OBTUSATA	Abundance of smooth periwinkle (*Littorina obtusata*) per plot	Count	None	Numeric
ROUGH_PERIWINKLE_LITTORINA_SAXATILIS	Abundance of rough periwinkle (*Littorina saxatilis*) per plot	Count	None	Numeric
DOGWHELK_NUCELLA_LAPILLUS	Abundance of dogwhelk (*Nucella lapillus*) per plot	Count	None	Numeric
LIMPET_TECTURA_TESTUDINALIS	Abundance of limpet (*Tectura testudinalis*) per plot	Count	None	Numeric
COMMON_PERIWINKLE_LITTORINA_LITTOREA_MEAN_MEASURE	Mean shell length of common periwinkle (average of up to 10 individuals if abundance > 10)	Millimeters (mm)	None	Numeric
SMOOTH_PERIWINKLE_LITTORINA_OBTUSATA_MEAN_MEASURE	Mean shell length of smooth periwinkle	Millimeters (mm)	None	Numeric
ROUGH_PERIWINKLE_LITTORINA_SAXATILIS_MEAN_MEASURE	Mean shell length of rough periwinkle	Millimeters (mm)	None	Numeric
DOGWHELK_NUCELLA_LAPILLUS_MEAN_MEASURE	Mean shell length of dogwhelk	Millimeters (mm)	None	Numeric
LIMPET_TECTURA_TESTUDINALIS_MEAN_MEASURE	Mean shell length of limpet	Millimeters (mm)	None	Numeric

References

Bäck, S., J. C. Collins, and G. Russell. 1992. Effects of salinity on growth of Baltic and Atlantic *Fucus vesiculosus*. British Phycological Journal 27:39–47.

Budzałek, G., and S. Śliwińska-Wilczewska. 2021. Allelopathic effect of macroalgae Fucus vesiculosus (Ochrophyta) and Coccotylus brodiei (Rhodophyta) on the growth and photosynthesis performance of Baltic cyanobacteria. Annales Universitatis Paedagogicae Cracoviensis Studia Naturae:81–94.

Cervin, G., M. Lindegarth, R. M. Viejo, and P. Åberg. 2004. Effects of small-scale disturbances of canopy and grazing on intertidal assemblages on the Swedish west coast. Journal of Experimental Marine Biology and Ecology 302:35–49.

Choi, H. G., and T. A. Norton. 2005. Competition and facilitation between germlings of Ascophyllum nodosum and Fucus vesiculosus. Marine Biology 147:525–532.

Clarke, K. R. 1993. Non-parametric multivariate analyses of changes in community structure. Australian Journal of Ecology 18:117–143.

Dayton, P. K. 1972. Toward an understanding of community resilience and the potential effects of enrichments to the benthos at McMurdo Sound, Antarctica. Page Proceedings of the colloquium on conservation problems in Antarctica. Allen Press Lawrence, KS.

Dexter, E., G. Rollwagen-Bollens, and S. M. Bollens. 2018. The trouble with stress: A flexible method for the evaluation of nonmetric multidimensional scaling. Limnology and Oceanography: Methods 16:434–443.

Hadlock, R. H. 1979. The distribution of Littorina obtusata (L.) in the rocky intertidal: effects of competition with Littorina littorea (L.). Master's thesis, Department of Zoology, University of Rhode Island, Kingston, RI.

Jueterbock, A., L. Tyberghein, H. Verbruggen, J. A. Coyer, J. L. Olsen, and G. Hoarau. 2013. Climate change impact on seaweed meadow distribution in the North Atlantic rocky intertidal. Ecology and evolution 3:1356–1373.

Linnaeus, C. 1753. Species plantarum, exhibentes plantas rite cognitas ad genera relatas cum differentiis specificis, nominibus trivialibus, synonymis selectis, locis natalibus, secundum systema sexuale digestas. Vol 1.

Linné, C. von, and L. Salvius. 1758. Caroli Linnaei...Systema naturae per regna tria naturae :secundum classes, ordines, genera, species, cum characteribus, differentiis, synonymis, locis. Impensis Direct. Laurentii Salvii, Holmiae.

Marbà, N., D. Krause-Jensen, B. Olesen, P. B. Christensen, A. Merzouk, J. Rodrigues, S. Wegeberg, and R. T. Wilce. 2017. Climate change stimulates the growth of the intertidal macroalgae Ascophyllum nodosum near the northern distribution limit. Ambio 46:119–131.

Minchin, P., and L. Rennie. 2010. Does the Hellinger transformation make PCA a viable method for community ordination? 95th Annual ESA Meeting Contributed Oral Papers.

Northeast Temperate Network. 2021, January 1. Long-term Rocky Intertidal - Database. https://irma.nps.gov/DataStore/Reference/Profile/2289832.

Parrot, D., M. Blümel, C. Utermann, G. Chianese, S. Krause, A. Kovalev, S. N. Gorb, and D. Tasdemir. 2019. Mapping the Surface Microbiome and Metabolome of Brown Seaweed Fucus vesiculosus by Amplicon Sequencing, Integrated Metabolomics and Imaging Techniques. Scientific Reports 9:1061.

Perry, F., E. d'Avack, and J. Hill. 2020. Ascophyllum nodosum and Fucus vesiculosus on variable salinity mid eulittoral rock.

Putnam, A. B., S. C. Endyke, A. R. Jones, L. A. D. Lockwood, J. Taylor, M. Albert, and M. D. Staudinger. 2024. Historical insights, current challenges: tracking marine biodiversity in an urban harbor ecosystem in the face of climate change. Marine Biodiversity 54.

Råberg, S., and L. Kautsky. 2007. A comparative biodiversity study of the associated fauna of perennial fucoids and filamentous algae. Estuarine, Coastal and Shelf Science 73:249–258.

Trott, T. J. 2022. Mesoscale Spatial Patterns of Gulf of Maine Rocky Intertidal Communities. Diversity 14:557.

Watson, D. C., and T. A. Norton. 1985. Dietary preferences of the common periwinkle, Littorinalittorea (L.). Journal of Experimental Marine Biology and Ecology 88:193–211.

Watson, D. C., and T. A. Norton. 1987. The habitat and feeding preferences of Littorina obtusata (L.) and L. mariae sacchi et rastelli. Journal of Experimental Marine Biology and Ecology 112:61–72.

Westerbom, M., and M. Koivisto. 2022. Mussels and canopy-forming algae as ecosystem engineers: their contribution to community organization in the rocky sublittoral. Frontiers in Marine Science 9.