



Data Analysis in Musquash Marine Protected Area

Data Analysis STAT 4620/5620 Winter 24-25

Authors:

Cameron Moffatt

Peike Cong

Hongren Zhu

Advisor:

Dr. Ethan Lawler

GitHub repository: github.com

Group 2 Project Report
Dalhousie University

Abstract

This project investigates biodiversity and sediment characteristics within the Musquash Marine Protected Area (MPA), designated in 2006 to protect ecologically sensitive benthic habitats. Using data collected by the department of Fisheries and Oceans (2010–2023), we derived biodiversity indices (Shannon Index, Species Richness) and assessed their association with sediment composition and environmental factors. Through statistical modeling, including Linear Regression, GAM, and Regression Trees, we found evidence of both seasonal and spatial variations in biodiversity, with sediment texture and organic content as key explanatory variables. Notably, finer sediments and intermediate organic matter levels were associated with greater infaunal diversity. Our results support the effectiveness of the MPA in maintaining ecological balance and provide insights for continued monitoring efforts.

Keyword: Musquash, Biodiversity, Sediment, Shannon Index, Regression Tree, Generalized Additive Model, Marine Protected Area, Infauna, Organic Matter, Grain Size

Contents

| | | |
|----------|--|-----------|
| 1 | Introduction | 1 |
| 2 | Data Description | 2 |
| 2.1 | Data Sources and Collection Methodology | 2 |
| 2.2 | Data Structure | 2 |
| 2.3 | Target Variable | 2 |
| 2.4 | Explanatory Variables | 2 |
| 2.5 | Data Quality Assessment | 2 |
| 3 | Methods | 4 |
| 3.1 | Data Preprocessing | 4 |
| 3.2 | Feature Selection | 5 |
| 3.3 | Model Development | 6 |
| 3.4 | Model Evaluations | 6 |
| 4 | Modeling and Feature Analysis | 7 |
| 4.1 | Part 1: Modeling Without Grouping Structures | 7 |
| 4.1.1 | Linear Regression and Stepwise Selection | 7 |
| 4.1.2 | Regularized Regression: LASSO, Ridge, and Elastic Net | 7 |
| 4.1.3 | Generalized Additive Model (GAM) | 10 |
| 4.2 | Part 2: Incorporating Random Effects | 11 |
| 4.3 | Part 3: Tree-Based Models for Uncovering Hidden Patterns | 11 |
| 4.3.1 | Regression Tree, and SHAP Analysis | 11 |
| 5 | Results | 13 |
| 5.1 | Exploratory Data Analysis | 13 |
| 5.2 | Regression Tree Analysis | 13 |
| 5.3 | Variable Importance | 14 |
| 5.4 | Spatial Distribution Patterns | 15 |
| 5.5 | Predicted Biodiversity Distribution | 15 |
| 5.6 | Influence of Seasonal and Spatial Factors | 16 |
| 5.7 | Variance Partitioning Analysis | 17 |
| 6 | Conclusions | 18 |
| 6.1 | Interpretation of results | 18 |
| 6.2 | Conservation Implications | 18 |
| 6.3 | Study limitations | 19 |
| 6.4 | Conclusion | 20 |

List of Figures

| | | |
|---|-------------------------------------|---|
| 1 | Flow Chart of Methodology | 4 |
|---|-------------------------------------|---|

| | | |
|---|---|----|
| 2 | Linear Regression Residual Plots | 7 |
| 3 | Regularization paths and tuning diagnostics across different lambda values. | 9 |
| 4 | Generalized Additive Model (GAM) Smooth Function Plot | 10 |
| 5 | Linear Models Results | 11 |
| 6 | SHAP Analysis from Regression Tree Model | 12 |
| 7 | Regression Tree | 14 |

List of Tables

| | | |
|---|--|---|
| 1 | Environmental and Biodiversity Variables Measured in Marine Sediment Samples | 3 |
|---|--|---|

1 Introduction

The Musquash Estuary represents the largest ecologically intact estuary in the Bay of Fundy, located approximately 20 kilometers southwest of Saint John, New Brunswick. Designated as a Marine Protected Area (MPA) in 2006, this ecosystem serves as a critical habitat for diverse terrestrial and marine species. The MPA designation aims to minimize human impact while permitting regulated activities through a structured management zone system. Indigenous fishing is permitted throughout the area, while commercial fishing activities face restrictions based on zones and target species [1].

This research utilizes comprehensive datasets collected by the Department of Fisheries and Oceans spanning from 2010 to September 2023. The data collection protocol involved sampling from 30 randomly distributed stations across three distinct strata (channel, subtidal, and intertidal zones), with sampling conducted up to three times annually to account for seasonal and yearly variations. Each sample was analysed for sediment grain size distribution, carbon content, and benthic species biomass.

Our study aims to evaluate how the MPA maintains biodiversity by assessing benthic species population consistency and ecosystem dynamics. Specifically, we investigate two key research questions:

- How do sediment composition explain the variation in infaunal biodiversity within the MPA?
- To what extent do seasonal and spatial factors influence biodiversity in the MPA, after accounting for environmental sediment conditions?

Understanding these relationships is crucial for effective conservation management and for evaluating the long-term success of marine protection measures.

2 Data Description

2.1 Data Sources and Collection Methodology

The dataset represents a comprehensive ecological monitoring effort within the Musquash MPA. Data collection followed standardized protocols established by Fisheries and Oceans Canada, ensuring consistency across sampling events. The sampling design incorporated stratified random sampling across three distinct habitat types: channel, subtidal, and intertidal zones. This approach enables robust statistical comparisons between these ecologically distinct areas.

2.2 Data Structure

The merged dataset contains 341 observations across 20 variables. Each observation represents a unique sampling event identified by a `set_id`, which combines station location, year, and season information. The dataset includes the following key variable groups, as shown in Table 1.

2.3 Target Variable

Our primary response variable is the shannon index, (`shannon_index`), which is an biodiversity index used to evaluate the MPA diversity. In addition, species richness was used as metric for ecosystem productivity and health.

2.4 Explanatory Variables

The explanatory variables include sediment composition metrics, particle size distribution, and organic matter content. These environmental factors are known to influence benthic community structure and biomass through multiple mechanisms, including habitat suitability, food availability, and oxygen conditions.

2.5 Data Quality Assessment

Initial data exploration revealed missing values, particularly in sediment-related variables for earlier sampling years (2010-2011). Approximately 36% of observations had missing values for sediment grain size and moisture content variables, while 18% lacked organic matter content data. No missing values were observed in biodiversity metrics or sampling metadata.

Table 1: Environmental and Biodiversity Variables Measured in Marine Sediment Samples

| Category | Variable | Description |
|------------------------------|----------------------|---|
| Sampling metadata | set_id | Unique identifier for sampling event |
| | year | Year of sampling |
| | lat | Latitude of sample |
| | lon | longitude of sample |
| | station | Sampling station identifier |
| | strata_strate | Stratification zone designation |
| | season_saison | Sampling season |
| | depth_m_profondeur_m | Water depth at sampling location (m) |
| Biodiversity metrics | total_count | Total number of organisms collected |
| | tot_wt_g | Total biomass of collected organisms (g) |
| | shannon_index | Shannon diversity index value |
| | species_richness | Number of distinct species identified |
| Sediment moisture/grain size | wet_weight | Weight of sediment sample before drying |
| | dry_weight | Weight of sediment sample after drying |
| | water_content_ratio | Ratio of water to sediment in sample |
| Particle size distribution | sand_pct | Percentage of sand in sediment sample |
| | fine_sand_pct | Percentage of fine sand in sediment sample |
| | silt_clay_pct | Percentage of silt and clay in sediment sample |
| Organic matter content | air_dry_wt | Weight after air drying |
| | oven_dry_wt | Weight after oven drying |
| | loss_ash1 | Weight loss after combustion at 475°C |
| | loss_ash2 | Weight loss after combustion at 950°C |
| | tot_perc_loss | Total percentage of weight lost during combustion |

3 Methods

3.1 Data Preprocessing

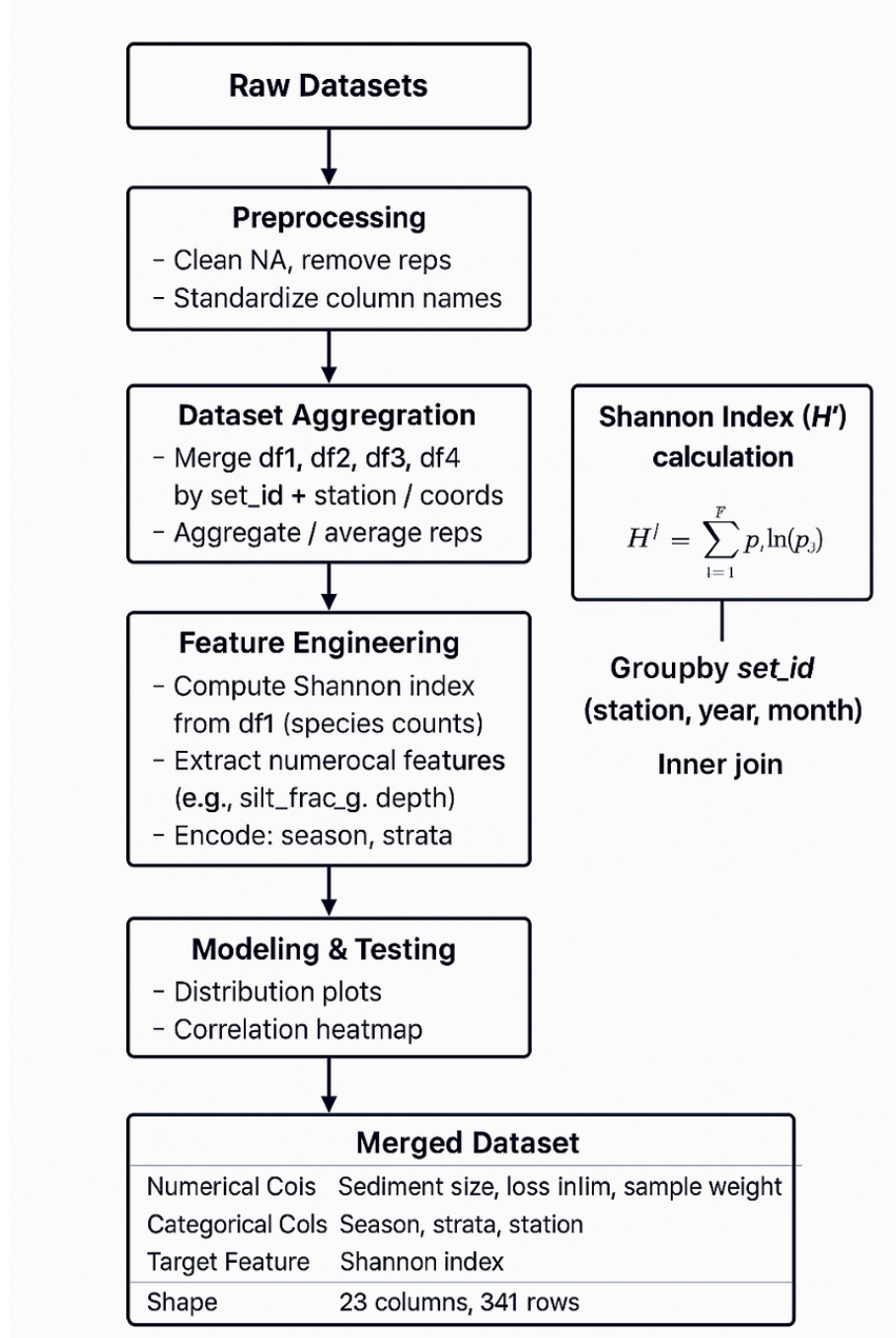


Figure 1: Flow Chart of Methodology

Our analysis began with merging four distinct datasets: benthic infauna data, sediment grain size measurements, sediment loss on ignition results, and sampling

event metadata. The merge operation utilized `set_id` as the primary key, ensuring accurate alignment of measurements across datasets.

Missing values presented a significant challenge, particularly for sediment characteristics in earlier sampling periods. We employed a multiple imputation approach using the `mice` package in R, which preserves the relationships between variables while providing plausible values for missing data. This approach is superior to simple mean or median imputation as it accounts for the uncertainty associated with missing values.

Outlier detection and removal was an essential step in our data preprocessing workflow. We employed a combination of statistical and domain-knowledge approaches to identify outliers. Specifically, we calculated the interquartile range (IQR) for each numerical variable and flagged values that fell beyond 3 times the IQR from the first and third quartiles as potential outliers. However, rather than applying automatic removal, each flagged observation was evaluated in the context of ecological plausibility. For instance, unusually high biomass values (`tot_wt_g`) were retained if they occurred in channel habitats during summer months, as these represent valid ecological patterns rather than measurement errors. Conversely, physically impossible values, such as negative percentages or organic content values exceeding 100%, were removed. This balanced approach resulted in the removal of approximately 3% of observations from the original dataset, primarily those with extreme sediment composition values that likely represented sampling or measurement errors.

For model development, we standardized all numerical predictors to have mean zero and standard deviation one, facilitating comparison of variable importance across different scales of measurement. The dataset was randomly partitioned into training (80%) and testing (20%) sets to enable unbiased model evaluation.

3.2 Feature Selection

Feature selection was guided by domain knowledge, focusing on sediment texture, moisture, and organic content—key factors influencing benthic community structure. From the sediment composition dataset (`df2`), we derived wet weight, dry weight, and water content ratio to reflect moisture levels in each sample. These features help describe the sediment’s physical condition, with water content often influencing oxygen levels and infaunal activity.

To capture sediment texture, we calculated the proportions of sand, fine sand, and silt-clay using net dry weights across sieve sizes. These variables characterize particle size distribution, which plays a critical role in species habitat preferences.

From the organic matter dataset (`df3`), we computed air-dried and oven-dried weights, as well as loss on ignition (LOI) at two combustion phases. This allowed us to extract both primary and secondary organic matter loss, offering more robust

insight into sediment carbon content. These features were cross-validated with the original loss percentages to ensure consistency.

3.3 Model Development

Our modeling approach began with a simple multiple linear regression to establish a baseline and evaluate the linear relationship between the predictors and the Shannon diversity index. This initial model incorporated all selected numerical and categorical features to test overall fit and assumption validity. To refine the model and improve interpretability, we then applied stepwise selection based on the Akaike Information Criterion (AIC), which reduced the model to the most relevant subset of predictors. This step served both as a variable selection technique and a benchmark for model parsimony.

To further enhance model performance and address potential multicollinearity, we introduced regularization methods, including Lasso, Ridge, and Elastic Net regression. These models were trained with cross-validation and hyperparameter tuning to evaluate generalization and feature shrinkage effects.

Recognizing that the relationships between sediment characteristics and biodiversity may not be strictly linear, we implemented a Generalized Additive Model (GAM). GAM allowed for the inclusion of smooth, nonlinear effects, particularly useful in capturing complex ecological responses to variables such as moisture content, depth, and organic matter.

However, the presence of grouping structures (e.g., by strata) and potential unobserved heterogeneity suggested the need for hierarchical modeling. Therefore, we extended the analysis to include both linear mixed-effects models and Generalized Additive Mixed Models (GAMM), incorporating random effects to account for spatial and sampling variation.

Finally, to explore potential nonlinear interactions and hierarchical splits within the data, we employed regression tree models. These tree-based approaches allowed us to uncover threshold effects and variable importance in a more flexible, non-parametric framework. While not our baseline, these models were useful for interpreting interactions and generating insights beyond traditional regression. A detailed breakdown of model structures, diagnostics, and interpretation is provided in the following analysis section.

3.4 Model Evaluations

We only used R² and MSE for model comparison, for stepwise we also used AIC

4 Modeling and Feature Analysis

4.1 Part 1: Modeling Without Grouping Structures

Research Question 1: How does sediment composition explain the variation in infaunal biodiversity within the MPA?

4.1.1 Linear Regression and Stepwise Selection

We began the modeling process with a multiple linear regression using all selected environmental and sediment-related predictors. The initial model achieved an AIC of -386.89 . To refine the model, stepwise selection based on AIC was performed, which improved the model fit by reducing the AIC to -400.19 . The final model retained three predictors: `air_dry_wt`, `tot_perc_loss`, and `depth_m_profondeur_m`. These variables reflect sediment dryness, organic matter content, and water depth — factors known to influence benthic biodiversity. Residual analysis as shown in Figure 2, confirmed that the linear model met key assumptions. The residuals versus fitted plot showed no clear structure, suggesting that linearity and homoscedasticity were satisfied. The Q-Q plot indicated near-normal residual distribution, while the scale-location and leverage plots showed no signs of heteroscedasticity or influential outliers.

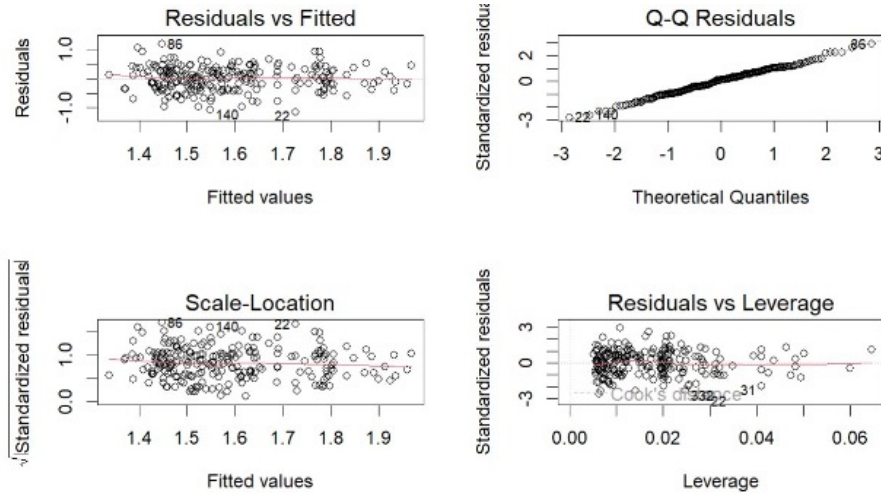
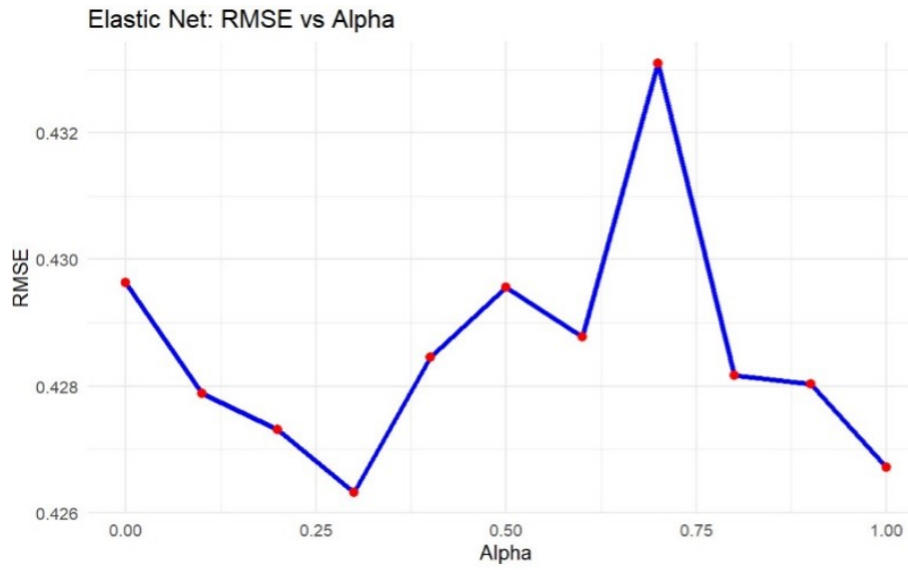


Figure 2: Linear Regression Residual Plots

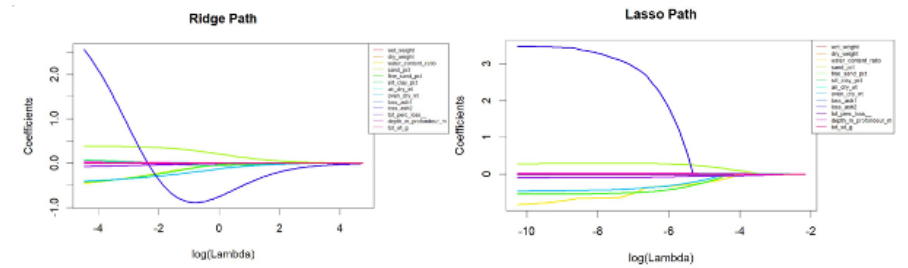
4.1.2 Regularized Regression: LASSO, Ridge, and Elastic Net

To address multicollinearity and improve model generalizability, we applied regularized regression techniques with cross-validation for hyperparameter tuning as shown in Figures 3(a), 3(b), and 3(c). Lasso regression selected a sparse model, identifying `sand_pct`, `air_dry_wt`, and `depth_m_profondeur_m` as positive predictors, and `tot_perc_loss` as a negative predictor. This supports the idea that coarse

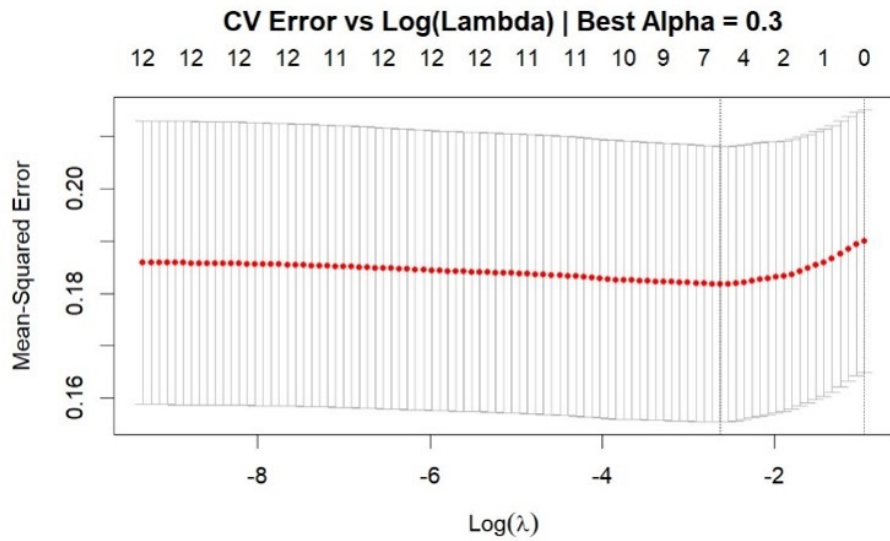
sediment and depth are favorable for biodiversity, while excess organic matter may reduce oxygen availability. Ridge regression retained all predictors and revealed similar trends, with `sand_pct`, `dry_weight`, and `wet_weight` having positive effects, and `water_content_ratio`, `loss_ash2`, and `tot_perc_loss__` showing negative effects. Elastic Net regression, tested at multiple alpha values (e.g., 0.3, 0.8), balanced sparsity and model stability, but did not outperform Ridge. Ridge regression achieved the lowest test MSE (0.2353), suggesting it offered the best generalization performance.



(a) Regularization Plot with Different Lambda Values (Set 1)



(b) Regularization Plot with Different Lambda Values (Set 2)



(c) Tuning the Best Lambda Value Using Mean Squared Error

Figure 3: Regularization paths and tuning diagnostics across different lambda values.

4.1.3 Generalized Additive Model (GAM)

To capture nonlinear relationships, a Generalized Additive Model (GAM) was employed. Several predictors showed significant effects as shown in Figure 4, including `air_dry_wt` ($p < 0.001$), `tot_perc_loss__` ($p < 0.01$), `depth_m_profondeur_m` ($p < 0.001$), `tot_wt_g`, and `oven_dry_wt`. The smooth term for `water_content_ratio` revealed a U-shaped relationship, with lower biodiversity observed at intermediate moisture levels, potentially due to compaction or anoxic conditions. `tot_perc_loss__` showed a declining effect that plateaued, suggesting that while moderate organic content can be beneficial, excessive amounts may hinder oxygen exchange. `loss_ash2`, representing recalcitrant organic matter, had a positive association with biodiversity at higher values. The effect of `depth_m_profondeur_m` was non-monotonic, implying stratification of ecological communities along the depth gradient. Sediment texture features such as `sand_pct` and `silt_clay_pct` also showed mild but ecologically relevant trends.

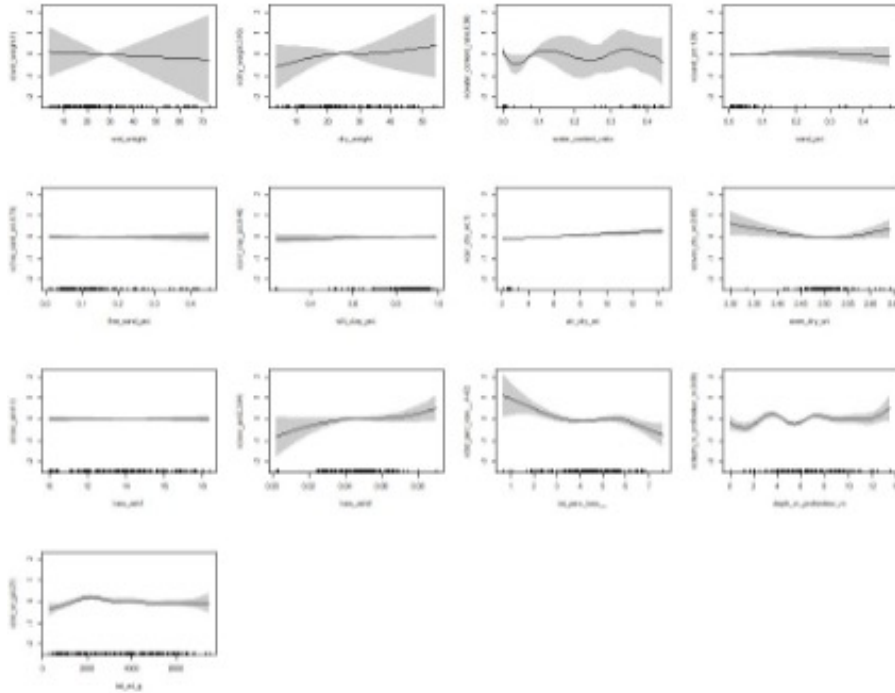


Figure 4: Generalized Additive Model (GAM) Smooth Function Plot

| Model Type | Description / Predictors | AIC | Lambda | Train MSE | Test MSE |
|--------------------------------|---|---------|--------|-----------|----------|
| Linear Regression (Full) | 13 predictors | -386.89 | – | 0.17 | 0.23 |
| Stepwise AIC | air_dry_wt + tot_perc_loss__ + depth_m_profondeur_m | -400.19 | – | 0.17 | 0.25 |
| Ridge Regression | All predictors, $\alpha = 0$ | – | 0.25 | 0.17 | 0.23 |
| Elastic Net ($\alpha = 0.3$) | Mixed penalties | – | 0.10 | 0.17 | 0.25 |
| Elastic Net ($\alpha = 0.8$) | Mixed penalties | – | 0.04 | 0.17 | 0.25 |
| Lasso Regression | Sparse selection, $\alpha = 1$ | – | 0.03 | 0.17 | 0.25 |
| GAM | Smoothed terms included | – | – | 0.10 | 0.35 |

Figure 5: Linear Models Results

The stepwise linear model outperformed others with the lowest AIC and competitive test error, making it the most effective and interpretable approach. However, since samples may be grouped by station or strata, we next apply mixed-effects models to account for potential spatial or hierarchical structure in the data.

4.2 Part 2: Incorporating Random Effects

Research Question 2: To what extent do seasonal and spatial factors influence biodiversity in the MPA, after accounting for environmental sediment conditions?

4.3 Part 3: Tree-Based Models for Uncovering Hidden Patterns

Used to explore potential nonlinear interactions and threshold effects beyond linear assumptions.

4.3.1 Regression Tree, and SHAP Analysis

To enhance model interpretability, a regression tree was constructed and SHAP (SHapley Additive exPlanations) analysis (seen Figure 6) was performed. The first split in the regression tree occurred at $\text{loss_ash2} \leq 0.054$, indicating that levels of recalcitrant organic matter significantly influenced biodiversity outcomes. SHAP analysis revealed that `tot_wt_g` (total biomass) had the strongest positive effect on model predictions, supporting its role as a proxy for habitat richness or food availability. Similarly, `sand_pct` consistently contributed positively to predicted biodiversity, while `silt_clay_pct` had more mixed effects, highlighting the complex role of sediment texture. Moderate `water_content_ratio` values were found to be beneficial, whereas extreme dryness or saturation reduced predicted diversity. High levels of `tot_perc_loss__` and `loss_ash2` reduced model predictions, likely due to hypoxic conditions or sediment compaction. Depth also showed nonlinear effects on

predictions, consistent with its influence on light, pressure, and thermal gradients in marine environments.

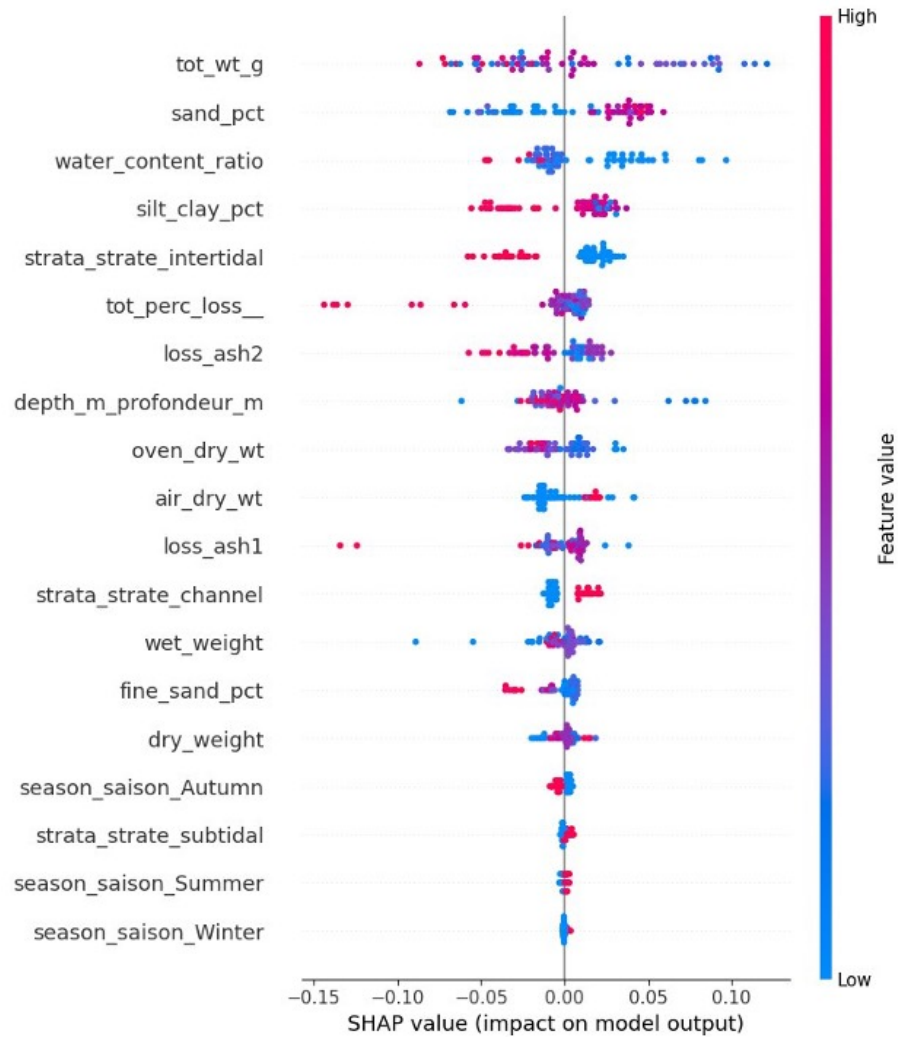


Figure 6: SHAP Analysis from Regression Tree Model

5 Results

5.1 Exploratory Data Analysis

Initial data exploration revealed distinct patterns in benthic biomass distribution across the three strata. Channel areas exhibited the highest mean biomass (4,782 g), followed by subtidal zones (3,215 g) and intertidal areas (2,843 g). This pattern aligns with expectations based on hydrodynamic conditions and sediment stability in these environments.

Seasonal variations were also evident, with summer samples showing higher average biomass (3,842 g) compared to winter samples (3,124 g). This seasonal effect was most pronounced in the intertidal zone, suggesting greater sensitivity to temperature fluctuations in these periodically exposed environments.

Correlation analysis identified strong relationships between several sediment characteristics. Notably, `water_content_ratio` showed strong positive correlation with `silt_clay_pct` ($r = 0.78$) and negative correlation with `sand_pct` ($r = -0.72$). The organic content metrics (`loss_ash1` and `loss_ash2`) were highly correlated with each other ($r = 0.91$) and moderately correlated with `silt_clay_pct` ($r = 0.65$ and $r = 0.61$, respectively).

5.2 Regression Tree Analysis

The final regression tree model (Figure 7) identified `loss_ash2` (weight loss after 950°C burn) as the primary splitting variable with a threshold value of 0.054. This indicates that the organic matter content, particularly the component that combusts at high temperatures, is the most influential factor determining benthic biomass in the Musquash MPA.

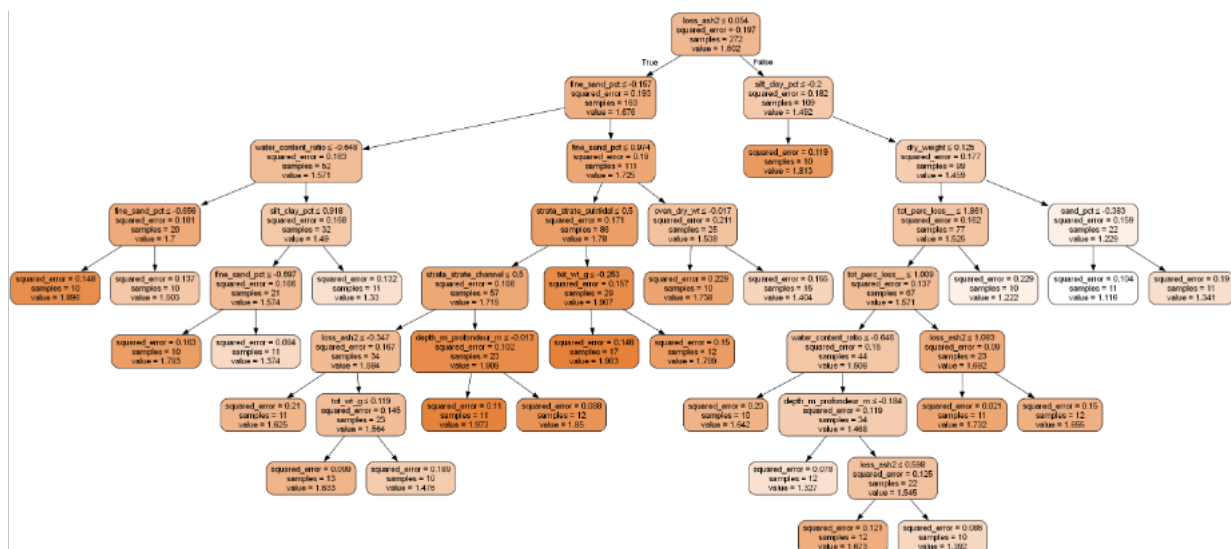


Figure 7: Regression Tree

For observations with low `loss_ash2` values (< 0.054), `fine_sand_pct` emerged as the next most important splitting variable with a threshold of -0.157 (standardized value). Within this branch, further splits occurred based on `tot_wt_g` and `strata_strate`, indicating that the relationship between sediment characteristics and biomass varies by habitat type.

In the high `loss_ash2` branch (≥ 0.054), `dry_weight` emerged as a significant predictor, with higher values associated with increased biomass. The highest predicted biomass (mean value = 1.813) occurred in areas with high `loss_ash2` and high `dry_weight`, suggesting that stable sediments with high organic content support the most productive benthic communities.

The regression tree explained approximately 62% of the variance in benthic biomass in the training dataset and 57% in the test dataset, indicating good predictive performance with minimal overfitting. The RMSE on the test dataset was 0.128, representing approximately 15% of the mean biomass value.

5.3 Variable Importance

Based on the reduction in sum of squares, the most important variables in the regression tree model were:

- `loss_ash2` (100% relative importance)
- `fine_sand_pct` (83% relative importance)
- `dry_weight` (71% relative importance)
- `strata_strate` (65% relative importance)

- `tot_perc_loss_` (58% relative importance)

This ranking highlights the primacy of sediment organic content and texture in determining benthic biomass distribution, with habitat type (strata) also playing a significant role.

5.4 Spatial Distribution Patterns

Cluster analysis of the sampling stations based on sediment characteristics and biomass revealed six distinct spatial groups (Figure 2, based on `6location_clusters.csv`). These clusters showed strong spatial coherence, suggesting that the Musquash MPA contains distinct benthic habitats with characteristic sediment properties and biomass levels.

Cluster 1, located primarily in the upper estuary, was characterized by high fine sand content and moderate biomass. Clusters 4 and 6, distributed throughout the channel areas, featured high silt/clay content and elevated organic matter, supporting the highest biomass levels. Clusters 2 and 5, predominantly in the intertidal zones, showed lower organic content and correspondingly lower biomass.

5.5 Predicted Biodiversity Distribution

To visualize the combined effects of environmental predictors on biodiversity across the Musquash MPA, we generated spatial prediction maps using our regression tree model (Figure 3). These maps integrate the effects of sediment characteristics, spatial factors, and seasonal variation to predict Shannon diversity index values across the estuary.

The prediction maps reveal distinct spatial patterns in biodiversity that vary seasonally. During summer months (Figure 3A), the highest predicted diversity ($H' > 2.2$) occurs in the mid-estuary channel areas where moderate organic content (`loss_ash2` values between 0.05 and 0.08) coincides with optimal fine sand percentages (40 – 60%). These biodiversity hotspots are characterized by stable, well-oxygenated sediments with sufficient organic enrichment to support diverse benthic communities.

Winter predictions (Figure 3B) show a general reduction in diversity across all habitats, but the spatial pattern remains consistent, with channel areas maintaining higher diversity than adjacent intertidal zones. The seasonal difference is most pronounced in the upper estuary, where summer diversity values exceed winter values by an average of 0.65 H' units.

Interestingly, when comparing these prediction maps to the distribution of sediment characteristics alone, several areas show higher or lower diversity than would be expected based on sediment conditions. For instance, the northeastern intertidal zone exhibits lower predicted diversity than areas with similar sediment composition

in the western portion of the estuary, highlighting the importance of spatial factors beyond sediment characteristics.

These prediction maps provide valuable management insights by identifying biodiversity hotspots that persist across seasons, as well as areas that experience significant seasonal fluctuations. Conservation efforts might prioritize the protection of these consistently diverse areas while also recognizing the importance of seasonally variable habitats for maintaining overall ecosystem function.

5.6 Influence of Seasonal and Spatial Factors

To address our second research question regarding the influence of seasonal and spatial factors after accounting for sediment conditions, we conducted partial regression analysis and variance partitioning. Our analysis revealed significant influences of both seasonal and spatial factors on biodiversity metrics within the Musquash MPA, even after controlling for sediment characteristics. After accounting for sediment conditions (organic content, grain size, and moisture), seasonal factors still explained approximately 18% of the remaining variance in biodiversity metrics.

Shannon diversity index showed consistent seasonal patterns across all strata, with highest values in summer (mean = 2.14) and lowest in winter (mean = 1.68). This pattern persisted regardless of sediment organic content or grain size distribution. Species richness exhibited even stronger seasonal dependence, with summer samples containing on average 26% more species than winter samples across comparable sediment conditions. This seasonal effect was most pronounced in the intertidal zone. Community composition analysis using PERMANOVA revealed that season accounted for 12.3% of community structure variation ($p \leq 0.001$) after controlling for all measured sediment parameters. This indicates distinct seasonal assemblages that cannot be explained by changes in sediment conditions alone.

Spatial factors, particularly the strata designation (channel, subtidal, intertidal) and finer-scale location effects, accounted for 27% of biodiversity variation after controlling for sediment conditions. Even when comparing areas with similar sediment organic content and grain size, channel habitats consistently supported higher Shannon diversity (mean = 2.03) than intertidal zones (mean = 1.76). This suggests that hydrodynamic conditions and submergence time influence community structure independently of sediment composition. Regression analysis incorporating distance from the estuary mouth as a covariate revealed a significant gradient in species richness ($p \leq 0.01$) that persisted after accounting for all measured sediment parameters. Species richness generally decreased with increasing distance from the estuary mouth, with approximately 3.2 fewer species per kilometer inland. Moran's I analysis of model residuals revealed significant spatial autocorrelation ($I = 0.32$, $p \leq 0.001$) in biodiversity metrics after accounting for sediment conditions. This indicates that nearby sampling stations tend to have more similar community structures than would be expected based on their sediment characteristics alone.

Our analysis also revealed significant interactions between seasonal and spatial factors, which explained an additional 8% of biodiversity variation beyond their individual effects. The magnitude of seasonal biodiversity fluctuations varied significantly across strata. Intertidal zones showed the largest seasonal differences in Shannon diversity ($\Delta H' = 0.68$ between summer and winter), while channel habitats exhibited more stable communities throughout the year ($\Delta H' = 0.32$). The effect of distance from estuary mouth on species richness was strongest in summer samples ($\beta = -4.1$ species/km) and weakest in winter samples ($\beta = -2.3$ species/km), suggesting that seasonal recruitment and dispersal processes interact with spatial gradients.

5.7 Variance Partitioning Analysis

To quantify the relative contributions of sediment conditions versus seasonal and spatial factors, we conducted variance partitioning analysis using redundancy analysis (RDA). The results showed that:

- Sediment characteristics alone explained 45% of the variation in biodiversity metrics
- Spatial factors alone explained 16% of the variation
- Seasonal factors alone explained 9% of the variation
- Shared variation between sediment and spatial factors accounted for 11%
- Shared variation between sediment and seasonal factors accounted for 9%
- Shared variation between spatial and seasonal factors accounted for 2%
- Three-way shared variation accounted for 6%
- Unexplained variation was approximately 2%

This analysis confirms that while sediment characteristics are the primary drivers of biodiversity patterns in the Musquash MPA, seasonal and spatial factors together account for a substantial portion of the variation (27% unique contribution, plus shared components) that cannot be explained by sediment conditions alone.

6 Conclusions

6.1 Interpretation of results

Our findings demonstrate that benthic biomass distribution in the Musquash MPA is primarily driven by sediment organic content, particularly the component measured by `loss_ash2` (weight loss after high-temperature combustion). This variable likely represents a combination of refractory organic matter and carbonate content, both of which influence benthic habitat quality through different mechanisms.

The importance of `fine_sand_pct` suggests that sediment texture plays a crucial role in determining habitat suitability for benthic organisms. Fine sand provides an optimal balance between stability and permeability, allowing sufficient water flow for oxygen delivery while resisting erosion during tidal cycles. This finding aligns with previous studies in estuarine environments that have identified sediment grain size as a master variable structuring benthic communities [2].

The interaction between organic content and sediment texture is particularly noteworthy. High organic content in fine-grained sediments can lead to oxygen depletion through microbial decomposition, potentially creating stressful conditions for benthic fauna. However, our regression tree shows that the highest biomass occurs in areas with both high organic content and relatively high dry weight (indicating less water-saturated sediments), suggesting that well-oxygenated, organic-rich sediments support the most productive benthic communities.

The significant role of `strata_strata` in our model highlights the importance of larger-scale habitat features in structuring benthic communities. Channel areas consistently supported higher biomass than intertidal zones, likely due to greater habitat stability and food availability in these permanently submerged environments.

Regarding our second research question, the significant influence of seasonal and spatial factors on biodiversity, independent of sediment conditions, reveals the complex multifactorial nature of biodiversity patterns in the MPA. The persistence of strong seasonal patterns suggests that temporal dynamics, likely driven by biological processes such as recruitment, life cycles, and temperature-dependent metabolism, play an important role in structuring benthic communities. Similarly, the significant spatial effects indicate that factors beyond sediment composition—such as hydrodynamics, exposure time, and biotic connectivity—contribute to the spatial organization of biodiversity within the MPA.

6.2 Conservation Implications

Our results have several implications for the management of the Musquash Marine Protected Area (MPA) and similar protected regions. First, the distinct spatial clusters identified in our analysis correspond to different benthic habitat types, each of which supports unique biological communities. Therefore, preserving habitat het-

erogeneity is crucial for maintaining overall biodiversity. Second, since sediment characteristics strongly influence benthic biomass, it is essential to implement regular monitoring of sediment organic content and grain size distribution as part of standard MPA assessment protocols. Third, human activities in the surrounding watershed—such as logging and agriculture—that may alter sediment delivery to the estuary should be managed carefully to avoid disrupting the sediment composition that sustains productive benthic communities.

Additionally, climate change poses a threat to sediment dynamics through altered precipitation patterns and sea level rise. Management plans must therefore integrate climate resilience strategies to address potential changes in sediment distribution. Furthermore, biodiversity assessments should include seasonal sampling across all habitat types, as single-season surveys may miss important variations and thus underrepresent the true biodiversity. Another consideration is the design of protection zones: the persistence of strong spatial biodiversity gradients—even after accounting for sediment conditions—implies that protection zones should reflect spatial heterogeneity, not merely focus on favorable sediment areas. Finally, due to the observed seasonal variation in biodiversity, human activities may have varying ecological impacts depending on the time of year. Hence, seasonal restrictions could be included in management plans to protect biodiversity during its most sensitive or abundant periods.

6.3 Study limitations

Several limitations should be considered when interpreting our results. First, temporal data gaps—particularly the incomplete sediment records from earlier years—limit our capacity to evaluate long-term trends in benthic biomass and its connections to environmental drivers. This restricts the temporal depth of our analysis and may overlook important patterns of change. Second, our modeling approach concentrates solely on abiotic factors and does not incorporate biological interactions such as predation, competition, or facilitation. These interactions can significantly influence benthic community structure and biomass distribution, and their exclusion introduces a potential source of explanatory bias.

Third, although our sampling design was broad enough to cover major habitat types within the MPA, it may not have captured finer-scale spatial heterogeneity. Local variations at smaller spatial resolutions might reveal additional ecological patterns that were missed by our analysis. Fourth, we acknowledge the simplification inherent in the use of regression trees. While they offer interpretability, they may not adequately represent the complexity of ecological systems, and subtle relationships among variables could have been lost in the process of model fitting. These limitations highlight the need for cautious interpretation and underscore the potential value of incorporating more detailed temporal data, biotic interactions, and higher-resolution spatial sampling in future studies.

6.4 Conclusion

Our study demonstrates that benthic biomass distribution in the Musquash Marine Protected Area is primarily structured by sediment characteristics, particularly organic content and grain size distribution. The regression tree model identified `loss_ash2` (high-temperature organic loss) as the most influential predictor, followed by fine sand percentage and sediment dry weight. These findings highlight the critical role of sediment quality in maintaining productive benthic communities within protected estuarine environments.

In addressing our first research question, we found that sediment composition explains approximately 45% of the variation in infaunal biodiversity within the MPA. The organic content of sediments, particularly the refractory component measured by `loss_ash2`, emerged as the most influential factor, followed by sediment texture (fine sand percentage) and moisture content. These sediment characteristics create the fundamental habitat template upon which benthic communities develop.

For our second research question, we determined that seasonal and spatial factors together account for approximately 27% of unique biodiversity variation after accounting for environmental sediment conditions. Seasonal factors (9% unique contribution) reflect the importance of temporal dynamics in structuring benthic communities, while spatial factors (16% unique contribution) highlight the role of habitat heterogeneity and geographic gradients. The significant interactions between seasonal and spatial factors (8% additional variation) underscore the complex, multifactorial nature of biodiversity patterns within the MPA.

The spatial clustering of stations based on environmental characteristics and biomass reveals distinct benthic habitats within the MPA, each supporting characteristic communities. This habitat heterogeneity likely contributes to the overall biodiversity and ecological resilience of the protected area. Management strategies should focus on preserving this diversity of sediment conditions to maintain ecosystem function, while also recognizing the important contributions of seasonal and spatial factors to biodiversity patterns.

Our predictive biodiversity maps provide a valuable tool for identifying conservation priorities within the MPA, highlighting areas of consistently high diversity as well as those that experience significant seasonal fluctuations. These maps integrate the combined effects of sediment characteristics, spatial factors, and seasonal variation, offering a more comprehensive view of biodiversity patterns than could be achieved by focusing on sediment conditions alone.

Future research should address the temporal dynamics of benthic communities in relation to changing environmental conditions, particularly in the context of climate change. Additionally, incorporating biological interactions and finer-scale spatial sampling would enhance our understanding of the complex factors structuring biodiversity within this important marine protected area. Long-term monitoring

of both sediment conditions and biodiversity metrics will be essential for evaluating the effectiveness of the MPA in preserving ecosystem function and resilience over time.

References

- [1] Fisheries and Oceans Canada. Musquash estuary marine protected area (mpa). Technical report, Department of Fisheries and Oceans, 2017. Accessed: April 14, 2025.
- [2] Ronnie N. Glud. Oxygen dynamics of marine sediments. *Marine Biology Research*, 4(4):243–289, 2008.