Quantitatively characterizing benthic community-habitat relationships in soft-sediment, nearshore environments to yield useful results for management

Flanagan A.M. [a*], Flood R.D. [a], Maher N.P. [b], and Cerrato R.M. [a]

[a] School of Marine and Atmospheric Sciences, Stony Brook University, Stony Brook, NY 11794-5000, USA

[b] The Nature Conservancy, Long Island Chapter, Cold Spring Harbor, NY 11724, USA

* Corresponding author. *E-mail address:* alisonmflanagan@gmail.com

Abstract

Effective management of benthic habitats is important for maintaining heathy and functional aquatic ecosystems. To provide managers with the best possible information, characterizing benthic habitats at the community level is essential; yet, acquiring the data sets needed to achieve this task is resource intensive and, at times, prohibitively expensive. Thus, thoughtful assessments of which data to collect and utilize in benthic habitat characterization studies are needed. Environmental data sets commonly used to characterize benthic habitats include a range of variables from water depth and sediment grain size to seabed features identified by sonar backscatter. The objective of this study was to identify the most useful environmental variables for characterizing infaunal benthic habitats and to determine how to best utilize these variables in analyses (e.g., by comparing continuous vs. categorical explanatory variables). The modeling approach used multivariate regression tree and redundancy analysis along with a critical cross-validation step for model evaluation. Results indicated that models with more than ~ 7 environmental predictors overfitted the data sets analyzed and that categorizing continuous

predictors into categorical ones influenced the proportion of infaunal community variation explained by each model. Habitats identified and characterized on the basis of sonar backscatter explained more of the infaunal community variation than any model that used a combination of other environmental variables (e.g., water depth & sediment grain size) or those constructed using categorical habitat classes from existing classification schemes. We therefore recommend maximizing the potential of sonar-derived variables for characterizing infaunal benthic habitats in nearshore, soft-sediment ecosystems.

Abstract

Effective management of benthic habitats is important for maintaining heathy and functional aquatic ecosystems. To provide managers with the best possible information, characterizing benthic habitats at the community level is essential; yet, acquiring the data sets needed to achieve this task is resource intensive and, at times, prohibitively expensive. Thus, thoughtful assessments of which data to collect and utilize in benthic habitat characterization studies are needed. Environmental data sets commonly used to characterize benthic habitats include a range of variables from water depth and sediment grain size to seabed features identified by sonar backscatter. The objective of this study was to identify the most useful environmental variables for characterizing infaunal benthic habitats and to determine how to best utilize these variables in analyses (e.g., by comparing continuous vs. categorical explanatory variables). The modeling approach used multivariate regression tree and redundancy analysis along with a critical cross-validation step for model evaluation. Results indicated that models with more than ~ 7 environmental predictors overfitted the data sets analyzed and that categorizing continuous predictors into categorical ones influenced the proportion of infaunal community variation explained by each model. Habitats identified and characterized on the basis of sonar backscatter explained more of the infaunal community variation than any model that used a combination of other environmental variables (e.g., water depth & sediment grain size) or those constructed using categorical habitat classes from existing classification schemes. We therefore recommend maximizing the potential of sonar-derived variables for characterizing infaunal benthic habitats in nearshore, soft-sediment ecosystems.

1. Introduction

The accurate characterization of coastal benthic habitats is a significant element of effective management because of the critical processes these habitats and associated communities contribute to ecosystems. Although numerous definitions of benthic habitats exist in the literature, they generally can be defined as areas of the seabed with distinct physical, chemical, and biotic characteristics (Lecours et al., 2015). The benthic communities within habitats have several important ecological functions including nutrient cycling (Welsh, 2003), providing structure and habitat for other organisms (Gray, 1974), planktonic food web interactions (Cloern, 1982; Cerrato et al., 2004), and as prey for higher trophic foragers including waterbirds (Richman and Lovvorn, 2009; Pérez-Vargas et al., 2016; Maceda-Veiga et al., 2017) and demersal fishes (Bottom and Jones, 1990; McCormick, 1995; Bizzarro et al., 2017). Since most benthic organisms are largely sedentary and cannot easily relocate to more suitable habitats, benthic communities are also particularly susceptible to natural and anthropogenic disturbances (Pearson and Rosenberg, 1978; Kröncke and Reiss, 2010) making them excellent bioindicators of changes in habitat quality (Borja et al., 2000; Borja et al., 2015; Pelletier et al., 2018).

Data sets that are commonly used to characterize soft-sediment (gravel, sand, and silt-clay) benthic habitats include biotic assemblage data in the form of counts of individual species, measurements of abiotic variables such as water depth (e.g., Smale, 2008; Marshall et al., 2018), sediment grain size (e.g., Sanders, 1958; Flanagan & Cerrato, 2015), sediment organic content (e.g., Silva et al., 2006; Ferraro & Cole, 2012), surficial percent cover of abiotic and biogenic materials (Taylor, 1998; Flanagan et al., 2018), and larger-scale geomorphological features detected by sidescan, single-beam, or multibeam sonar (Bell et al., 2000; Diaz et al., 2004; Weaver et al., 2013; Lecours et al., 2016). One way to utilize these data is to categorize the biotic and abiotic variables using criteria from benthic habitat classification schemes, which are designed to provide a common language for describing and managing submerged habitats (e.g., Davies et al., 2004; Auster et al., 2009; Marshall et al., 2018). Whether the goal is to implement a habitat classification scheme or to quantitatively model variation in community-habitat relationships, a trade-off exists between the quality of the data collected (level of accuracy and precision), sample size, and spatial scale (resolution and extent) of the data sets used (Lecours et al., 2015; Flanagan et al., 2018). Yet, it is unclear how each of these characteristics should be prioritized in a study design (Lecours et al., 2015) or in an analysis.

Habitat managers are tasked with making decisions based on the data available to them at the time, which have probably been analyzed to some extent, potentially resulting in important habitat information being lost from the original, “raw” data. Using an unsuitable approach to an analysis could result in a flawed interpretation regardless of the quality and completeness of the underlying data. To provide managers with meaningful information that can be used to protect, monitor, and/or restore benthic habitats and the ecological functions they provide, it is essential to characterize habitats at the community level (Allee et al., 2000; Parks, 2002; Palumbi et al., 2003; Maher, 2006). Questions related to communities or "sets of co-occurring species" are some of the most difficult to address (Sutherland et al., 2013) especially in aquatic environments where community-habitat relationships are less easily observed, data are expensive to acquire (Deborde et al., 2016; Marshall et al., 2018) and studies are more limited relative to terrestrial ecosystems (Diaz et al., 2004; Lecours et al., 2015). In addition, subtidal, infaunal benthic communities in substrates dominated by gravel, sand, and silt-clay, the focus of the current study, are relatively less studied than intertidal communities (Fraschetti et al., 2005), epifaunal assemblages (Lecours et al., 2015), and subtidal, hard substrate reef communities (Fraschetti et al., 2005; Marshall et al., 2018). Thus, careful consideration of the limitations and the best utilization of benthic data sets for characterizing infaunal community-habitat relationships is a worthwhile undertaking.

Numerous researchers have suggested that the environmental variables used to characterize benthic habitats should serve as proxies for discerning patterns in benthic community assemblages (Stevens & Connolly, 2004; Auster et al., 2009; McGonigle et al., 2009). It follows from this rationale that the environmental data or any categorical variables derived from them (including habitat classes) should quantitatively predict variation in community structure – a notion that can be tested explicitly and rigorously in a statistical modeling framework (Flanagan & Cerrato, 2015). Of the commonly collected environmental data used to characterize soft-sediment benthic habitats, several questions arise regarding the number of environmental variables to include and the identity of the essential ones (Lecours et al., 2015). Many of the variables are correlated (e.g., sediment organic content and silt-clay), suggesting some level of redundant information will be present and that a subset of the variables may be utilized with little to no loss of explanatory power. Some variables that can be measured may or may not be useful proxies and may be relegated to secondary status under most circumstances. For example, some surficial cover characteristics (e.g., percent cover of sand, seaweed, shell, etc.) may be poor predictors of infaunal abundance. Inclusion of too many variables in a statistical model can result in overfitting, leading to deceptively high r^2 estimates and the inclusion of unnecessary, spurious variables (Burnham & Anderson, 2002).

Despite long-term studies of coastal benthic fauna (Petersen, 1913; Sanders, 1958; Flanagan & Cerrato, 2015), it also remains uncertain whether habitats are generally better described as varying along continuous gradients in underlying environmental factors or have discrete boundaries with corresponding abrupt changes in community structure (Smale, 2008; Flanagan et al., 2018). Thus, in addition to the type of explanatory variables utilized, there are numerous options for how they are incorporated into a modeling framework to evaluate the underlying structure of the benthic community-habitat relationships in question. Water depth, for instance, can be used as a continuous explanatory variable in an analysis or be included as a categorical one (e.g., shallow, moderate, deep; Auster et al., 2009; FGDC, 2012). The Folk (1954) sediment classification scheme is also often used to categorize sediment grain size data (percent gravel, sand, silt, and clay) into a fixed set of classes, and it has been adopted by some habitat classification schemes for characterizing sediment composition across different habitats (e.g., Auster et al., 2009; FGDC, 2012). However, with water depth, sediment grain size, and other commonly collected environmental variables, it is unclear whether the categories can be identified *a priori* without compromising model performance or if one variable form (i.e., continuous vs. categorical) is superior to another.

Here, we evaluated the extent to which environmental variables (water depth, sediment grain size, percent cover of cobble, shell, seaweeds, or other material, and regions with similar acoustic properties identified by sonar backscatter) that are typically measured in benthic habitat characterization studies explain infaunal community-habitat relationships in soft-sediment, nearshore environments ranging from brackish to near marine. The overarching objective of this study was to identify which types of environmental variables commonly collected in benthic habitat studies best explain relationships between infaunal communities and their environment and the most useful form these environmental data should take in statistical analyses. Our intent is to inform both managers and practitioners so that these variables may be prioritized in study designs relevant to benthic habitat characterization. To this end, we address questions of 1) variable form (e.g., continuous vs. categorical variables), 2) variable complexity (the number and types of explanatory variables utilized), 3) uniformity (consistent selection of explanatory variables across data sets), and 4) priority (critical to measure).

2. Methods

2.1. Study areas

This study was carried out using benthic data sets collected from five areas surrounding Long Island, New York: two within the Hudson River Estuary (Haverstraw Bay and Tappan Zee), one on Long Island's north shore (Huntington Harbor), and two areas within Long Island's Peconic Bays Estuary (Robins Island and Shelter Island; Figure 1). These areas are moderate in spatial extent ranging from 3.3 to 30.0 km^2 and were selected to represent a variety of habitats from brackish to near marine. The Hudson River areas comprise the mesohaline portion of the estuary while the others are polyhaline. Detailed descriptions of the study areas and the data sets analyzed are available from the studies cited in Table 1.

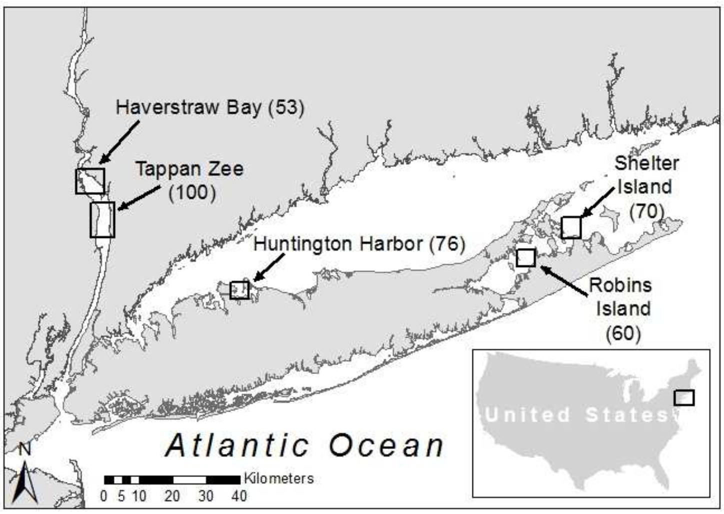


Figure 1. Study area map. The number of *in situ* environmental and faunal samples collected from each area is indicated in parentheses.

Table 1. The area of each study location and the sources of the data sets analyzed with references.

Study Location	Area (km^2)	Benthic assemblage, water depth, & sediment grain size data	Surficial percent cover from underwater video	Provinces	CMECS Geoforms
Haverstraw Bay	30.0	Cerrato et al. (2015)	Not collected	Bell et al. (2000)	Bell et al. (2000)
Tappan Zee	9.4	Maher & Cerrato (2004)	Flanagan (2016)	Bell et al. (2000)	Bell et al. (2000)
Huntington Harbor	16.5	Cerrato & Holt (2008)	Flanagan (2016)	Cerrato & Holt (2008)	Weaver et al. (2013)
Robins Island	3.3	Cerrato & Maher (2007)	Flanagan (2016)	Cerrato & Maher (2007)	Weaver et al. (2013)
Shelter Island	12.5	Cerrato & Maher (2007)	Flanagan (2016)	Cerrato & Maher (2007)	Weaver et al. (2013)

2.2. Summary of the data sets analyzed

The benthic data sets consisted of infaunal assemblage abundance data and commonly collected environmental variables including water depth, sediment grain size (percent gravel, sand, and silt-clay), surficial percent cover (e.g., percent cover of sand, seaweed, shell, etc.), and non-sonar and sonar-derived areas of the seabed with presumably homogeneous bottom types (hereafter called geoforms and provinces depending on the data source; Table 1). In addition, we created categorical environmental variables and habitat classes from the data collected at each area using criteria from two habitat classification schemes relevant to New York waters: a habitat classification scheme for the Long Island Sound (LIS) region (Auster et al., 2009), hereafter referred to as the LIS scheme, and the Coastal and Marine Ecological Classification Standard, hereafter designated as CMECS (FGDC, 2012).

2.3. Environmental and benthic assemblage data

Faunal and sediment samples were collected using a modified van Veen grab sampler (0.04 m^2) at the Haverstraw Bay (n = 51), Tappan Zee (n = 100), Huntington Harbor (n = 76), Robins Island (n = 60), and Shelter Island (n = 70) sites. Sampling locations were random but stratified by province to ensure full coverage of the range of infaunal habitats likely to be present. Water depth was recorded at the time of faunal sample collection. Subsamples of sediments for grain size analysis were drawn from each grab, and remaining material was washed through a 0.5 mm sieve for fauna. In the lab, sediment samples were partitioned into major size fractions (gravel, sand, & silt-clay) following Folk (1974). Individual organisms were identified to species whenever possible, and species abundance per grab sample (0.04 m^2) was enumerated.

Surficial percent cover of seabed materials (e.g., percent cover of sand, seaweed, shell, etc.) was obtained through supervised maximum likelihood analysis (ArcGIS 10.1 ESRI, Redlands, CA) of still images extracted from underwater videos, which were collected at four of the five study areas as in Flanagan (2016): Tappan Zee (n = 100), Huntington Harbor (n = 76), Robins Island (n = 60), and Shelter Island (n = 70). Image dimensions were 17.5 x 30 cm and comparable to those of the modified van Veen grab sampler (20 x 30 cm) used to collect the sediment and faunal samples. At Tappan Zee, recordings captured smaller areas (13.5 x 23.5 cm) because the camera had to be lowered closer to the seabed due to high turbidity (Flanagan, 2016).

Provinces were created through visual analysis of the backscatter data from Haverstraw Bay, Huntington Harbor, Robins Island, and Shelter Island (Maher, 2006; Cerrato & Maher, 2007; Cerrato & Holt, 2008; Cerrato et al., 2015). Provinces for the Tappan Zee area were taken from Bell et al. (2000) who supplemented sidescan sonar data with multibeam bathymetry, chirp sub-bottom seismics, sediment cores, and sediment grabs. Maps illustrating the configuration of the provinces identified at each area are provided in Supplementary Material.

2.4. Habitat classes

The LIS scheme and CMECS habitat classes were assigned as fixed classes to all sampling locations using the water depth, sediment grain size, surficial percent cover, and geoform data collected from each area. A detailed description of the methods used to assign the habitat classes is provided in Supplementary Material. Water depth measurements from each station were categorized as shallow ($\leq$ 4 m) or deep (> 4 m) explanatory variables. The sediment data were categorized using the Folk (1954) sediment classification system, and categorical biogenic components of the LIS scheme and CMECS were assigned using the surficial percent cover data. Geoforms, or structural regions of the seabed in the CMECS scheme, were taken from Weaver et al. (2013) for the Huntington Harbor, Robins Island, and Shelter Island areas. For the Tappan Zee area, bottom types described in Bell et al. (2000) were matched with the “Level 1” geoforms described in CMECS (FGDC, 2012). These included mollusc reef, flat, channel, and wave field. Four “Level 1” geoforms were identified for Haverstraw (mollusc reef, flat, channel, and dredged channel) from bathymetry, backscatter, and grain size data using CMECS criteria (FGDC, 2012).

2.5. Multivariate analyses of the benthic community-habitat relationships

Direct analyses (Legendre & Legendre, 1998) using a combination of multivariate regression tree (MRT; De’ath, 2002) and redundancy analysis (RDA; Jongman et al., 1995), along with a critical cross-validation step for model evaluation, were used to develop models of benthic community-habitat relationships. MRT and RDA are robust to collinearity in the explanatory variables since both are forward selection stepwise procedures that remove explained variance at each step before considering subsequent explanatory variables (Jongman et al, 1995). Faunal data were Hellinger transformed prior to analysis by calculating the square root of the relative abundance of each taxon within a sample. This transformation produces ecologically reasonable measures of compositional differences when coupled with Euclidean distance (Legendre & Gallagher, 2001), the metric utilized by both MRT and RDA; thus, this common metric allowed for direct comparisons of fit across statistical models, a feature that would not be possible if MRT were combined with other direct ordination methods (e.g., canonical correspondence analysis with its chi-square metric).

Since one of the main goals of this study was to identify which environmental variables best characterized the benthic habitats examined and whether variable form (e.g., continuous vs. categorical explanatory variables) influenced model performance, we developed models consisting of single types of environmental variables in addition to models with multiple types of explanatory variables (Tables 2-4). Wherever possible, we compared the continuous version of each explanatory variable (e.g., water depth in meters) to categorical analogs (e.g., the shallow/deep categories from the LIS scheme). Categorical explanatory variables were created in three ways (Tables 2 & 4): 1) defined *a priori*, e.g., using criteria from the LIS and CMECS habitat classification schemes (hereafter referred to as fixed categorical variables), 2) created by combining the fixed categorical variables into larger sets based on MRT results (hereafter referred to as simplified categorical variables; Figure 2), and 3) identified from the results of MRT on the continuous explanatory variables (e.g., sediment grain size), breaking the continuous variable up into intervals (hereafter referred to as flexible categorical variables; Tables 2 & 4; Figure 2). In addition, we included RDA models that can use continuous explanatory variables to examine potential linear relationships between benthic community structure and gradients in water depth, sediment composition, or surficial percent cover (Table 2). Models containing multiple explanatory variables utilized the LIS and CMECS habitat classes as fixed (or simplified) categorical explanatory variables, a combination of flexible explanatory variables and simplified provinces or geoforms, and a combination of continuous, linear explanatory variables with simplified provinces or geoforms (Table 3).

Table 2. Framework for assessing the extent to which single types of commonly collected environmental variables explain benthic community-habitat relationships when they are used as fixed categorical, simplified categorical, flexible categorical, or continuous explanatory variables in an analysis. Figure 2 provides an example that illustrates the MRT-based approach used to create the fixed, simplified, and flexible categorical variables and Table 4 describes the different variable types & forms utilized.

Variable type & form	Fixed categorical (all possible categories) →	Simplified categorical (determined by MRT on fixed categorical variables)	Flexible categorical variables (determined by MRT on continuous variables)	Continuous linear (RDA on continuous variables)
Water depth	Shallow/deep (± 4 m) from FGDC (2012)	NA	Flexible depth intervals	Continuous depth gradient
Sediment grain size	Fixed sediment categories from Folk (1954)	Simplified sediment categories from Folk (1954)	Flexible sediment grain size intervals (%gravel, sand & silt-clay)	Continuous sediment gradient (%gravel, sand & silt-clay)
Percent cover	NA	NA	Flexible percent cover intervals	Continuous percent cover (% cover of sand, fauna, shell, etc.)
Geoforms	Bell et al. (2000); FGDC (2012); Weaver et al. (2013)	Simplified geoforms	NA	NA
Provinces	Bell et al. (2000); Maher & Cerrato (2004); Maher (2006); Cerrato & Maher (2007); Cerrato & Holt (2008); Cerrato et al (2015)	Simplified provinces	NA	NA

Table 3. Framework for assessing the extent to which models with multiple types of categorical and/or continuous explanatory variables explain benthic community-habitat relationships. Habitat classes were evaluated as part of the multiple variable models since they integrate information from all of the other environmental variables collected in this study.

Explanatory variables from Table 1 for models with multiple explanatory variables	Fixed categorical	MRT: Flexible environmental categories & simplified geoforms or provinces	RDA: Continuous (linear) environmental & simplified categorical geoforms or provinces
Habitat classes	FGDC (2012)	Simplified habitat classes	NA
Environmental variables & geoforms	NA	Flexible depth, flexible sediment grain size, flexible percent cover & simplified geoforms	Continuous depth, continuous sediment grain size, continuous percent cover & simplified geoforms
Environmental variables & provinces	NA	Flexible depth, flexible grain size, flexible percent cover & simplified provinces	Continuous depth, continuous grain size, continuous percent cover & simplified provinces

MRT was used in the present study to create groups of faunal assemblage sample data based on repetitive, binary splitting of the categorical (e.g., the fixed habitat classes and provinces) and/or continuous (e.g., water depth, grain size, and surficial percent cover) explanatory variables. Binary splits were selected based on one of the explanatory variables to minimize differences in community structure within sample groups while maximizing differences between groups (De'ath, 2002). When MRT was run using fixed categorical variables, the criterion split the data into subsets containing samples from one or more categories (simplified categories). For continuous variables (water depth, sediment grain size, and percent cover), the criterion partitioned the range of the variable into intervals (flexible categories) and membership consisted of all samples in the interval. For all models, the binary splitting process was repeated until a stopping rule was met. For deriving simplified and flexible categorical variables, the stopping criterion was based on 10-fold cross-validation. Figure 2 provides an example illustrating this process. The final groups, called terminal nodes, are represented by the multivariate mean of all taxa belonging to that group. MRT models were run using the *rpart* function from the *mvpart* package in R (De'ath, 2014; R Core Development Team, 2017).

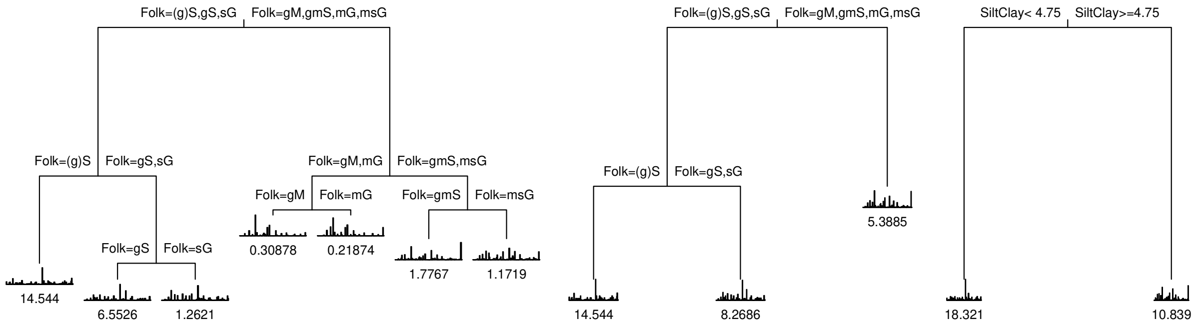


Figure 2. An example of models with fixed categorical variables (left), simplified categorical variables (middle), and flexible categorical variables (right) created using MRT. Response variables are Hellinger transformed species abundances and explanatory variables are either the Folk (1954) grain size categories or percent gravel, sand and silt-clay. This example is from Shelter Island and utilized the fixed grain size categories from the Folk (1954) sediment classification system to create simplified categories by combining the fixed categories into larger groups (e.g., gM, gmS, mG, and msG were grouped). For the flexible categories, percent gravel, sand, and silt-clay were the explanatory variables and a single split (based on percent silt-clay) was selected in the final MRT model. The abbreviations refer to descriptive names from Folk (1954): (g)S = slightly gravelly sand, gS = gravelly sand, sG = sandy gravel, gM = gravelly mud, gmS = gravelly muddy sand, mG = muddy gravel, and msG = muddy sandy gravel. The histograms at the bottom of each node are the average Hellinger transformed abundances of each taxon. The deviance (i.e. the sum of squared differences between Hellinger transformed species abundances) is indicated for each group.

RDA is a multivariate technique that combines ordination of sample species abundance data with multiple linear regression on the explanatory variables (Jongman et al., 1995). RDA was implemented in Canoco 4.5 (Microcomputer Power, Ithaca, NY, USA), and it can utilize both continuous and categorical explanatory variables. Categorical variables were incorporated into the analysis by representing the n categorical levels with (n - 1) binary (0, 1) variables. A script for 10-fold cross-validation of RDA results was created using the functions *rda* and *predict* in the *vegan* package (Oksanen et al., 2017; R Core Development Team, 2017).

RDA is limited to linear relationships between the response (benthic assemblage data) and explanatory variables (environmental data), but MRT can effectively explain a variety of community-habitat relationships including those that are nonlinear or that contain discontinuities and interactions (Crawley, 2007; Hastie et al., 2001). Thus, utilizing MRT and RDA together allowed the exploration of a wide range of community-habitat relationships (Flanagan & Cerrato, 2015). By comparing fit between the MRT and RDA models that were run using the same set of explanatory variables, it also becomes possible to infer whether community-habitat relationships are better described as varying along continuous (linear) gradients in underlying environmental factors (i.e., where RDA outperformed MRT) or by discrete boundaries with corresponding changes in community structure (i.e., where MRT outperformed RDA; Flanagan & Cerrato, 2015).

Cross-validation was a critical element in this study to protect against generating overfitted models and to provide a rational means of comparing models with large differences in the number of explanatory variables. Details regarding our approach to cross-validation in MRT and RDA, including an explanation of why cross-validation was favored over AIC, are provided in Supplementary Material. Table 4 provides a guide to the terminology used to describe the various data types, variable forms, and modeling definitions used in this study, highlighting those that are the most essential to understanding our approach and for interpreting the results.

Table 4. Definitions of the terms used to describe the variables and analysis methods utilized in this study.

Term	Definition in the context of this study
Variable type	Refers to the type of response (faunal data) and explanatory variables (environmental data) collected and used in the analyses.
Fauna	The abundances of each taxa within a van Veen grab sample.
Water depth	Water depth of each sampling location in meters.
Sediment grain size	The percentage of gravel, sand, and silt-clay in sediment samples.
Surficial percent cover	Percent surficial cover of abiotic and biotic seabed features (e.g., sand, seaweeds).
Provinces	Areas of the seabed that consist of relatively homogeneous bottom types (e.g., sandy vs. muddy areas) identified from visual analysis of sonar data (e.g., backscatter).
Geoforms	Presumably uniform regions of the seabed that are defined using the criteria from CMECS. Unlike provinces, geoforms can be derived with or without the use of sonar data.
Habitat classes	Habitat names derived using the criteria from the LIS scheme and CMECS. Used as fixed and simplified categorical (nominal) variables in the analyses.
Variable form	Refers to the structure of the explanatory variables used in the analyses (categorical vs. continuous).
Fixed categorical variables	Categorical explanatory variables that were defined prior to analyses. In most cases the fixed categorical variables in this study are from the LIS and CMECS habitat classification schemes.
Simplified categorical variables	Categorical explanatory variables that were created by combining the fixed categorical variables into larger sets using MRT analysis.
Flexible categorical variables	Categorical explanatory variables that were defined using MRT analysis on continuous variables. In other words, flexible categories were defined by analysis rather than being defined *a priori* using fixed criteria.
Continuous variables	Non-categorical (numeric) explanatory variables. Examples include water depth in meters and the percentage of gravel, sand, and silt-clay in a sample.
Benthic community-habitat relationships	The relationship between benthic communities and the biotic and abiotic components of their environment.
Single variable models	MRT and RDA models with single explanatory variables (e.g., models with water depth as the only explanatory variable).
Multiple variable models	MRT and RDA models with multiple types and combinations of explanatory variables (e.g., models with water depth, sediment grain size, percent cover, and provinces as explanatory variables).
Explained variance	The proportion of benthic community variation explained by the explanatory variables used in MRT or RDA.
r^2	The coefficient of determination. Used in this study to illustrate differences between r^2 and cross-validated r^2. This comparison is important because large differences between r^2 and cross-validated r^2 indicate that models are overdetermined; thus, the ability of the environmental data used in these models to explain benthic-community habitat relationships is unreliable.
Cross-validated r^2	The r^2 estimated by cross-validation analysis. A more honest or conservative measure of how well the environmental data explain benthic community-habitat relationships.

3. Results

3.1. General environmental and faunal characteristics

All sites were nearshore with maximum water depths ranging from 11 to 20 m. Minimum sampling depths were determined by the draft of the vessel used in the study (~ 2.6 m). The percentage of samples $\leq$ 4 m, the depth criterion from the LIS scheme, ranged from 3.3% at Robins Island to 48.1% at Haverstraw Bay. Sediment grain size varied broadly among study areas and, once classified, occupied 7 to 13 of 15 possible Folk (1954) grain size categories. A total of 15 cover classes were identified across the study areas surveyed (Table S3). Sites had between one and four CMECS geoforms present, and the number of provinces defined based on acoustic backscatter at each area ranged from 5 at Haverstraw Bay to 15 at Huntington Harbor. The Supplementary Material provides example surficial percent cover classifications derived from underwater videos and maps illustrating the configuration of the provinces at each study area.

Mean infaunal abundances and species richness per 0.04 m^2 grab sample in each study area were 106 individuals and 9 taxa for Haverstraw Bay, 103 individuals and 11 taxa in the Tappan Zee, 349 individuals and 15 taxa in Huntington Harbor, 279 individuals and 25 taxa in Robins Island, 380 individuals and 24 taxa per sample in Shelter Island. Overall species richness also varied across study areas with a total of 25 taxa for Haverstraw Bay, 40 taxa identified in the Tappan Zee, 82 in Huntington Harbor, 71 in Robins Island and 95 taxa in Shelter Island.

3.2. Summary of the benthic habitat classes identified using the LIS scheme and CMECS

The LIS scheme and CMECS produced identical habitat classes when applied to the data sets examined in this study (Tables S1 & Table S2). Consequently, models that used the categorical habitat classes created from these schemes will simply be referred to as habitat class models. The habitat classes predominantly included shallow ($\leq$ 4 m) and deep (> 4 m) areas with muddy, sandy, or slightly gravelly sediments that were generally devoid of biogenic features, such as mussel beds or reefs. A range of 10 to 26 habitat classes were identified at each study area, and the top 2 to 3 habitat classes with the greatest frequency of occurrence represented between 26.7 to 63.3% of the stations sampled. At Haverstraw Bay, habitat classes consisted of 1 to 6 sampling stations. Eleven of the 23 habitat classes had only one sample. The three habitat classes with the most sampling stations (5 to 6) were shallow ($\leq$ 4 m) with Folk (1954) sediment classes consisting of slightly gravelly sandy mud, deep (> 4 m) with slightly gravelly sandy mud, and deep with slightly gravelly mud. The two most common habitat classes found for the Tappan Zee data set included nearly half (45%) of the 100 stations sampled and consisted of deep areas with either mud or mud with slightly coarser sediments. The two habitat classes most common to Huntington Harbor represented 38% of the 76 stations sampled and included deep areas with gravelly muds or gravelly muddy sands. Robins Island had the largest fraction of its sampling stations represented by two habitat classes (63% of the 60 stations sampled). These consisted of deep areas with sediments composed of slightly gravelly muddy sands or slightly gravelly sandy muds. The two most common habitat classes found for the Shelter Island data set represented 42.9% of the 70 stations sampled and included deep areas with either slightly gravelly sands or gravelly muddy sands with *Crepidula fornicata* (slipper snail) beds.

3.3. Fixed categorical models & overfitting

The expected relationship between the coefficient of determination (r^2) and cross-validated r^2 was observed across all models (Figures 3-5), particularly those that used fixed categorical explanatory variables. In particular, r^2 increased monotonically with each binary split in MRT and as each variable was added in RDA. Conversely, cross-validated r^2 initially increased, reached a maximum, and then either leveled off or declined with subsequent additions of explanatory variables. Figure 3 provides examples of an MRT and RDA analysis illustrating this pattern. Models with large differences between r^2 and cross-validated r^2 overfitted the data sets analyzed and are therefore inadequate for reliably explaining patterns in benthic community-habitat relationships.

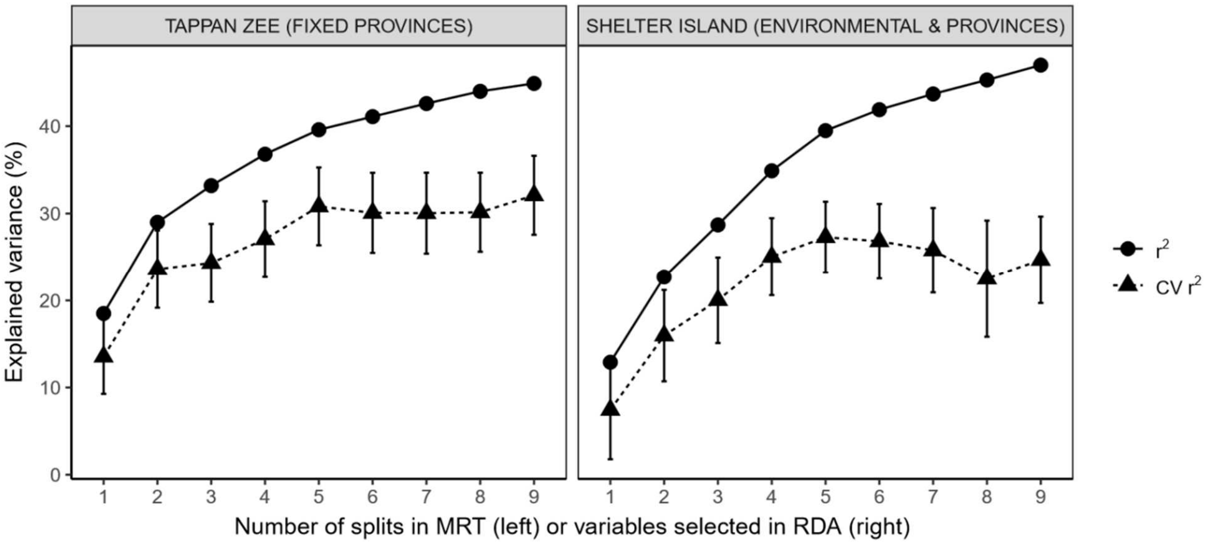


Figure 3. Differences between r^2 and cross-validated (CV) $r^2 \pm 1$ SE with increasing numbers of groups (terminal nodes) in MRT and variables selected in RDA. Examples are plotted using the results from the fixed province model at Tappan Zee and the linear environmental variables with province model at Shelter Island.

The fixed categorical sediment models (i.e., those with 7 to 13 Folk categories) consistently overfitted the data sets analyzed with r^2 ranging from 23.7 to 35.5% and cross-validated r^2 values that never exceeded 10%. Full versions of the LIS and CMECS habitat class models had r^2 values accounting for 34.1 to 69.0% of the total community variation (Table S4), but also overfitted all of the data sets analyzed, as did the models that combined water depth, sediment grain size, percent cover, and fixed geoforms or acoustic provinces. These models contained 7 to

24 variables and had r^2 values ranging from 32.0 to 73.9%, but cross-validated r^2 values were often < 0% (Table S4). Because of this consistent overfitting in the fixed categorical models (with the exception of fixed depth), only detailed results of the simplified categorical, flexible categorical, and continuous models will be reported in subsequent sections.

3.4. Models with single types & forms of environmental variables

Of the environmental variable types included in the single variable models, provinces explained the largest proportion of the benthic community variation (Figure 4). In the simplified categorical models, cross-validated r^2 for provinces was on average 4.8 times greater than water depth, 3.8 times greater than sediment grain size, 7.1 times greater than surficial percent cover, and 3.1 times greater than the CMECS geoforms (Figure 4). The CMECS geoforms at three of the five sites (Huntington Harbor, Robins Island, and Shelter Island) were derived without the use of sonar data and these models explained little to no community variation. However, the two sites with geoforms that were derived from sonar data based on CMECS criteria (Haverstraw Bay and Tappan Zee) fared better with cross-validated r^2 exceeding 20% like the provinces (Figure 4).

The water depth, sediment grain size, surficial percent cover, and geoform models rarely (8 of 44 cases) exceeded 10% cross-validated r^2 (Figure 4). The exceptions were the flexible water depth model at Huntington Harbor (11.0%), the linear sediment models at Tappan Zee (11.2%) and Huntington Harbor (10.8%), and the simplified categorical sediment model at Shelter Island (10.9%). In some cases, variable form (i.e., continuous vs. categorical) consistently influenced the proportion of community variation explained by the environmental variables across the data sets analyzed (Figure 4). For instance, models that utilized sediment grain size as a continuous explanatory variable always explained more of the community variation than the ones that used sediment grain size as fixed or simplified categorical explanatory variables based on the cross-validated r^2 (Figure 4). However, categorizing the continuous water depth, sediment grain size, and percent cover data did not substantially increase or decrease cross-validated r^2, with the exception of the fixed categorical models that overfitted the data sets analyzed. There are some additional structural patterns worth noting in regard to the comparative analyses of the fixed categorical, simplified categorical, flexible categorical, and continuous variables, which are addressed in the Discussion.

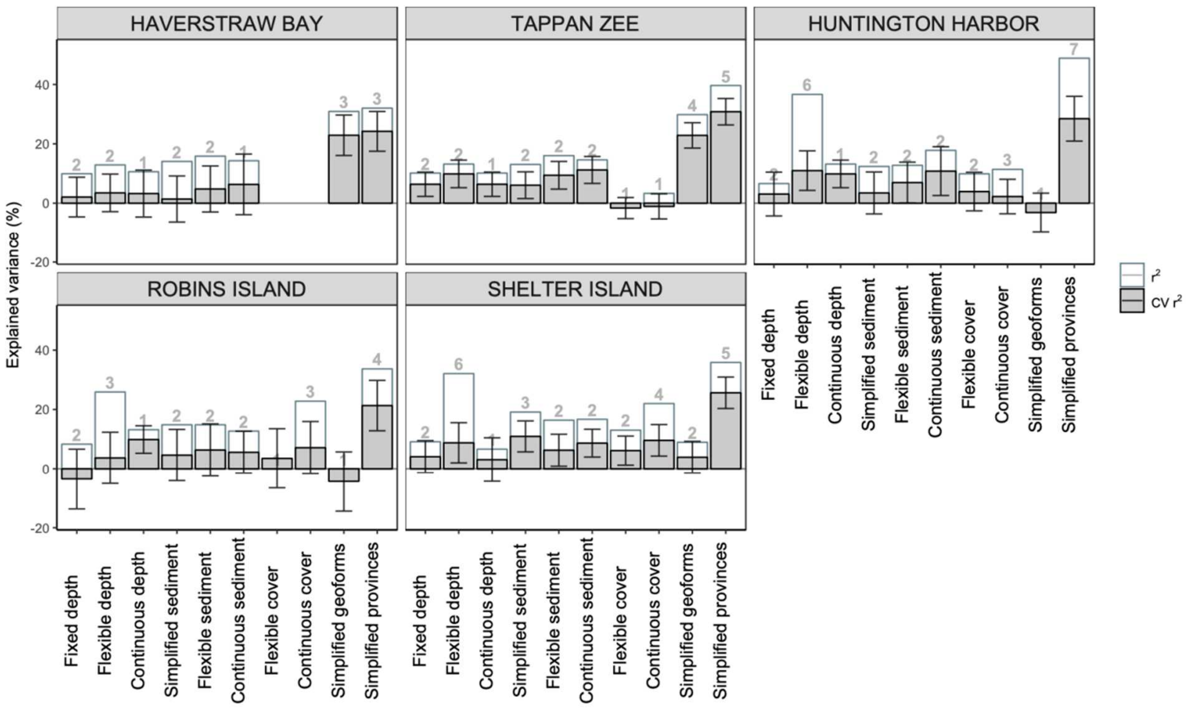


Figure 4. Infaunal community variation explained by models with single types of environmental variables at each study area. The variable type (water depth, sediment, etc.) and form (e.g., categorical vs. continuous) of each explanatory variable was evaluated using the framework outlined in Table 2. Both r^2 and cross-validated (CV) r^2 values ± 1 SE are plotted for comparison. The number of groups in MRT and variables selected in RDA by the cross-validation analysis is indicated for each model. Results for each data set were standardized by dividing the

explained variance by the total variance in the faunal data. Details about the MRT splits and variables selected in RDA are provided in Supplementary Material (Table S4).

3.5. Models using multiple types & forms of environmental variables

Multiple variable models with simplified provinces had 5.3 times greater cross-validated r^2 compared to the simplified habitat class models (Figure 5), and in models with simplified provinces and other types of environmental variables, cross-validated r^2 was 2.2 times greater on average than models without provinces. The models that utilized flexible categorical or continuous environmental variables with simplified geoforms had 3.2 times greater cross-validated r^2 than the simplified habitat class models (Figure 5). Moreover, models with water depth, sediment grain size, and/or surficial percent cover variables with geoforms as the explanatory variables were more comparable to those with provinces when the geoforms were based on applying the CMECS criteria using sonar backscatter data, as was the case for Haverstraw Bay and Tappan Zee. Models with provinces at these two areas only had 1.3 times greater cross-validated r^2 on average when compared to the models with geoforms but were 2.2 times greater than those at the other sites where geoforms were not derived from sonar (Figure 5). This reflects the same pattern found in the analyses of geoforms as a single type of explanatory variable.

The multiple variable models with simplified provinces and geoforms contained 1 to 6 explanatory variables after applying the cross-validation step (Figure 5). While these models were dominated by the simplified province and geoform variables, the other environmental variables selected varied across each study area (Table S4). Water depth and sediment grain size were selected across all data sets, except for Haverstraw Bay, and the simplified province and geoform models at Robins Island were the only ones with percent cover variables (percent cover of mud and *Microciona prolifera*). Moreover, explanatory variable form did not substantially impact the ability of the variable to explain community variation in the multiple variable models, aside from the fixed categorical variables that overfitted the data sets analyzed (Table S4).

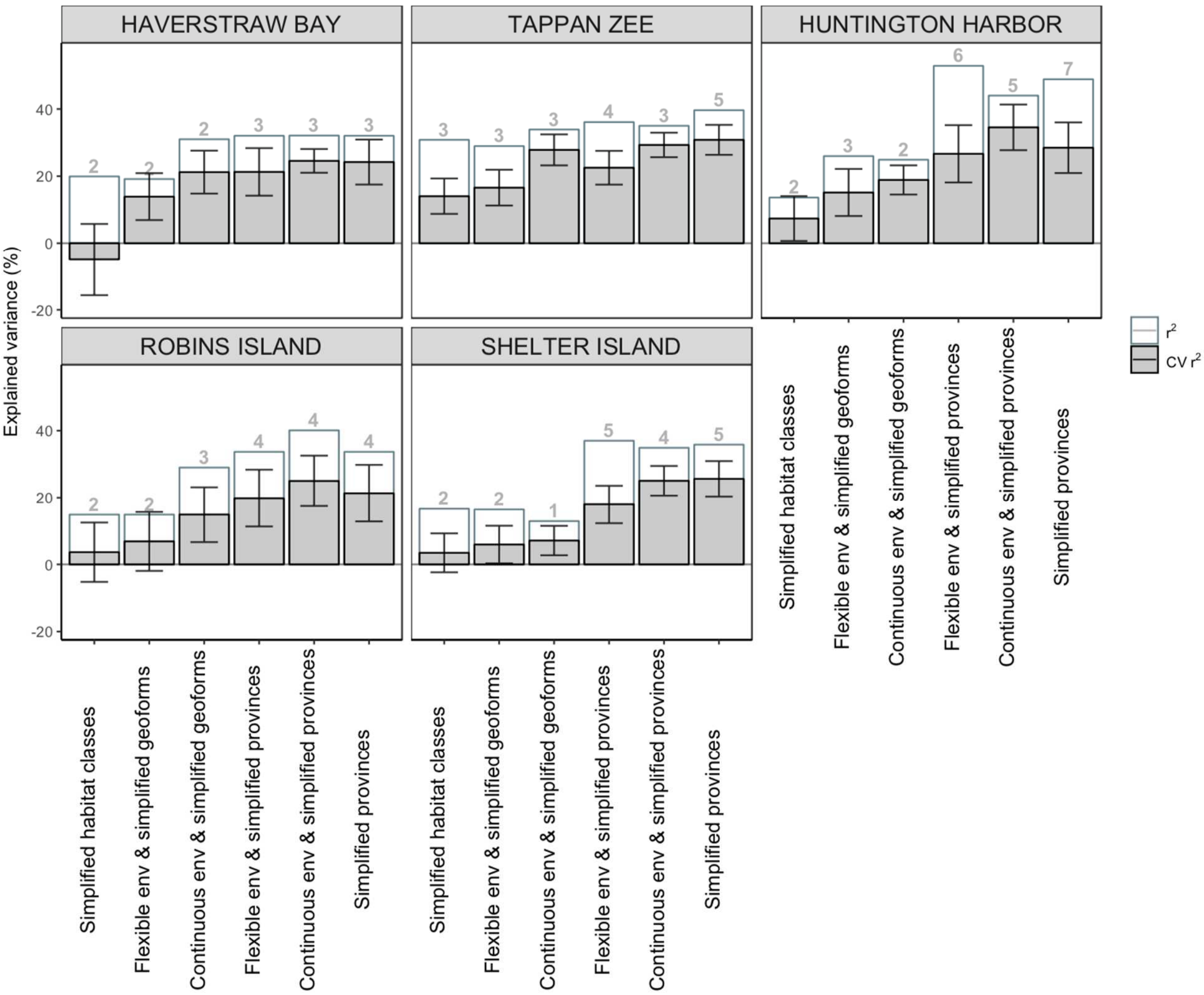


Figure 5. Infaunal community variation explained by models with multiple environmental variables at each study area. Each model was constructed using the framework outlined in Table 3. Results for models with fixed categorical variables are not presented since they consistently overfitted the data sets analyzed. Both r^2 and cross-validated (CV) r^2 values ± 1 SE are plotted. The number of groups in MRT and variables selected in RDA by the

cross-validation analysis are indicated for each model. Results for each data set were standardized by dividing the explained variance by the total variance in the faunal data. Simplified province models are plotted for comparison. More information on each model including details about the MRT splits and variables selected in RDA is provided in Supplementary Material.

4. Discussion

Our findings provide insight into questions of 1) variable form (e.g., continuous vs. categorical variables), 2) variable complexity (the number and types of explanatory variables utilized), 3) uniformity (consistent selection of explanatory variables across data sets), and 4) priority (critical to measure) that can be used to guide study designs and modeling efforts concerned with benthic habitat characterization and management in nearshore, soft-sediment ecosystems. First, we found that variable form influenced model performance but often not substantially. For instance, the continuous sediment grain size models generally explained slightly more of the community variation than the fixed or simplified categorical sediment models (Figure 4), and the models with continuous environmental variables and geoforms always had a greater cross-validated r^2 than the models with flexible environmental variables and geoforms (Figure 5). The one instance where variable form had a substantial impact involved the fixed categorical models because they overfitted the data sets analyzed and explained little cross-validated r^2. Second, models of moderate complexity with 2 to 6 explanatory variables outperformed those with more complex structures that generally overfitted the data. Third, provinces were uniformly selected across all data sets analyzed. Water depth and/or sediment grain size variables were also selected across most of the data sets in models with multiple types of environmental variables (Table S4), but provinces always explained the largest proportion of the community variation.

There are some general structural patterns worth noting in regard to the comparative analyses of the fixed categorical, simplified categorical, flexible categorical, and continuous environmental variables. For instance, the ± 4 m water depth criterion from the LIS scheme, i.e., the fixed categorical depth model, was almost always less effective at explaining community variation than the flexible or continuous linear depth models. Water depths identified by single binary splits in MRT ranged from 5 to 12 m, i.e., a depth always a greater than the ± 4 m depth criterion from the LIS scheme (Table S4).

The simplified categorical sediment models always consisted of only 2 to 3 categorical groups that were created by combining several of the fixed Folk categories in MRT into larger sets, and these simplified sets always had a cross-validated r^2 comparable to or greater than the fixed categorical sediment models (not shown). None of the sediment models had cross-validated r^2 values exceeding 11% and the average cross-validated r^2 was 6.1 $\pm$ 3.5% (sd), suggesting that sediment grain size is a moderate predictor of infaunal community structure in the data sets examined regardless of variable form.

Considering that the study sites examined were predominantly soft-bottom areas with a very low frequency of surficial biotic structure (e.g., in the form of shell, seaweed, etc.), it is not surprising that cross-validated r^2 never exceeded 10% in the models that utilized percent cover as the only explanatory variable. Just 10 of 100 samples were classified as mollusc reefs at Tappan Zee, 2 of 60 at Robins Island, and 8 of 50 at Haverstraw Bay. Only the Shelter Island site, with 38 of 70 samples characterized as a “Crepidula Reef”, had substantial biogenic structure. Although the slipper snails (*Crepidula fornicata*) occurred in consolidated aggregations of 5 to 10 individuals formed by preferential larval settlement (Zhao & Qian, 2002), the aggregations occurred on sand and were not attached to hard substrate, nor were they permanent structures.

Based on the comparison between the geoforms derived from applying the CMECS criteria using sonar data at the Haverstraw Bay and Tappan Zee sites and those taken from the Weaver et al. (2013) study that were derived with no sonar basis, it is clear that the habitat classification schemes would benefit from utilizing sonar data to define "Level 1" geoforms in CMECS (Table S1) and perhaps in applying the "Class Level" of the LIS scheme (Table S2). This does not, however, fully explain the generally poor performance of the habitat class models. It is important to note that this overall poor quantitative assessment of the habitat class models does not negate the value of these classification schemes in producing a common nomenclature for describing habitats, only that the habitat classes cannot be used at face value as a quantitative, categorical representation of the habitat structure for infauna in soft-sediment, nearshore environments.

There was strikingly little difference in cross-validated r^2 values for models using only provinces compared to those that combined provinces with other environmental variables (Figure 5), suggesting that there was little additional explanatory contribution to including depth, grain size, and percent cover variables to models with provinces. This outcome raises questions relevant to the spatial scale of the explanatory variables included in our analyses. Depth, grain size, and percent cover are *in situ* variables, i.e., collected at the same locations as the faunal samples. Provinces are seascape scale variables representing broad areas of the seafloor. Flanagan et al (2018) examined scaling questions at four of the five sites in the current study (Haverstraw Bay, Tappan Zee, Robins Island, and Shelter Island). They found that the within-province explanatory value of water depth, grain size, and percent cover was weak, and these variables primarily contributed to explaining between-province variability in fauna, i.e., something that the categorical province variables were also representing. They suggested further that one explanation for their result might be that within-province faunal variation was being regulated by patchy rather than gradational factors (e.g., water depth gradients) that were not being measured.

In this study, the provinces far outperformed the LIS scheme and CMECS habitat classes as well as all other environmental data collected. This outcome is particularly useful from a management perspective. Provinces represent habitats at seascape scales, which are mappable and therefore easier to identify and monitor. Moreover, in the case of the study areas examined, the provinces were relatively easy and cost-efficient to derive from backscatter intensity data. The provinces were created using visual analysis of the backscatter data from each study area, an inherently subjective approach. Thus, in the future, it may be worthwhile to compare the provinces in this study to those derived using more objective image processing techniques such as pixel-based methods (e.g., Jenks natural breaks for unsupervised classification of backscatter intensity images (Jenks, 1967 but also see Janowski et al., 2018) and/or object-based image analysis (e.g., Ismail et al., 2015; Janowski et al., 2018), which can be implemented using various classification algorithms including classification and regression tree analyses (Breiman et al., 1984), random forests (Breiman, 2001), support vector machines (Cortes & Vapnik, 1995), and k-nearest neighbor analyses (e.g., Janowski et al., 2018). However, based on our assessment of the backscatter data collected from the sites in this study, it seems unlikely that these more sophisticated and objective approaches would have yielded a different outcome.

Other measures derived from sonar data (e.g., rugosity, slope, aspect, etc.) have been useful in describing and predicting clear patterns in the abundance and distribution of benthic epifauna

(Kostylev et al., 2001; Holmes et al., 2008; Rattray et al., 2009; Pierdomenico et al., 2015) and demersal fishes (Iampietro et al., 2005; Wedding & Friedlander, 2008; Young et al., 2010). In previous work, sonar data have been used as proxies for natural phenomena (e.g., exposure to wave action, subtidal currents, and vulnerability to sedimentation) that could conceivably govern patterns in benthos, but which are typically unmeasured (Rattray et al., 2009, 2013). While the utility of sonar in characterizing habitats for benthic epifauna and demersal fishes has been well-documented, the explanatory value of sonar in characterizing benthic infauna has been relatively unexplored (Brown et al., 2011). Using improved techniques for the classification of infaunal habitats is particularly important since large areas of the ocean floor primarily consist of soft sediments (Rhoads, 1974), and thus infaunal habitats are conceivably the most widespread in nature.

A significant finding to understanding the ecology and distribution of benthic infaunal assemblages in the present study is that it can be easy to overfit benthic community-habitat models using environmental data sets that are commonly included as part of habitat characterization studies (e.g., water depth, sediment grain size, etc.). This was clearly and repeatedly indicated in the analyses by the large differences between the coefficient of determination (r^2) and cross-validated r^2 especially when full models were examined (Table S4). The overfitting problem was also illustrated by the large differences in r^2 and the numbers of variables selected between the fixed and simplified models. The cross-validation procedure indicated, with few exceptions, that models with 5 or fewer variables tended to be selected.

This outcome, consistent across all five study sites, has several consequences for data sets of comparable size and composition. First, a fully structured model made up of multiple types of environmental variables will give deceptively high r^2 values that cannot be validated. This holds true for all models considered in this study that used combinations of environmental variables (Figure 5), and includes the categorical habitat class models. Second, collecting larger numbers of benthic samples is needed in order to increase the chance of validation; however, this may be prohibitively expensive (Deborde et al., 2016; Marshall et al., 2018). The data sets examined in the present study each required 9 to 18 months effort to produce. With the possible exception of Haverstraw Bay, where sampling density was 1.7 samples per km^2, sampling density ranged from 5 to 18 samples per km^2. Yet, there was no obvious relationship between sample size, sample density, and the number of selected parameters. For example, the RDA model selection process that used all explanatory data resulted in 3 variables for both the Haverstraw Bay and Tappan Zee sites despite the differences in sample size (51 vs. 100) and sample density (1.7 samples per km^2 vs. 10.6 samples per km^2). Burnham and Anderson (2002) suggested that effect sizes taper in biological systems with a few large effects followed by progressively smaller ones, and that the smaller effects require a very large number of samples to identify. This statement is applicable to the data sets in the present study, and results suggest that the larger effects are being captured. Unfortunately, there was no clear indication of how much additional effort would be required to detect smaller effects.

Third, given the practical limitations of identifying smaller effects in these data sets, it seems imperative to use existing and to discover new environmental variables that integrate ecological processes in order to “package” the largest amount of explanatory value in the fewest number of variables. Since benthic fauna are connected to the environment at very fine scales (e.g., tubes engineered by infauna may influence benthic boundary layer flow; Shumchenia & King, 2010), it is curious that the acoustic provinces, by far the largest-scale explanatory variable used in the present study, explained substantially more of the community variance in the data sets relative to other types of environmental variables. This curious feature is explained by the fact that sonar backscatter reflects multiple physical (e.g., grain size, compaction, porosity, sorting, volume scattering) and biogenic (e.g., shell beds, shell hash, bioturbation, and tube mats) features and processes (Jackson & Briggs, 1992; Borgeld et al., 1999; Goff et al., 2000; Brown et al., 2002, 2011; Urgeles et al., 2002; Cutter et al., 2003; Ferrini, 2004; Nitsche et al., 2004, 2007). Thus, obvious candidates for new explanatory variables include metrics derived from detailed analysis of sonar backscatter and bathymetry data (Huvenne et al., 2002; Maher, 2006; Fonseca & Mayer, 2007; Holmes et al., 2008; Rattray et al., 2009), and perhaps others that integrate grain size, water content, and shallow sediment structure such as a bottom penetrometer (Stark & Wever, 2009).

Finally, the current study results emphasize why ground truth sampling of the fauna is absolutely essential. Quantitative macrofaunal community data are "expensive and time consuming" (Verfaillie et al., 2009) to collect and often lacking in habitat classification and mapping studies (Ismail et al., 2015). Geologic or geophysical features detectable by surveys that appear to characterize spatially distinct sedimentary regions (e.g., sand veneers, large sand waves, rippled sand) are not necessarily ecologically relevant (Snelgrove & Butman, 1994; Brown et al., 2002; McBreen et al., 2008). It is clear from the results of the current study that complex abiotic-biotic data sets have limitations, and without careful analysis, models can easily over-characterize by fitting spurious variation (i.e., "noise"). Thus, habitat identification and characterization models must be carefully and rigorously tested. The model selection process used in this study focused on identifying patterns in the data sets rather than on identifying a model with high r^2.

Results presented in this paper are broadly applicable to studies and management efforts concerned with explaining benthic community-habitat relationships, highlight the importance of maximizing the use of sonar data in terms of its ability to identify and characterize benthic systems without overfitting the data, and draw attention to the problem of over-characterization in the context of habitat classification. Future emphasis should be placed on deriving new variables or measures from sonar that enhance our ability to explain community structure. Useful variables would include those that explain a substantial proportion of community variation or at least match that explained by the acoustic provinces in this study (i.e., a minimum of 20%). Sonar data are particularly useful in this context since they can be segmented across multiple spatial scales and used to create new variables (e.g., rugosity as a proxy for habitat complexity, slope as a proxy for larval dispersal and settlement, etc.). In addition, segmenting sonar data across multiple spatial scales enables efforts that test the impact of observational scale on one's ability to explain variation in biological communities – an area of inquiry that is broadly relevant to and critical for habitat identification and characterization efforts within and outside of benthic marine systems.

Acknowledgements

This manuscript results from projects "Integrating Fish Trawl, Water Quality, and Benthic Mapping Data in the Peconic Estuary", "Northern Harbors of Long Island", and "Acoustic Mapping and Habitat Characterization in Haverstraw Bay" funded by New York Department of Environmental Conservation, from projects R/CO-6 and R/CO-7 funded under award 22915 from the National Sea Grant College Program of the U.S. Department of Commerce's National Oceanic and Atmospheric Administration, to the Research Foundation of the State University of

New York on behalf of New York Sea Grant, and from awards from Suffolk County, NY, and The Nature Conservatory, Long Island Chapter, for studies in the Peconic Bays. The statements, findings, conclusions, views, and recommendations are those of the authors and do not necessarily reflect the views of any of those organizations.

References

Akaike, H., 1973. Information theory as an extension of the maximum likelihood principle. In: Petrov BN and Csaki F, eds. Second International Symposium on Information Theory. Akademiai Kiado, Budapest.

Allee, R. J., M. Dethier, D. Brown, L. Deegan, G. R. Ford, T. R. Hourigan, J. Maragos, et al., 2000. Marine and Estuarine Ecosystem and Habitat Classification. National Oceanic and Atmospheric Administration technical memorandum. Silver Spring, Maryland.

Auster, P. J., K. B. Heinonen, C. Witharana, & M. Mckee, 2009. A Habitat Classification Scheme for The Long Island Sound Region Long Island Sound Study Technical Report. Stamford, Connecticut, USA.

Bell, R.E., Flood, R.D., Carbotte, S.M, Ryan W.B.F., McHugh, C., Cormier, M., Versteeg, R., Chayes, D., Bokuniewicz, H., F., 2000. Hudson River Estuary Program Benthic Mapping Project Revised Final Report. Submitted to the New York State Department of Environmental Conservation. .

Bizzarro, J. J., M. M. Yoklavich, & W. W. Wakefield, 2017. Diet composition and foraging ecology of U.S. Pacific Coast groundfishes with applications for fisheries management.

Environmental Biology of Fishes Environmental Biology of Fishes 100: 375–393.

Borgeld, J. C., J. E. Hughes Clarke, J. A. Goff, L. A. Mayer, & J. A. Curtis, 1999. Acoustic backscatter of the 1995 flood deposit on the Eel shelf. Marine Geology 154: 197–210.

Borja, Á., J. Franco, & V. Pérez, 2000. A Marine Biotic Index to Establish the Ecological Quality of Soft-Bottom Benthos Within European Estuarine and Coastal Environments. Marine Pollution Bulletin 40: 1100–1114, http://linkinghub.elsevier.com/retrieve/pii/S0025326X00000618.

Borja, Á., S. L. Marín, I. Muxika, L. Pino, & J. G. Rodríguez, 2015. Is there a possibility of ranking benthic quality assessment indices to select the most responsive to different human pressures?. Marine Pollution Bulletin 97: 85–94, https://linkinghub.elsevier.com/retrieve/pii/S0025326X15003975.

Bottom, D. L., & K. K. Jones, 1990. Species composition, distribution, and invertebrate prey of fish assemblages in the Columbia River Estuary. 25: 243–270.

Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C. G., 1984. Classification and Regression Trees. Wadsworth International Group, Belmont, California, USA.

Breiman, L., 2001. Random Forests. Macine Learning 5–32.

Brown, C. J., K. M. Cooper, W. J. Meadows, D. S. Limpenny, & H. L. Rees, 2002. Small-scale mapping of sea-bed assemblages in the eastern English Channel using sidescan sonar and remote sampling techniques. Estuarine, Coastal and Shelf Science 54: 263–278.

Brown, C. J., S. J. Smith, P. Lawton, & J. T. Anderson, 2011. Benthic habitat mapping: A review of progress towards improved understanding of the spatial ecology of the seafloor using acoustic techniques. Estuarine, Coastal and Shelf Science 92: 502–520, http://linkinghub.elsevier.com/retrieve/pii/S0272771411000485.

Burnham, David A. Anderson, K. P., 2002. Model Selection and Multi-Model Inference : A Practical Information-Theoretic Approach (2nd Edition). Ecological Modelling. Springer, New York, http://linkinghub.elsevier.com/retrieve/pii/S0304380003004526.

Cerrato, R. M., Flanagan, A.M., Flood, R. D., 2015. Haverstraw Bay Benthic Habitat Characterization. Marine Sciences Research Center Special Report No. 141. Stony Brook, New York, USA.

Cerrato, R. M., Holt, L., 2008. North Shore Bays Benthic Mapping: Ground Truth Studies. Final Report to the NYS Department of Environmental Conservation. Marine Sciences Research Center Special Report No. 135. Stony Brook, New York, USA.

Cerrato, R. M., Maher, N. P., 2007. Benthic Mapping for Habitat Classification in the Peconic Estuary: Phase I Ground Truth Studies. Marine Sciences Research Center Special Report No. 134. Stony Brook, New York, USA.

Cerrato, R. M., D. A. Caron, D. J. Lonsdale, J. M. Rose, & R. A. Schaffner, 2004. Effect of the northern quahog Mercenaria mercenaria on the development of blooms of the brown tide alga Aureococcus anophagefferens. Marine Ecology Progress Series 281: 93–108.

Cloern, J., 1982. Does the Benthos Control Phytoplankton Biomass in South San Francisco

Bay?. Marine Ecology Progress Series 9: 191–202, http://www.int-res.com/articles/meps/9/m009p191.pdf.

Cortes, C., & V. Vapnik, 1995. Support-vector networks. Machine Learning 20: 273–297, http://link.springer.com/10.1007/BF00994018.

Crawley, M. J., 2007. The R Book. John Wiley & Sons, Ltd, Chichester, UK, http://doi.wiley.com/10.1002/9780470515075.

Cutter, G. R., Y. Rzhanov, & L. A. Mayer, 2003. Automated segmentation of seafloor bathymetry from multibeam echosounder data using local fourier histogram texture features. Journal of Experimental Marine Biology and Ecology 285–286: 355–370.

Davies, C. E., D. Moss, & M. O. Hill, 2004. EUNIS Habitat Classification Revised 2004. Technology. .

De’ath, G., 2002. Multivariate regression trees: a new technique for modeling species-environment relationships. Ecology 83: 1105–1117, https://doi.org/10.1890/0012-9658(2002)083[1105:MRTANT]2.0.CO;2.

De’ath, G., 2014. mvpart: Multivariate partitioning. R package version 1.6-2. https://CRAN.R-project.org/package=mvpart. .

Deborde, D. D. D., M. B. M. Hernandez, & F. S. Magbanua, 2016. Benthic Macroinvertebrate Community as an Ind icator of Stream Health : The Effects of Land Use on Stream Benthic Macroinvertebrates. Science Diliman 28: 5–26.

Diaz, R. J., M. Solan, & R. M. Valente, 2004. A review of approaches for classifying benthic habitats and evaluating habitat quality. Journal of Environmental Management 73: 165–181.

Ferraro, S. P., & F. A. Cole, 2012. Ecological periodic tables for benthic macrofaunal usage of estuarine habitats: Insights from a case study in Tillamook Bay, Oregon, USA. Estuarine, Coastal and Shelf Science Elsevier Ltd 102–103: 70–83, http://dx.doi.org/10.1016/j.ecss.2012.03.009.

Ferrini, V., 2004. Nearshore sedimentary dynamics revealed through the analysis of high frequency multibeam sonar data, Ph.D. Thesis. Stony Brook University, Stony Brook, New York.

FGDC, 2012. Coastal and Marine Ecological Classification Standard. Charleston, South Carolina.

Flanagan, A. M., 2016. Quantitative Benthic Community Models: The Relationship Between Explained Variance and Scale, Ph.D. Thesis. Stony Brook University, Stony Brook, New York.

Flanagan, A. M., & R. M. Cerrato, 2015. An approach for quantifying the efficacy of ecological classification schemes as management tools. Continental Shelf Research 109: 55–66.

Flanagan, A. M., R. D. Flood, M. G. Frisk, C. D. Garza, G. R. Lopez, N. P. Maher, & R. M. Cerrato, 2018. The relationship between observational scale and explained variance in benthic communities. PLoS ONE 13: 1–25.

Folk, R. L., 1954. The distinction between grain size and mineral composition in sedimentary-rock nomenclature The University of Chicago Press. The University of Chicago Press 62: 344–

359.

Folk, R. L., 1974. Petrology of Sedimentary Rock. Hemphill Publishing Company, Austin, Texas, USA.

Fonseca, L., & L. Mayer, 2007. Remote estimation of surficial seafloor properties through the application Angular Range Analysis to multibeam sonar data. Marine Geophysical Researches 28: 119–126.

Fraschetti, S., A. Terlizzi, & L. Benedetti-cecchi, 2005. Patterns of distribution of marine assemblages from rocky shores: Evidence of relevant scales of variation. Marine Ecology Progress Series 296: 13–29.

Goff, J. A., H. C. Olson, & C. S. Duncan, 2000. Correlation of side-scan backscatter intensity with grain-size distribution of shelf sediments, New Jersey margin. Geo-Marine Letters 20: 43–49.

Golub, G. H., M. Heath, & G. Wahba, 1979. Generalized cross-validation as a method for choosing a good ridge parameter. 21: 215–223.

Hastie, T., R. Tibshirani, & J. Friedman, 2001. The Elements of Statistical Learning. The Mathematical Intelligencer. Springer, New York, http://www.springerlink.com/index/D7X7KX6772HQ2135.pdf%255Cnhttp://www-stat.stanford.edu/~tibs/book/preface.ps.

Holmes, K. W., K. P. Van Niel, B. Radford, G. A. Kendrick, & S. L. Grove, 2008. Modelling distribution of marine benthos from hydroacoustics and underwater video. Continental Shelf

Research 28: 1800–1810.

Huvenne, V. A. I., P. Blondel, & J. P. Henriet, 2002. Textural analyses of sidescan sonar imagery from two mound provinces in the Porcupine Seabight. Marine Geology 189: 323–341.

Iampietro, P. J., R. G. Kvitek, & E. Morris, 2005. Recent Advances in Automated Genus-specific. Marine Technology Society Journal 39: 83–93.

Ismail, K., V. A. I. Huvenne, & D. G. Masson, 2015. Objective automated classification technique for marine landscape mapping in submarine canyons. Marine Geology 362: 17–32.

Jackson, D. R., & K. B. Briggs, 1992. High-frequency bottom backscattering: Roughness versus sediment volume scattering. The Journal of the Acoustical Society of America 92: 962–977, http://asa.scitation.org/doi/10.1121/1.403966.

Janowski, L., K. Trzcinska, J. Tegowski, A. Kruss, M. Rucinska-Zjadacz, & P. Pocwiardowski, 2018. Nearshore Benthic Habitat Mapping Based on Multi-Frequency, Multibeam Echosounder Data Using a Combined Object-Based Approach: A Case Study from the Rowy Site in the Southern Baltic Sea. Remote Sensing 10: 1983.

Jenks, G., 1967. The data model concept in statistical mapping. International Yearbook of Cartography 7: 186–190.

Jongman, R. H. G., C. J. F. ter Braak, & O. F. R. Van Tongeren, 1995. Data Analysis in Community and Landscape Ecology. Biometrics 46: 287, https://www.jstor.org/stable/2531665?origin=crossref.

Kostylev, V. E., B. J. Todd, G. B. J. Fader, R. C. Courtney, G. D. M. Cameron, & R. A. Pickrill, 2001. Benthic habitat mapping on the Scotian Shelf based on multibeam bathymetry, surficial geology and sea floor photographs. Marine Ecology Progress Series 219: 121–137.

Kröncke, I., & H. Reiss, 2010. Influence of macrofauna long-term natural variability on benthic indices used in ecological quality assessment. Marine Pollution Bulletin 60: 58–68, http://linkinghub.elsevier.com/retrieve/pii/S0025326X09003671.

Lecours, V., C. J. Brown, R. Devillers, V. L. Lucieer, & E. N. Edinger, 2016. Comparing selections of environmental variables for ecological studies: A focus on terrain attributes. PLoS ONE 11: 1–18.

Lecours, V., R. Devillers, D. Schneider, V. Lucieer, C. Brown, & E. Edinger, 2015. Spatial scale and geographic context in benthic habitat mapping: review and future directions. Marine Ecology Progress Series 535: 259–284, http://www.int-res.com/abstracts/meps/v535/p259-284/.

Legendre, P., & E. D. Gallagher, 2001. Ecologically meaningful transformations for ordination of species data. Oecologia 129: 271–280.

Legendre P, L. L., 1998. Numerical Ecology. Amsterdam, The Netherlands.

Maceda-Veiga, A., R. López, & A. J. Green, 2017. Dramatic impact of alien carp Cyprinus carpio on globally threatened diving ducks and other waterbirds in Mediterranean shallow lakes. Biological Conservation 212: 74–85, https://linkinghub.elsevier.com/retrieve/pii/S0006320716309776.

Maher, N. P., 2006. A New Approach for Benthic Biotope Identification and Mapping. Ph.D.

thesis. Stony Brook University, Stony Brook, New York.

Marshall, J. E., D. J. Bucher, & S. D. A. Smith, 2018. Patterns of infaunal macromollusc assemblages in a subtropical marine park: Implications for management. Marine and Freshwater Research 69: 502–513.

McBreen, F., J. G. Wilson, A. S. Y. Mackie, & C. Nic Aonghusa, 2008. Seabed mapping in the southern Irish Sea: Predicting benthic biological communities based on sediment characteristics. Hydrobiologia 606: 93–103.

McCormick, M. I., 1995. Fish feeding on mobile benthic invertebrates: influence of spatial variability in habitat associations. Marine Biology 121: 627–637, http://link.springer.com/10.1007/BF00349298.

McGonigle, C., C. Brown, R. Quinn, & J. Grabowski, 2009. Evaluation of image-based multibeam sonar backscatter classification for benthic habitat discrimination and mapping at Stanton Banks, UK. Estuarine, Coastal and Shelf Science 81: 423–437.

Nitsche, F. O., R. Bell, S. M. Carbotte, W. B. F. Ryan, & R. Flood, 2004. Process-related classification of acoustic data from the Hudson River Estuary. Marine Geology 209: 131–145.

Nitsche, F. O., W. B. F. Ryan, S. M. Carbotte, R. E. Bell, A. Slagle, C. Bertinado, R. Flood, T. Kenna, & C. McHugh, 2007. Regional patterns and local variations of sediment distribution in the Hudson River Estuary. Estuarine, Coastal and Shelf Science 71: 259–277.

Oksanen, J., Blanchet, F.G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., & H. Minchin, P.R., O'Hara, R. B., Simpson, G.L., Solymos, P., Henry, M., Stevens, H., Szoecs, E., Wagner,

2017. vegan: Community Ecology Package. R package version 2.4-4. https://CRAN.R-project.org/package=vegan. .

Palumbi, S. R., S. D. Gaines, H. Leslie, & R. R. Warner, 2003. New Wave: High-Tech Tools to Help Marine Reserve Research. Frontiers in Ecology and the Environment 1: 73, http://doi.wiley.com/10.2307/3868033.

Parks, N., 2002. A lingua franca for marine habitat classification - an idea whose time has come. Bioscience 52: 324, isi:000174881000002.

Pearson, T. H., R. Rosenberg, & R. Pearson, T.H., Rosenberg, 1978. Macrobenthic succession in relation to organic enrichment and pollution of the marine environment. Oceanography and marine biology annual review 16: 229–311.

Pelletier, M. C., D. J. Gillett, A. Hamilton, T. Grayson, V. Hansen, E. W. Leppo, S. B. Weisberg, & A. Borja, 2018. Adaptation and application of multivariate AMBI (M-AMBI) in US coastal waters. Ecological Indicators 89: 818–827, https://linkinghub.elsevier.com/retrieve/pii/S1470160X17305514.

Pérez-Vargas, A. D., M. Bernal, C. S. Delgadillo, E. F. González-Navarro, & M. F. Landaeta, 2016. Benthic food distribution as a predictor of the spatial distribution for shorebirds in a wetland of central Chile. Revista de biología marina y oceanografía 51: 147–159, http://www.scielo.cl/scielo.php?script=sci_arttext&pid=S0718-19572016000100014&lng=en&nrm=iso&tlng=en.

Petersen, C. G. J., 1913. Valuation of the sea. II. The animal communities of the sea bottom and their importance for marine zoogeography. Report of the Danish Biological Station 16: 229–311.

Pierdomenico, M., V. G. Guida, L. Macelloni, F. L. Chiocci, P. A. Rona, M. I. Scranton, V. Asper, & A. Diercks, 2015. Sedimentary facies, geomorphic features and habitat distribution at the Hudson Canyon head from AUV multibeam data. Deep-Sea Research Part II: Topical Studies in Oceanography 121: 112–125.

R Core Development Team, 2017. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/. .

Rattray, A., D. Ierodiaconou, L. Laurenson, S. Burq, & M. Reston, 2009. Hydro-acoustic remote sensing of benthic biological communities on the shallow South East Australian continental shelf. Estuarine, Coastal and Shelf Science Elsevier Ltd 84: 237–245, http://dx.doi.org/10.1016/j.ecss.2009.06.023.

Rattray, A., D. Ierodiaconou, J. Monk, V. L. Versace, & L. J. B. Laurenson, 2013. Detecting patterns of change in benthic habitats by acoustic remote sensing. Marine Ecology Progress Series 477: 1–13.

Rhoads, D. C., 1974. Organism-sediment relations on the muddy sea floor. Oceanography and marine biology annual review 12: 263–300.

Richman, S. E., & J. R. Lovvorn, 2009. Predator size, prey size and threshold food densities of diving ducks: does a common prey base support fewer large animals?. Journal of Animal Ecology 78: 1033–1042, http://doi.wiley.com/10.1111/j.1365-2656.2009.01556.x.

Sanders, H. L., 1958. Benthic Studies in Buzzards Bay. I. Animal-Sediment Relationships.

Limnology and Oceanography 3: 245–258.

Shumchenia, E. J., & J. W. King, 2010. Comparison of methods for integrating biological and physical data for marine habitat mapping and classification. Continental Shelf Research Elsevier 30: 1717–1729, http://dx.doi.org/10.1016/j.csr.2010.07.007.

Silva, G., J. L. Costa, P. R. de Almeida, & M. J. Costa, 2006. Structure and Dynamics of a Benthic Invertebrate Community in an Intertidal Area of the Tagus Estuary, Western Portugal: A Six Year Data Series. Hydrobiologia 555: 115–128, http://link.springer.com/10.1007/s10750-005-1110-8.

Smale, D. A., 2008. Continuous benthic community change along a depth gradient in Antarctic shallows: Evidence of patchiness but not zonation. Polar Biology 31: 189–198.

Snelgrove, P. V. R., & C. A. N. N. Butman, 1994. Animal-sediment relationships revisited: cause versus effect. 111–177.

Stark, N., & T. F. Wever, 2009. Unraveling subtle details of expendable bottom penetrometer (XBP) deceleration profiles. Geo-Marine Letters 29: 39–45.

Stevens, T., & R. M. Connolly, 2004. Testing the utility of abiotic surrogates for marine habitat mapping at scales relevant to management. Biological Conservation 119: 351–362.

Sutherland, W. J., R. P. Freckleton, H. C. J. Godfray, S. R. Beissinger, T. Benton, D. D. Cameron, Y. Carmel, D. A. Coomes, T. Coulson, M. C. Emmerson, R. S. Hails, G. C. Hays, D. J. Hodgson, M. J. Hutchings, D. Johnson, J. P. G. Jones, M. J. Keeling, H. Kokko, W. E. Kunin, X. Lambin, O. T. Lewis, Y. Malhi, N. Mieszkowska, E. J. Milner-Gulland, K. Norris, A. B.

Phillimore, D. W. Purves, J. M. Reid, D. C. Reuman, K. Thompson, J. M. J. Travis, L. A. Turnbull, D. A. Wardle, & T. Wiegand, 2013. Identification of 100 fundamental ecological questions. Journal of Ecology 101: 58–67.

Taylor, R. B., 1998. Density, biomass and productivity of animals in four subtidal rocky reef habitats: The importance of small mobile invertebrates. Marine Ecology Progress Series 172: 37–51.

Urgeles, R., J. Locat, T. Schmitt, & J. E. Hughes Clarke, 2002. The July 1996 flood deposit in the Saguenay Fjord, Quebec, Canada: Implications for sources of spatial and temporal backscatter variations. Marine Geology 184: 41–60.

Verfaillie, E., S. Degraer, K. Schelfaut, W. Willems, & V. Van Lancker, 2009. A protocol for classifying ecologically relevant marine zones, a statistical approach. Estuarine, Coastal and Shelf Science 83: 175–185.

Weaver, K.J., E.J. Shumchenia, K.H. Ford, M.A. Rousseau, J.K. Greene, M. G. A. and J. W. K., 2013. Application of the Coastal and Marine Ecological Classification Standard (CMECS) to the Northwest Atlantic. The Nature Conservancy, Eastern Division Conservation Science, Eastern Regional Office. Boston, MA. , http://nature.ly/EDcmecs.

Wedding, L. M., & A. Friedlander, 2008. Determining the influence of seascape structure on coral reef fishes in Hawaii using a geospatial approach. Marine Geodesy 31: 246–266.

Welsh, D. T., 2003. It’s a dirty job but someone has to do it: The role of marine benthic macrofauna in organic matter turnover and nutrient recycling to the water column. Chemistry and Ecology 19: 321–342, http://www.tandfonline.com/doi/abs/10.1080/02757540310001554 74.

Young, M. A., P. J. Iampietro, R. G. Kvitek, & C. D. Garza, 2010. Multivariate bathymetry-derived generalized linear model accurately predicts rockfish distribution on Cordell Bank, California, USA. Marine Ecology Progress Series 415: 247–261.

Zhao, B., & P. Y. Qian, 2002. Larval settlement and metamorphosis in the slipper limpet Crepidula onyx (Sowerby) in response to conspecific cues and the cues from biofilm. Journal of Experimental Marine Biology and Ecology 269: 39–51.

Supplementary Material

Assignment of benthic habitat classes using the criteria from the Long Island Sound classification scheme (Auster et al., 2009) and the Coastal and Marine Ecological Classification Standard (FGDC, 2012)

Two habitat classification schemes were applied to the data sets in this study: (1) A classification scheme for the Long Island Sound (LIS) region (Auster et al. 2009), hereafter referred to as the LIS scheme and (2) the Coastal and Marine Ecological Classification Standard (CMECS) (FGDC 2012). The LIS scheme is largely hierarchical and is an adaptation of a deep water scheme by Greene et al. (1999). The broadest unit, the System level, divides an area into multiple seascapes (e.g., western, central LIS, etc.). Below the System level, Subsystem divides habitats into intertidal, shallow subtidal ($\leq$ 4 m), and deep subtidal (> 4 m). Classes within the Subsystems are based on large-scale morphological features such as channels, basins, and sand waves. The next two levels partition Classes based on sediment grain size and small-scale morphological features. At the final level, Modifiers, the scheme becomes non-hierarchical and requires characterizing a variety of physical, chemical, geological, biological, and anthropogenic features. The LIS scheme was implemented by assigning the environmental data from each sample a unique category code at the System through the Secondary Subclass levels (Table S1). The codes were then joined together by concatenation in a manner similar to that recommended by FGDC (2014). The fully formed code was used to produce unique categorical variables for analysis. As suggested by Weaver et al. (2013), the Class level in the LIS scheme is equivalent or nearly equivalent to Geoform in CMECS (described below), so the Geoform categories were used for Class.

Table S1. Classification of sampling stations using the LIS scheme (Auster et al. 2009). The Folk (1954) sediment categories under Primary Subclass level are represented by abbreviations that refer to descriptive names: G = gravel, mG = muddy gravel, msG = muddy sandy gravel, sG = sandy gravel, gmS = gravelly muddy sand, gM = gravelly mud, gmS = gravelly muddy sand, gS = gravelly sand, (g)M = slightly gravelly mud, (g)sM = slightly gravelly sandy mud, (g)S = slightly gravelly sand, M = mud, sM = sandy mud, mS = muddy sand, S = sand.

Study Area	System	Subsystem	Class	Primary Subclass (Folk 1954)				Secondary Subclass
Haverstraw Bay	Hudson River Estuary	Subtidal shallow	Mollusc Reef	G				None
		Subtidal deep	Flat	mG	msG	sG		
			Channel	gM	gmS	-		
			Dredged Channel	(g)M	(g)sM	(g)mS	-	
				M	sM	mS	S	
Tappan Zee	Hudson River Estuary	Subtidal shallow	Mollusc Reef	–				None
		Subtidal deep	Flat	mG	msG	sG		
			Wave Field	gM	-	-		
			Channel	(g)M	(g)sM	(g)mS	-	
				M	sM	mS	-	
Huntington Harbor	Western Long Island Sound	Subtidal shallow	Basin	–				None
		Subtidal deep		mG	msG	sG		
				gM	gmS	gS		
				(g)M	(g)sM	(g)mS	(g)S	
				M	-	-	-	
Robins Island	Peconic Bay Estuary	Subtidal shallow	Basin	–				Biogenic reef
		Subtidal deep		-	-	-		None
				gM	gmS	-		
				–	(g)sM	(g)mS	(g)S	
				-	sM	mS	S	
Shelter Island	Peconic Bay Estuary	Subtidal shallow	Basin	–				Biogenic reef
		Subtidal deep	Flat	mG	msG	sG		None
				gM	gmS	gS		
				–	–	–	(g)S	
				–	–	–	-	

CMECS is in its fourth version (FGDC 2012) and was preceded by a classification system by Allee et al. (2000). Environments in this scheme are initially assigned to one of three systems: Marine, Estuarine, and Lacustrine (Table S2). The Estuarine System, which represents all of the study areas, is divided into four Subsystems: Coastal ($\leq$ 4 m), Open Water (> 4 m), Tidal Riverine Coastal ($\leq$4m), and Tidal Riverine Open Water (> 4m; Table 2). All sampling locations were subtidal. Data sets were further characterized within four Components: Water Column, Geoform, Substrate, and Biotic. Classification within a Component is hierarchical, but each Component can be investigated independently of the others. In the present study, the Water Column Component was not utilized. As with the LIS scheme, environmental data were used to assign unique category codes to elements in the scheme, and these codes were joined by concatenation into categorical variables for analysis. Geoforms were defined to Level 1, the Substrate Component to Subgroup, and the Biotic Component to Subclass (Table S2).

Table S2. Classification of sampling stations using CMECS (FGDC 2012). The Folk (1954) sediment categories under Substrate Subgroup are represented by abbreviations that refer to descriptive names: G = gravel, mG = muddy gravel, msG = muddy sandy gravel, sG = sandy gravel, gmS = gravelly muddy sand, gM = gravelly mud, gmS = gravelly muddy sand, gS = gravelly sand, (g)M = slightly gravelly mud, (g)sM = slightly gravelly sandy mud, (g)S = slightly gravelly sand, M = mud, sM = sandy mud, mS = muddy sand, S = sand.

Study		Aquatic Setting	Geoform Component		Substrate Component								Biotic Component		
Area	System	Subsystem	Geoform Origin	Geoform	Substrate Origin	Substrate Class	Substrate Subclass	Substrate Group	Substrate Subgroup (Folk 1954)				Biotic Setting	Biotic Class	Biotic Subclass
Haverstraw Bay	Estuarine	Tidal Riverine Coastal	Anthropogenic	Dredged Channel	Geologic	Unconsolidated Mineral	Coarse Unconsolidated	Gravelly	G				Benthic/Attached Biota	Faunal Bed	Soft Sediment Fauna
		Tidal Riverine Open Water	Geologic	Flat			Fine Unconsolidated	Slightly Gravelly	mG	msG	sG				
				Channel				Mud	gM	gmS	-				
			Biogenic	Mollusc Reef					(g)M	(g)sM	(g)mS	-			
									M	sM	mS	S			
					Biogenic	Shell	Shell Rubble	Oyster Rubble							
Tappan Zee	Estuarine	Tidal Riverine Coastal	Geologic	Flat	Geologic	Unconsolidated Mineral	Coarse Unconsolidated	Gravelly	–				Benthic/Attached Biota	Faunal Bed	Soft Sediment Fauna
		Tidal Riverine Open Water		Channel				Gravel Mixes	mG	msG	sG				
				Wave Field			Fine Unconsolidated	Slightly Gravelly	gM	-	-				
			Biogenic	Mollusc Reef				Sandy Mud	(g)M	(g)sM	(g)mS	-			
								Muddy Sand	M	sM	mS	-			
								Mud							
					Biogenic	Shell	Shell Rubble	Oyster Rubble							
Huntington Harbor	Estuarine	Coastal	Geologic	Basin	Geologic	Unconsolidated Mineral	Coarse Unconsolidated	Gravelly	–				Benthic/Attached Biota	Faunal Bed	Soft Sediment Fauna
		Open Water						Gravel Mixes	mG	msG	sG				
							Fine Unconsolidated	Slightly Gravelly	gM	gmS	gS				
								Muddy Sand	(g)M	(g)sM	(g)mS	(g)S			
								Sandy Mud	M	-	-	-			
								Mud							
Robins Island	Estuarine	Coastal	Geologic	Basin	Geologic	Unconsolidated Mineral	Coarse Unconsolidated	Gravelly	–				Benthic/Attached Biota	Faunal Bed	Soft Sediment Fauna
		Open Water		Slope			Fine Unconsolidated	Slightly Gravelly	-	-	-			Reef Biota	Mollusc Reef Biota
								Sand	gM	gmS	-				
								Sandy Mud	–	(g)sM	(g)mS	(g)S			
								Muddy Sand	-	sM	mS	S			
Shelter Island	Estuarine	Coastal	Geologic	Basin	Geologic	Unconsolidated Mineral	Coarse Unconsolidated	Gravelly	–				Benthic/Attached Biota	Faunal Bed	Soft Sediment Fauna
		Open Water		Flat				Gravel Mixes	mG	msG	sG			Reef Biota	Mollusc Reef Biota
							Fine Unconsolidated	Slightly Gravelly	gM	gmS	gS				
									–	–	–	(g)S			
									–	–	–	-			

Estimating surficial percent cover from underwater video

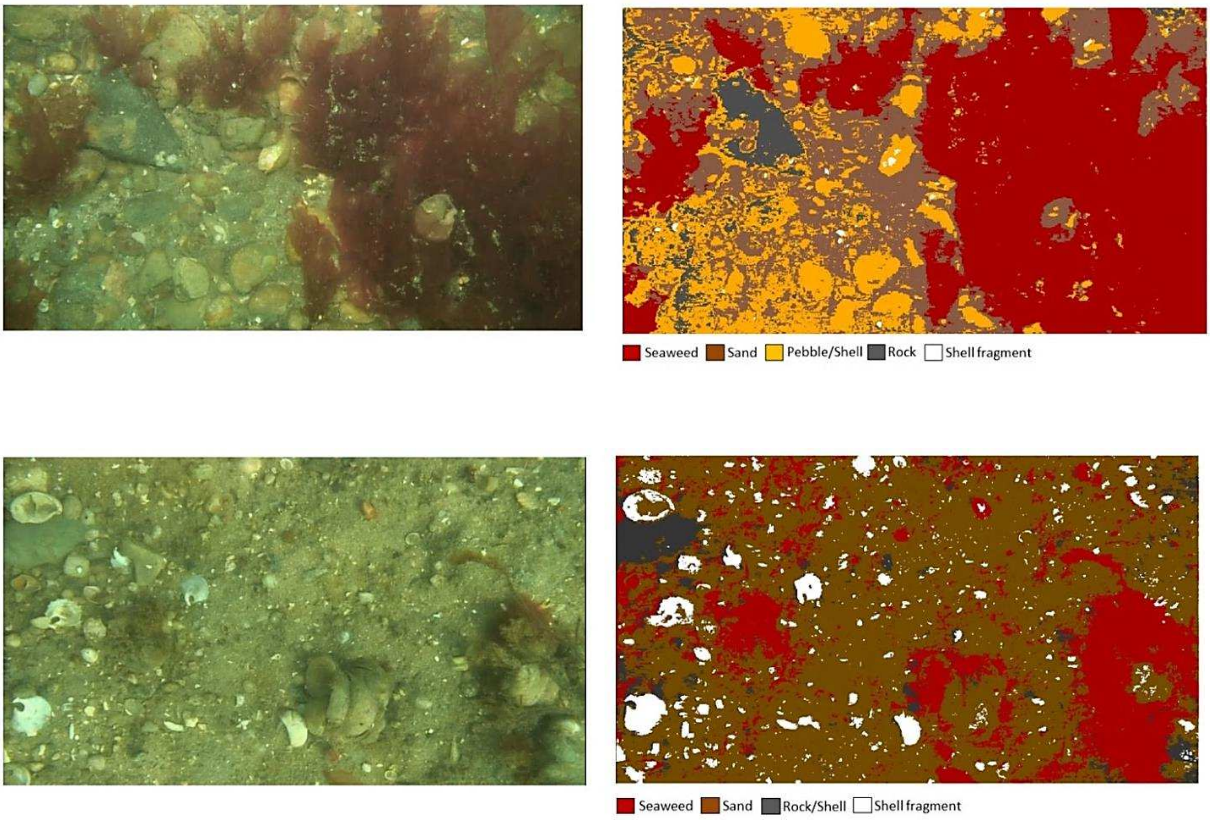


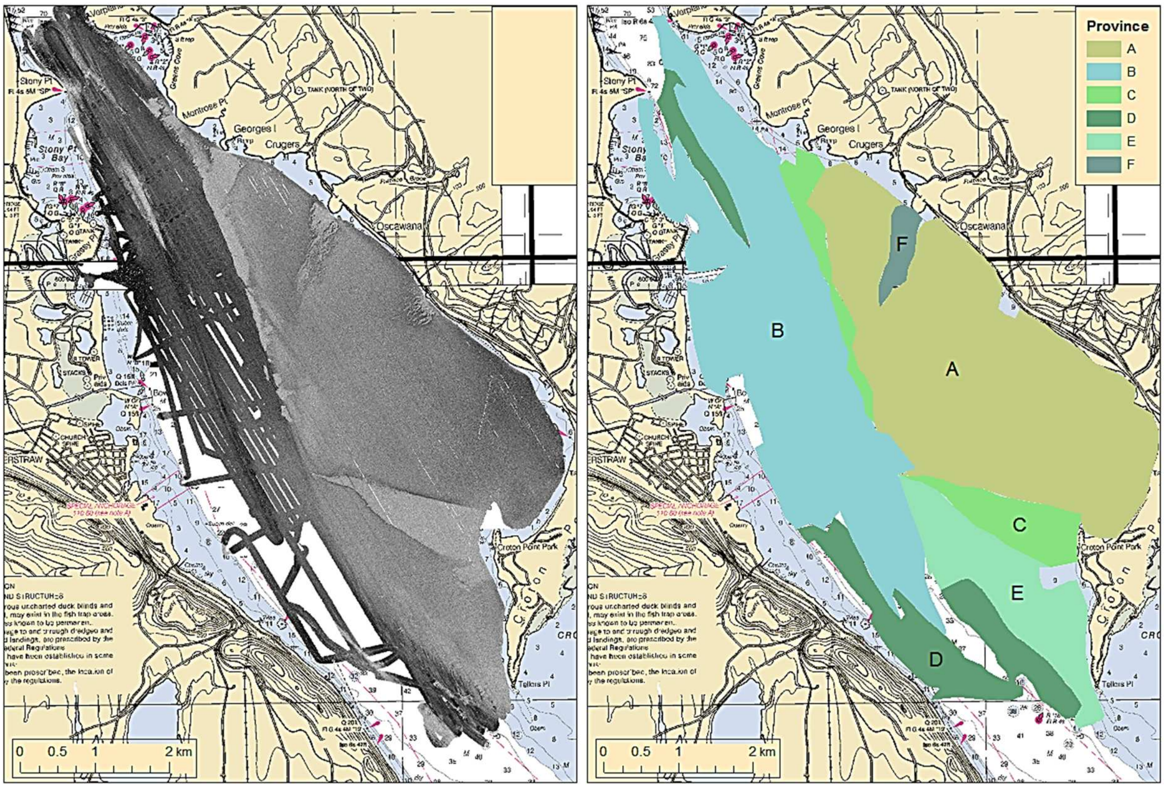
Figure S1. An example of a pebble and seaweed bottom (top) and a sand and shell bottom (bottom) from Shelter Island. The left panels are still images extracted from underwater videos, and the right ones are the results obtained by maximum likelihood classification. These images cover 17.5 x 30 cm portions of the seabed.

Table S3. Percent cover classes identified at each study area using maximum likelihood image analysis.

PERCENT COVER CLASS	ABBREVIATION	PERCENT COVER OF EACH CLASS AT THE STUDY AREAS SURVEYED (MEAN ± 1 SD)			
		TAPPAN ZEE	HUNTINGTON	ROBINS ISLAND	SHELTER ISLAND
Sand	PCSa		1.99 ± 12.19		40.19 ± 43.17
Mud	PCMu	97.60 ± 9.63	83.25 ± 33.37	81.88 ± 37.11	3.17 ± 14.2
Shell Fragment	PCShFg		0.97 ± 2.65	1.65 ± 3.12	1.58 ± 2.91
Shell	PCSh	0.08 ± .61	0.06 ± 0.41	0.26 ± 1.17	7.10 ± 20.68
Rock	PCR				0.67 ± 2.85
Pebble	PCPb		1.29 ± 11.22		1.19 ± 8.18
Seaweed	PCSw		0.04 ± 0.24	0.44 ± 2.35	4.94 ± 8.74
Silty Shell	PCSiSh	0.97 ± 8.36	7.57 ± 22.37	0.66 ± 2.41	4.85 ± 16.01
Shell Pebble	PCShPb		0.35 ± 2.30		10.53 ± 25.53
Muddy Sand	PCMuSa			14.74 ± 33.51	13.5 ± 28.00
Silty Material	PCSiCovered	1.07 ± 3.61			
Anthropogenic	Anthro				0.004 ± .040
Unknown	Unk	0.21 ± 1.16	3.95 ± 19.6	0.01 ± .05	0.09 ± .67
M. prolifera	Mpor		0.04 ± 0.24	0.02 ± .10	0.14 ± .32
Crepidula	Crep		0.38 ± 1.72	0.15 ± .84	11.87 ± 27.28

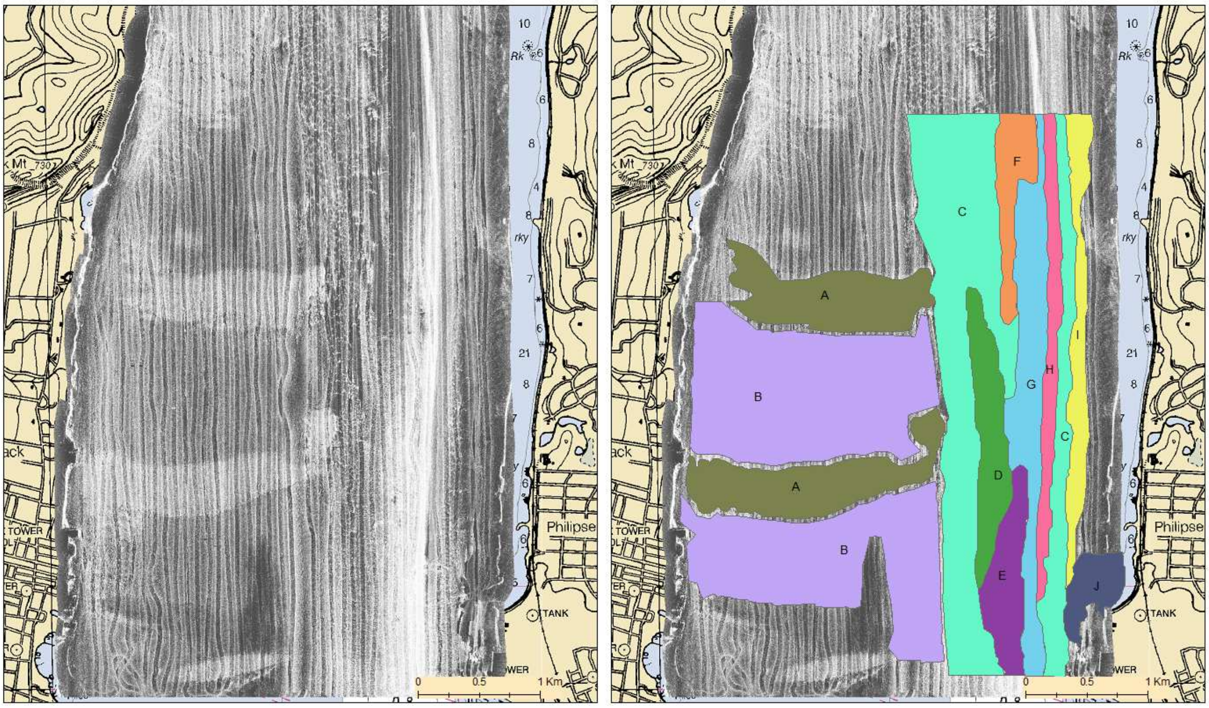
Provinces identified at each study area

<u>Haverstraw Bay</u>



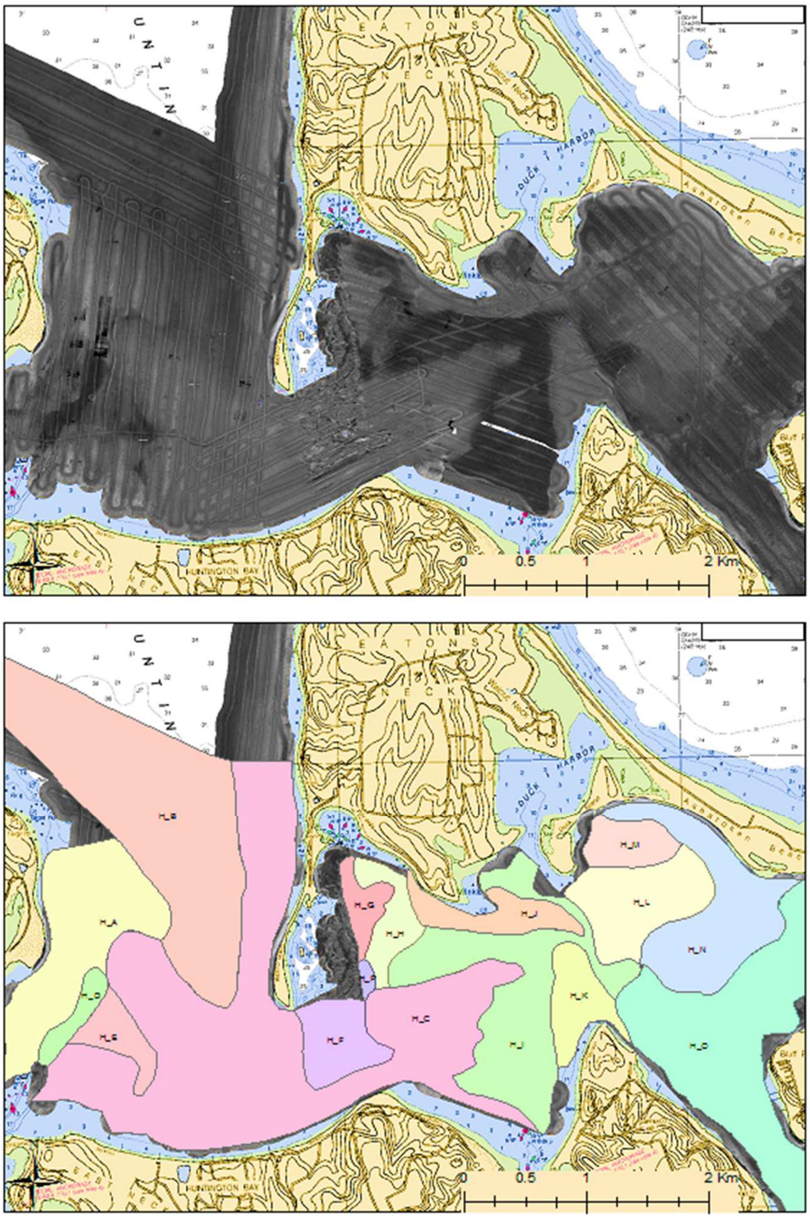
Multibeam backscatter data (left) and visual interpretation of acoustic provinces (right) at the Haverstraw Bay study area in the Hudson River Estuary, NY. Additional acoustic data sets were used to delineate the provinces at this site. Basemap from https://nationalmap.gov.

Tappan Zee



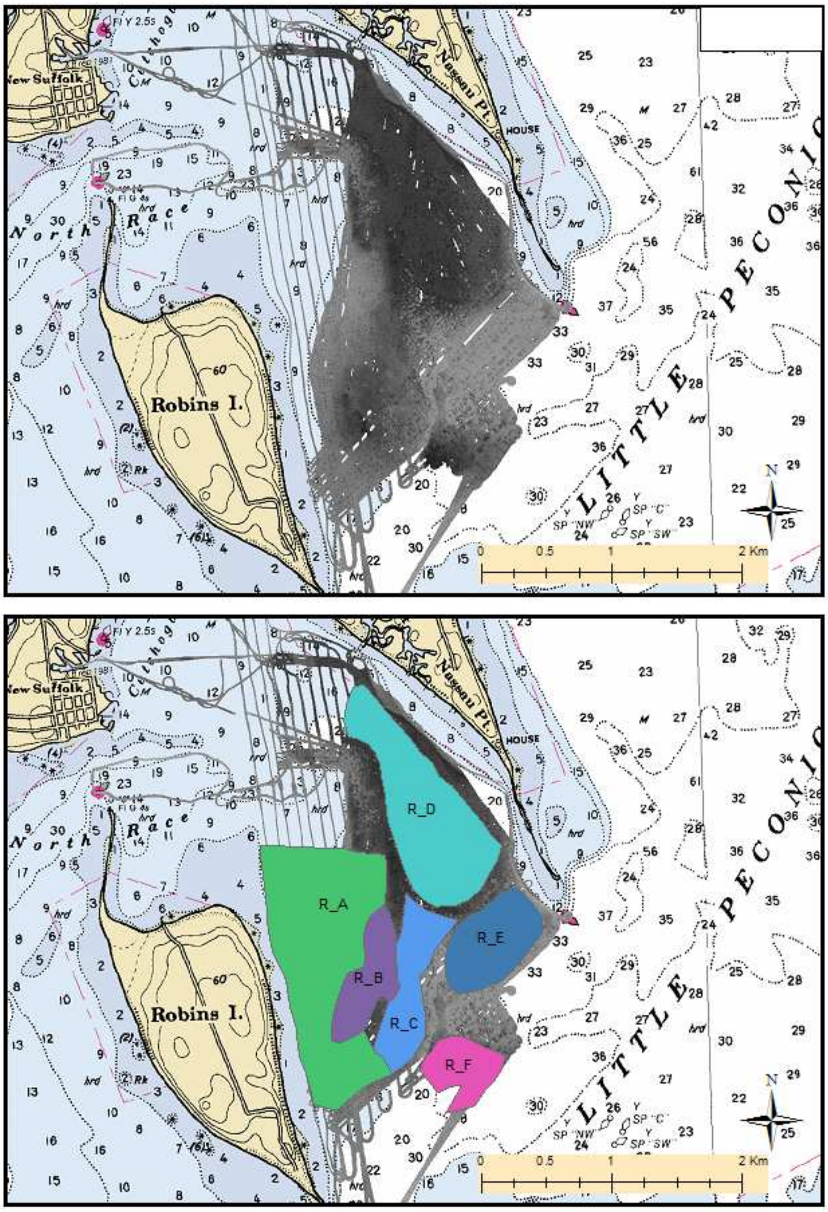
Sidescan sonar backscatter data (left) and visual interpretation of acoustic provinces (right) at the Tappan Zee study area in the Hudson River Estuary, NY. Basemap from https://nationalmap.gov.

Huntington Harbor



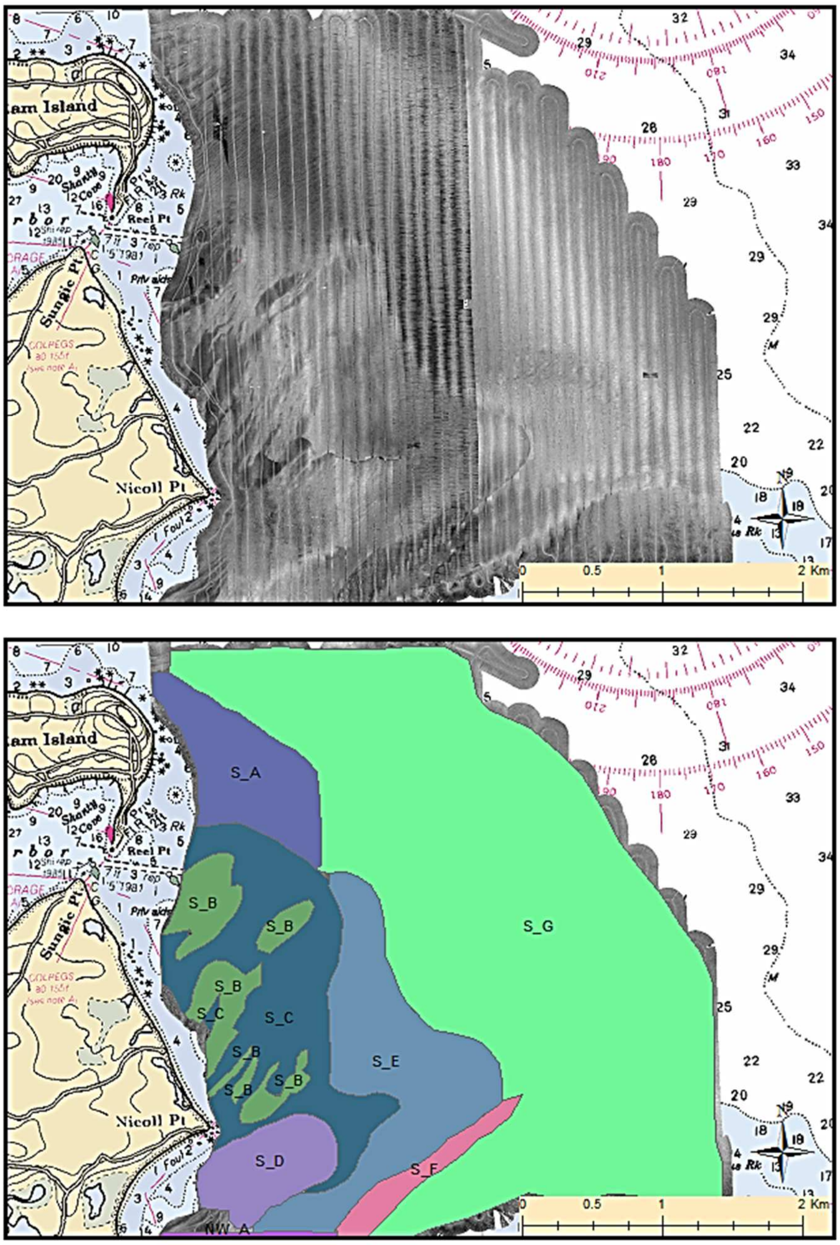
Sidescan sonar backscatter data (top) and visual interpretation of acoustic provinces (bottom) at the Huntington Harbor study area on the north shore of Long Island, NY. Basemap from https://nationalmap.gov.

Robins Island



Multibeam sonar backscatter data (top) and visual interpretation of acoustic provinces (bottom) at Robins Island on the east end of Long Island, NY. Basemap from https://nationalmap.gov.

Shelter Island



Sidescan sonar backscatter data (top) and visual interpretation of acoustic provinces (bottom) at Shelter Island on the east end of Long Island, NY. Basemap from https://nationalmap.gov.

Cross-validation vs. AIC

For MRT, 10-fold cross-validation was applied to each binary split, and splitting continued until the minimum cross-validated error was reached. The tree was then pruned back to the simplest one whose cross-validated error was within one standard error of the minimum cross-validated error (Breiman et al., 1984; Hastie et al., 2001). In RDA, forward selection in Canoco 4.5 was used to identify the variable that explained the largest fraction of faunal variation, and this variable was added to the RDA first. Subsequent explanatory variables were added to the analysis in the order of their explanatory value. Cross-validation was applied to each step of this sequence to identify the minimum cross-validated error. The model was then trimmed as in MRT to the model with the fewest variables whose cross-validated error was within one standard error of the minimum cross-validated error (Breiman et al., 1984). To account for variation due to random data partitioning, the median result from at least 5 cross-validation runs is reported.

Using the minimum cross-validated r^2 instead of a model with the smallest number of parameters within one standard error of the minimum or using Akaike's information criterion (AIC; Akaike, 1973) would have resulted in models with a larger number of environmental variables. With the former, it would have been difficult to justify that the additional model variables added real explained variance. The latter is appropriate and perfectly suitable under normal circumstances in applications where there is a chance of missing some important property and where cross-validation is not implemented. Unfortunately, assessing the performance of explanatory variables across multiple data sets is not possible using AIC since absolute values of this measure have no interpretation (Burnham and Anderson 2002), unlike r^2 whose meaning extends across different data sets. It should also be noted that AIC asymptotically coincides with generalized cross-validation in evaluating subsets of explanatory variables (Golub et al., 1979).

Detailed summary of model results

Table S4. Detailed summary of the single and multiple variable model results for each study area: k refers to the total number of groups in MRT or the total number of variables selected in RDA, r^2 is the coefficient of determination, CV r^2 is cross-validated r^2, and S.E. is the standard error of cross-validated r^2 for each model. All three measures are reported as a percentage of the total variance. The last column lists grouping criteria in MRT or the variables selected in RDA. MRT groups are listed as sets ({ }). Provinces are listed as letters and the grain size categories for the simplified sediment models are listed using the abbreviations from Folk (1954).

Study area		Model	k	r^2	CV r^2	S.E.	MRT grouping criteria & variables selected in RDA
HAVERSTRAW BAY	Single variable	Fixed water depth	2	9.9	2.1	6.7	{shallow} {deep}
		Flexible depth	2	12.9	3.5	6.3	{water depth >= 4.8} {water depth < 4.8}
		Continuous depth	1	10.6	3.2	7.9	continuous water depth
		Fixed sediment	11	35.5	1.1	8.7	All Folk categories
		Simplified sediment	2	14.1	1.4	7.8	{(g)M, (g)mS, (g)sM, gM, gmS, M, sM} {G, mG, msG, sG}
		Flexible sediment	2	15.9	4.8	7.8	{% silt-clay >= 29.6} {% silt-clay < 29.6}
		Continuous sediment	1	14.3	6.3	10.2	% gravel
		Flexible percent cover	NA	NA	NA	NA	no % cover data were collected
		Continuous percent cover	NA	NA	NA	NA	no % cover data were collected
		Fixed geoforms	4	34.6	20.4	7.3	All geoforms
		Simplified geoforms	3	30.9	22.9	6.8	{dredged channel, channel, flat} {mollusc reef}
		Fixed provinces	5	36.8	21.6	7.0	All provinces
		Simplified provinces	3	32.0	24.2	6.7	{AE} {BD} {C}
	Multiple variables	Fixed habitat classes	23	69.0	-8.1	11.3	All 23 LIS-CMECS habitat classes
		Simplified habitat classes	2	19.9	-4.9	10.7	MRT combined 23 LIS-CMECS classes into 2 groups/sets of classes
		Flexible env & fixed geoforms	12	67.4	0.3	10.0	All flexible environmental & geoforms
		Flexible env & simplified geoforms	2	19.2	13.9	7.0	{dredged channel, channel, flat} {mollusc reef}
		Continuous env & fixed geoforms	7	39.6	17.8	8.3	All continuous environmental & geoforms
		Continuous env & simplified geoforms	3	31.0	21.2	6.4	{mollusc reef} {flat} {dredged channel} {channel}
		Flexible env & fixed provinces	13	63.9	7.0	9.5	All flexible environmental & provinces
		Flexible env & simplified provinces	3	32.0	21.3	7.1	{BD} {AE} {C}
		Continuous env & fixed provinces	7	39.2	21.4	7.4	All continuous environmental & provinces
		Continuous env & simplified provinces	3	32.1	24.6	3.5	{C} {AE} {BD}

TAPPAN ZEE	Single variable	Fixed water depth	2	10.1	6.4	4.1	{shallow} {deep}
		Flexible depth	2	13.2	9.9	4.6	{water depth >= 5.8} {water depth < 5.8}
		Continuous depth	1	10.1	6.4	4.1	continuous water depth
		Fixed sediment	10	23.7	6.6	4.3	All Folk categories
		Simplified sediment	2	13.1	6.1	4.5	{(g)M, (g)mS, (g)sM, M, mS, sM} {gM, mG, msG, sG}
		Flexible sediment	2	16.0	9.4	4.7	{% gravel >= 6.7} {% gravel < 6.7}
		Continuous sediment	2	14.6	11.2	4.5	% gravel
		Flexible percent cover	1	0.0	-1.7	3.6	none
		Continuous percent cover	1	3.3	-1.1	4.2	% silt cover
		Fixed geoforms	4	29.8	22.8	4.3	All geoforms
		Simplified geoforms	4	29.8	22.8	4.3	{flat} {mollusc reef} {channel} {wave field}
		Fixed provinces	10	44.9	32.1	4.5	All provinces
		Simplified provinces	5	39.6	30.8	4.5	{A} {BIJ} {CH} {DF} {EG}
	Multiple variables	Fixed habitat classes	26	53.3	11.9	5.8	All 26 LIS-CMECS habitat classes
		Simplified habitat classes	3	30.8	14.0	5.3	MRT combined 26 LIS-CMECS classes into 3 groups/sets of classes
		Flexible env & fixed geoforms	19	66.9	20.5	5.5	All flexible environmental & geoforms
		Flexible env & simplified geoforms	3	29.0	16.6	5.3	{flat} {mollusc reef, channel, wave field & % silt-clay >= 68.4} {mollusc reef, channel, wave field & % silt-clay < 68.4}
		Continuous env & fixed geoforms	14	43.1	-43.0	67.7	All continuous environmental & geoforms
		Continuous env & simplified geoforms	3	33.9	27.8	4.6	{flat} {% silt-clay} {water depth}
		Flexible env & fixed provinces	17	65.6	19.1	6.8	All flexible environmental & provinces
		Flexible env & simplified provinces	4	36.1	22.5	5.0	{BIJ} {ACDEFGH & % silt-clay < 68.4} {% silt-clay >= 68.4 & EH} {% silt-clay >= 68.4 & ACDFG}
		Continuous env & fixed provinces	17	45.3	-46.4	70.6	All continuous environmental & provinces
		Continuous env & simplified provinces	3	35.0	29.3	3.6	{BIJ} {% silt-clay} {DF}

HUNTINGTON HARBOR	Single variable	Fixed water depth	2	6.6	3.1	7.4	{shallow} {deep}
		Flexible depth	6	36.7	11.0	6.7	{water depth < 4.0} {4.0-5.3} {5.3-5.9} {5.9-8.8} {8.8-9.2} {> 9.2}
		Continuous depth	1	13.2	9.9	4.6	continuous water depth
		Fixed sediment	10	29.7	3.0	7.9	All 13 LIS-CMECS habitat classes+I75:I82
		Simplified sediment	2	12.4	3.4	7.1	{(g)M, (g)mS, (g)sM, gM, gmS, M} {gS, mG, msG, sG}
		Flexible sediment	2	12.8	6.9	6.9	{% gravel >= 25.6} {% gravel < 25.6}
		Continuous sediment	2	17.8	10.8	8.2	% gravel
		Flexible percent cover	2	9.9	3.9	6.5	{% mud cover >= 97.2} {% mud cover < 97.2}
		Continuous percent cover	3	11.4	2.2	5.8	{% mud cover} {% unknown cover} {% M. porifera cover}
		Fixed geoforms	1	0.0	-3.2	6.5	All geoforms
		Simplified geoforms	1	0.0	-3.2	6.5	Only 1 Geoform identified at this site: basin
		Fixed provinces	15	58.7	36.7	6.2	All provinces
		Simplified provinces	7	48.9	28.5	7.5	{O} {I} {HLMN} {E} {D} {AC} {FGJK}
	Multiple variables	Fixed habitat classes	13	34.1	2.7	7.5	All 13 LIS-CMECS habitat classes
		Simplified habitat classes	2	13.6	7.4	6.7	MRT combined 13 LIS-CMECS classes into 2 groups/sets of classes
		Flexible env & fixed geoforms	17	64.0	-0.8	9.9	All flexible environmental & geoforms
		Flexible env & simplified geoforms	3	26.0	15.1	7.0	{% gravel >= 25.6} {% gravel < 25.6 & water depth < 6.9} {% gravel < 25.6 & water depth >= 6.9}
		Continuous env & fixed geoforms	16	32.0	-58.3	53.0	All continuous environmental & geoforms
		Continuous env & simplified geoforms	2	24.9	18.9	4.3	{% gravel} {water depth}
		Flexible env & fixed provinces	17	73.9	15.6	8.2	All flexible environmental & provinces
		Flexible env & simplified provinces	6	52.9	26.7	8.5	{O} {AC} {FGJK} {HILMN} {BDE & water depth < 9.4} {BDE and water depth >= 9.4}
		Continuous env & fixed provinces	23	63.4	-31.0	56.6	All continuous environmental & provinces
		Continuous env & simplified provinces	5	44.0	34.5	6.8	{% gravel} {water depth} {O} {AC} {BD}

967

ROBINS ISLAND	Single variable	Fixed water depth	2	8.3	-3.5	10.1	{shallow} {deep}
		Flexible depth	3	25.9	3.7	8.7	{water depth < 6.0} {water depth >= 6.0 & < 12.5} {water depth >= 12.5}
		Continuous depth	1	13.2	9.9	4.6	continuous water depth
		Fixed sediment	8	26.6	-1.4	9.9	All Folk categories
		Simplified sediment	2	14.8	4.6	8.7	{(g)mS, (g)sM, gM, gmS, mS, sM} {(g)S, S}
		Flexible sediment	2	14.8	6.3	8.8	{% sand >= 85.2} {% sand < 85.2}
		Continuous sediment	2	12.7	5.6	7.1	% silt-clay
		Flexible percent cover	1	0.0	3.5	10.0	none
		Continuous percent cover	3	22.8	7.1	8.8	% shell fragment cover, % mud cover, % M. porifera cover
		Fixed geoforms	2	1.9	-5.2	10.0	All geoforms
		Simplified geoforms	1	0.0	-4.4	10.0	{basin & slope}
		Fixed provinces	6	40.4	25.7	8.7	All provinces
		Simplified provinces	4	33.7	21.3	8.5	{A} {BC} {DF} {E}
	Multiple variables	Fixed habitat classes	10	35.6	0.4	8.9	All 10 LIS-CMECS habitat classes
		Simplified habitat classes	2	14.8	3.6	8.8	MRT combined 10 LIS-CMECS classes into 2 groups/sets of classes
		Flexible env & fixed geoforms	17	68.9	3.6	10.0	All flexible environmental & geoforms
		Flexible env & simplified geoforms	2	14.8	6.9	8.8	{% sand >= 85.2} {% sand < 85.2}
		Continuous env & fixed geoforms	16	47.2	-101.5	75.9	All continuous environmental & geoforms
		Continuous env & simplified geoforms	3	29.0	14.9	8.2	{% silt-clay} {% mud cover} {water depth}
		Flexible env & fixed provinces	15	67.8	21.9	8.0	All flexible environmental & provinces
		Flexible env & simplified provinces	4	33.7	19.8	8.5	{A} {BC} {DF} {E}
		Continuous env & fixed provinces	21	60.1	-113.2	76.4	All continuous environmental & provinces
		Continuous env & simplified provinces	4	40.1	25.0	7.6	{A} {BC} {% mud cover} {% M. Porifera cover}

968

SHELTER ISLAND	Single variable	Fixed water depth	2	9.1	4.1	5.4	{shallow} {deep}
		Flexible depth	6	32.1	8.8	6.8	{water depth < 4.7} {4.7-5.7} {5.7-5.9} {5.9-8.0} {8.0-8.6} {>8.6}
		Continuous depth	1	6.6	3.1	7.4	continuous water depth
		Fixed sediment	7	25.9	9.9	5.6	All Folk categories
		Simplified sediment	3	19.1	10.9	5.2	{(g)S} {gS, sG} {gM, gmS, mG, msG}
		Flexible sediment	2	16.4	6.3	5.4	{% silt-clay >= 4.8} {% silt-clay < 4.8}
		Continuous sediment	2	16.7	8.7	4.7	% sand
		Flexible percent cover	2	13.0	6.1	4.9	{% sand cover >= 55.5} {% sand cover < 55.5}
		Continuous percent cover	4	22.0	9.6	5.3	{% shell/pebble cover} {% shell cover} {% sand cover} {% muddy sand cover}
		Fixed geoforms	2	8.9	3.9	5.4	All geoforms
		Simplified geoforms	2	8.9	3.9	5.4	{basin} {flat}
		Fixed provinces	7	40.3	26.0	5.2	All provinces
		Simplified provinces	5	35.8	25.6	5.3	{A} {B} {CEG} {D} {F}
	Multiple variables	Fixed habitat classes	18	47.0	3.0	7.0	All 18 LIS-CMECS habitat classes
		Simplified habitat classes	2	16.6	3.4	5.8	MRT combined 18 LIS-CMECS classes into 2 groups/sets of classes
		Flexible env & fixed geoforms	19	71.8	-1.0	7.4	All flexible environmental & geoforms
		Flexible env & simplified geoforms	2	16.4	5.9	5.6	{% silt-clay >= 4.8} {% silt-clay < 4.8}
		Continuous env & fixed geoforms	19	44.9	-114.1	80.1	All continuous environmental & geoforms
		Continuous env & simplified geoforms	1	12.9	7.1	4.4	% sand
		Flexible env & fixed provinces	20	72.1	6.0	6.8	All flexible environmental & provinces
		Flexible env & simplified provinces	5	37.0	17.9	5.6	{CE} {D} {ABF & % sand < 92.7} {ABF & % sand >= 92.7 and depth >= 5.7} {ABF & % sand >= 92.7 and depth < 5.7}
		Continuous env & fixed provinces	24	56.4	-146.8	78.7	All continuous environmental & provinces
		Continuous env & simplified provinces	4	34.9	25.0	4.4	{% sand} {CEG} {D} {A}