Hydrobiologia (2015) 756:117–137
DOI 10.1007/s10750-014-2139-3

OFFSHORE WIND FARM IMPACTS

Toward wind farm monitoring optimization: assessment of ecological zones from marine landscapes using machine learning algorithms

Annette R. Grilli · Emily J. Shumchenia

Received: 19 March 2014 / Revised: 17 November 2014 / Accepted: 29 November 2014 / Published online: 12 December 2014
© Springer International Publishing Switzerland 2014

Abstract Within the perspective of siting wind farms offshore of Rhode Island, USA, the State and National Environmental Agencies had requested a local marine ecological assessment, which led to an ecological zoning of the area. In view of expanding this zoning outside its limit of the test area and filling gaps in ecological zones, an effort to model those ecological zones using marine landscape or abiotic features was carried out. This study tests the accuracy of selected machine learning algorithmic models, decision tree, and random forest, for relating marine landscapes features to ecological sub-regions. Both models show to be good predictive tools with accuracy after cross validation of the order of 5–3%. Key abiotic variables to provide an accurate model were investigated. The study demonstrates the importance of the distance to coast, the sediment characteristics (fraction of clay, median size of the sediments), the hydrodynamic features, in particular not only tidal current/drag force, but also wave drag force, and finally the oceanographic characteristics such as stratification and sea surface temperature to built a good predictive model. Those findings provide some insight on the pre-monitoring effort optimization.

Keywords Random forest · Decision tree · Marine landscape · Ecological zoning · Offshore wind farm siting optimization · Cluster analysis

Guest editors: Steven Degraer, Jennifer Dannheim, Andrew B. Gill, Han Lindeboom, and Dan Wilhelmsson/ Environmental impacts of offshore wind farms

Electronic supplementary material The online version of this article (doi:10.1007/s10750-014-2139-3) contains supplementary material, which is available to authorized users.

A. R. Grilli (✉)
Department of Ocean Engineering, University of Rhode Island, Kingston, USA
e-mail: agrilli@egr.uri.edu

E. J. Shumchenia
Graduate School of Oceanography, University of Rhode Island, Kingston, USA

Introduction

Background

Within the perspective of siting wind farms offshore of Rhode Island (RI), USA, the State and National Environmental Agencies, the RI coastal resource management council (CRMC), and the bureau of ocean energy management (BOEM), respectively, have requested that a regional marine ecological assessment be carried out. In response to this request, the development of a RI ocean special area management plan (RIOSAMP) was initiated in 2008, leading to extensive oceanographic and ecological data campaigns and analyses (SAMP, 2010). On this basis, a wind farm siting protocol was developed, including an

optimization tool between the wind resource and technical, societal, and ecological constraints, addressing both macro- and micro-siting issues (Spaulding et al., 2010; Grilli et al., 2013; O'Reilly et al., 2013). While the macro-siting tool seeks to optimize wind farm location within the region of interest, the micro-siting tool optimizes the relative position of individual turbines within the area predefined at the macro-siting stage. Both macro- and micro-siting stages are distinct in their scale of interest and in the issues and processes involved. The ecological cost issue of offshore wind farm development was raised at the macro-siting stage, which led to establishing ecological sub-regions.

The RIOSAMP project included a full ecological assessment for the area of interest, leading to extensive fish and mammal data bases (Bohaboy et al., 2010; Kenney & Vigness-Raposa, 2010; Malek et al., 2010) and providing the initial data necessary for zoning the RIOSAMP area in a species sensitivity to wind farm impacts perspective. Seasonal *ecological typologies* were established based on species relative abundance and richness, providing a marine zoning based on homogeneous species assemblages corresponding to geographical patterns, labeled as *ecological sub-regions* (Grilli et al., 2013). More recently, we developed a *marine landscape* typology representing a zoning based on abiotic variables, contributing to establishing the initial ecological structural framework describing the area (Shumchenia & Grilli, 2012). Both typology methods were based on a combination of principal component and cluster analyses. Ecological typologies based on species assemblage similarities were developed using a similar methodology, in particular, for the ecological zoning of estuaries (Buddemeier et al., 2008), or for zoning based on marine pelagic assemblages (Jordaan, 2010). Marine landscape typologies based on similar methods have been established in other seas, in particular, in the North Sea with the work of Verfaillie et al. (2009).

Abiotic and biotic patterns

In our initial typologies, we observed similarities between biotic and abiotic patterns. However, despite both sets of variables being clearly functionally related based on ecological theory (Austin, 2002; Connor et al., 2003), the relationship between both patterns could not be quantified using "*classical*" inferential statistical methods (e.g., multivariable nonlinear regressions or discriminant analysis). By "*classical*" statistics, we refer to the *data modeling* culture, versus the *algorithmic modeling* culture, as discussed by Breiman (2001). Both cultures are fundamentally opposite in their conceptual approach to natural processes. The former assumes a stochastic data model describing the functional processes relating patterns, while the latter considers these functional processes as being complex and unknown. The latter approach led to the development of *learning machine* algorithms, including neural networks (e.g., Bishop, 1995), support vector machines (e.g., Vapnik, 1998; Hamel, 2009), boosting (e.g., Schapire, 2013), and decision tree (DT) and random forest (RF) algorithms (e.g., Breiman, 2001; Biau, 2012).

The limitations of "*classical*" inferential statistical methods for relating biotic and abiotic variables result from many entwined reasons, which ultimately fall under the overarching issues of having multi-spatial and temporal scales, as well as many sources of nonlinearity, in the involved processes. Indeed, applying standard statistics requires making several assumptions about the variables, in particular, that they are *well* represented by a theoretical distribution. This may not be the case when considering a fixed scale, such as the sub-regional scale in the present case. In addition, the processes relating biotic and abiotic variables are often complex and expressed by nonlinear relationships that are not well captured by multivariate correlations. In particular, the data modeling approach has difficulties capturing three key ecological phenomena: (1) the multi-scale aspect of the processes involved (Cowen et al., 2006); (2) the fragmentation (Fahrig, 2003); and (3) the edge effect (Ewers & Didham, 2006).

Fragmentation and edge effect indeed create nonlinearity and spatial discontinuities in species spatial distributions, which alter the statistical distribution of the species, in particular, by including an often unacceptable number of zeros (i.e., no occurrence); in parallel, the relationship between abiotic and biotic variables might be complex and difficult, if not impossible to grasp, with standard statistics, in particular, if the relationship between both datasets is not clearly established at each sampling site, but is shifted in space. Methods based on machine learning algorithms, on the other hand, do not require any well-behaved distribution; thus, a niche with few

occurrences of a rare species can be detected. Additionally, these methods detect dissimilarities between sites and do not require occurrences to be statistically distributed in a particular way. By multi-scale, we refer to the superimposition of biomes, in particular, to the question of migratory species traveling across established biomes. Similarly, migratory species create spatial discontinuities superimposed on a structured primary ecosystem. Mathematically, again, this will create zeros in the spatial distribution, undesirable in classical statistics, but perfectly acceptable in machine learning algorithms.

While the use of classical statistics to model the spatial distribution of benthic communities from abiotic variables can be successful in specific environments, such as the North Sea (Degraer et al., 2008), it is often limited to being a good tool for extracting causality factors. As a predictive tool of spatial distributions, classical statistics is often poor, especially in complex environments such has tropical reefs (Richmond & Stevens, 2014), or for modeling pelagic communities (Malcolm et al., 2011). These limitations motivated the use of either a combination of data and ecological models (Guisan & Thuiller, 2005; Guisan & Rahbek, 2011), or algorithmic models, for habitat mapping (Bahn & McGill, 2013). The term 'conceptual model' is used to include both the ecological and data models (Austin, 2002). Algorithmic models have been successfully used for about a decade to predict continental biotic environments at the community or species levels (e.g., Elith et al., 2006; Phillips et al., 2006). Cutler et al. (2007) demonstrated the accuracy of the RF algorithm as a classification tool in ecology. Drake et al. (2006) modeled ecological niches using support vector machines. The use of learning machine algorithms in marine environments for assessing marine communities from abiotic factors has similarly emerged as a useful predictive tool. De'Ath & Fabricius (2000) showed the superiority of regression trees to linear models, for assessing relationships between physical variables, location, and soft coral taxa, in the Great Barrier Reef. Wiley et al. (2003) used the "Genetic Algorithm for Rules and Prediction" toolbox (GARP; Stockwell, 1999) in the marine environment, and demonstrated its ability to predict geographic distributions of fish across a heterogeneous ocean region. Rinne et al. (2014) used a RF algorithm (Breiman, 2001) and the "Maximum Enthropy" toolbox (MAXENT; Phillips et al., 2006) to predict the occurrence of rocky reefs from abiotic variables, and estimate fish species spatial distributions. Pesch et al. (2011) used a "Classification and Regression Tree" method (CART; Breiman et al., 1984) to predict eco-regions in the North Sea, based on benthic organisms as well as on grain size, temperature, salinity, nutrients, and bathymetry.

Study motivation

Marine landscape abiotic variables are generally available on relatively fine spatial and temporal resolutions, while the spatial and temporal resolution of species abundance is often too coarse or irregular, and becomes questionable. A reliable model relating abiotic features and ecological zones would thus provide an accurate method for either filling data gaps or extrapolating data outside the limits of the test site (within a similar ecological region). Additionally, the identification of key abiotic variables, sufficient and necessary to predict specific ecological zones, would allow both optimizing and cutting the cost of environmental monitoring campaigns, which is particularly critical in the context of offshore wind development over large areas.

Study objectives

The main objective of this study is to test the accuracy of selected machine learning algorithm models for relating marine landscapes to ecological zones. Two models are presented, based on DT and RF algorithms, and their results are discussed in the "Data and methods" section. As a second and parallel objective, we aim at identifying the key abiotic variables sufficient and necessary to provide an accurate model.

The proposed models are seasonal; they were developed for fall and spring seasons. In the present paper, we will only present results for the spring season. The proposed methods and algorithms were applied and validated for the RIOSAMP area, in the Atlantic Ocean, offshore of Rhode Island, USA.

Terminology

In the hierarchical framework of Zacharias & Roff (2000) (modified from Noss, 1990), at the sub-regional scale, biotic entities are referred to as *communities* and abiotic entities as *ecosystems*. A community represents a species assemblage; from an abiotic

perspective, ecosystem features (i.e., abiotic variables) describe habitats or habitat types at any scale (Roff & Taylor, 2000). More recently, the "landscape" terminology initially used by Noss for continental ecosystems has been made specific to the marine environment, with *marine landscape* referring to a relatively broad scale ecosystem described by abiotic features, rather than *habitat* (Van Lancker & Foster-Smith, 2007; Verfaillie et al., 2009). In this paper, we use *marine landscape* to define abiotic subregional entities, and *ecological sub-regions* to refer to the geographical patterns defined by our biotic communities (Fig. 1). For sake of brevity, we will also refer to these patterns as *zones.*

Data and methods

The approach developed in this study is summarized in Fig. 2. Two independent typologies were preliminary established for biotic and abiotic variables, providing geographic patterns: the biotic zones, or ecological subregions, corresponding to specific species assemblage are characterized by a sensitivity index to wind farm impact (Grilli et al., 2013); and the abiotic zones, or marine landscapes. The learning machine models are used to relate a given site to its corresponding ecological subregions based on measured abiotic variables and, correlatively, to its wind farm sensitivity zone, in relation to a representative assemblage. Two types of models are developed based either on DT or RF algorithms.

Study area

The selected study area is that of the RIOSAMP (Fig. 3), which is bounded to the North by the Rhode Island coast, to the west by the Connecticut and New York state borders (Long Island Sound), to the east by the Massachusetts state border (Vineyard Sound), and to the south, 50 miles offshore, by an open ocean boundary. This area includes Block Island Sound and Rhode Island Sound, west and east of Block Island, respectively, and has a bathymetry that is gradually deepening with increasing distance from shore, reaching depths of 60 m at the open ocean boundary. Two deeper areas, between 50 and 60 m deep, intrude in Block Island and Rhode Island Sounds, altering this monotonic decrease in depth. Quaternary moraine deposits result in relatively shallower depth shoals being located in some deeper areas, such as southwest of Block Island on the Southwest Ledge (on the order of 10 m depth within a 25 m deep area), as well as further offshore in the Rhode Island Sound, on the Cox Ledge (on the order of 30 m depth within a 45 m deep area). Indeed, the geology of the area is characterized by the late quaternary advance, pulsing, and retreat, of the Wisconsin Laurentide ice sheet over cretaceous sedimentary layers consisting of a lower layer of semi-consolidated sandstone and very compact clay and silt, and an upper unconsolidated layer, which is a combination of inter-bedded quartz, sand, clay, and lignite (coastal plain strata). The quaternary layer, however, can be moraine, glaciofluvial, or glaciolacustrine sediment, or simply holocene marine deposits, which reflects in the bathymetry (Needell & Lewis, 1984; Stone & Sirkin, 1996; McMullen et al., 2008). Geological maps, as well as interpreted geological maps in terms of (wind farm) *construction effort,* were given in LaFrance et al. (2010) and Spaulding et al. (2010), respectively. The oceanographic characteristics of the area, such as salinity and temperature profiles, as well as currents, are described in details in Codiga & Ullman (2010) and Grilli et al. (2010).

Typologies

Marine Landscape and ecological typologies were established using a combination of principal component

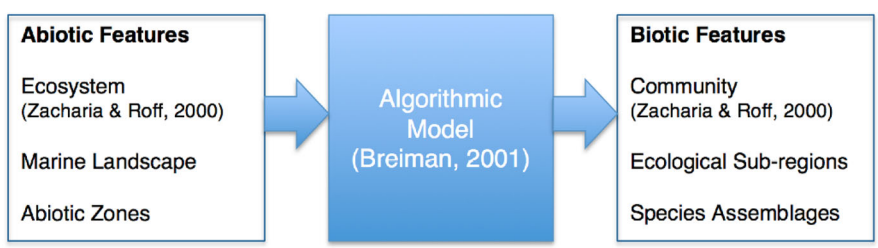

Fig. 1 Terminology used in this study to define structural and geographical biotic and abiotic entities, at the ecosystem and community levels

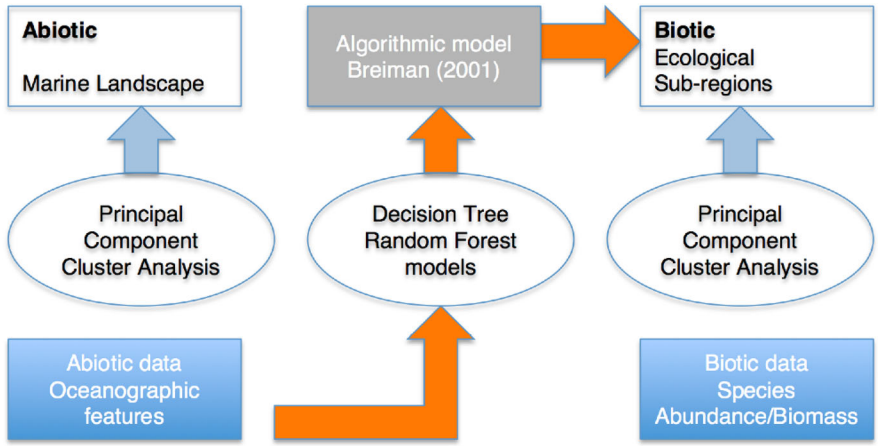

Fig. 2 Schematic of the structural ecological assessment. The approach used in the present analysis, which allocates an ecological zone label to a site defined by its abiotic components, is shown in *orange*

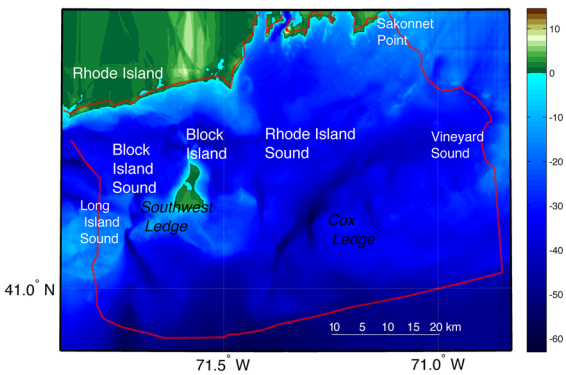

Fig. 3 Test site: the Rhode Island Ocean special area management plan (RIOSAMP) area (*red solid line*). Bathymetry is in background in meters. (Color figure online)

and k-means cluster analysis. A 200-m grid was selected for the analysis, a resolution recognized as fine enough to identify regional patterns in marine landscapes (Verfaillie et al., 2009) and biotic communities (Derous et al., 2007). The purpose of the multivariate spatial analysis is to regroup similar grid cells based on their descriptor or feature values into homogeneous zones or regions (we have the values of 17 and 21 features at each grid cell, for the marine landscape and the ecological typologies, respectively). The principal component analysis (PCA) is used to simplify the grouping of cells occurring in a large multi-dimensional space, since PCA reduces the multivariate-space dimension while keeping most of the information using fewer variables, referred to as the principal components (Zuur et al., 2007). The cluster analysis regroups similar cells in the principal component space; the k-mean clustering method was used in the present analyses. This method yields a set of clusters, as compact and well separated as possible, with each cluster reflecting, here, either a specific ecological region or a specific marine landscape.

Ecological regions

Our ecological database is composed of values of biomass for 16 fish species (Bohaboy et al., 2010) and abundance for three mammal group: whales, dolphins, and porpoises (Kenney & Vigness-Raposa, 2010), interpolated on the 200 m grid. Data were provided seasonally; data sources and a list of species included in the analysis are provided in the Appendix. A discussion on trawl types and methods used to convert data into biomass per unit area is given in the RIOSAMP relevant reports (Kenney & Vigness-Raposa, 2010; Bohaboy et al., 2010). The ecological typology resulting from the principal component and cluster analyses was presented in Grilli et al. (2013), of which we summarize the principal findings in the following.

The geographical zoning resulting from the ecological typology is shown for the spring season in Fig. 4. The representative assemblages of each subregion are schematized in Fig. 5. In this schematic view, species are grouped, and their relative frequency is compared to their general distribution in the whole SAMP area, on a scale of 0–3 (a score above 1 indicates a mean abundance of the group in the cluster higher than the median abundance in the SAMP area; a score above two indicates that the mean abundance of the group in the cluster is higher than 75% of the

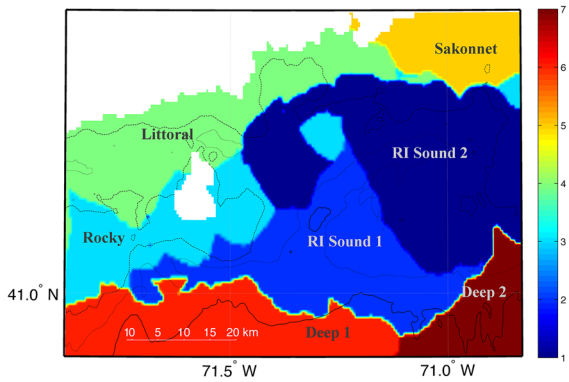

Fig. 4 Ecological sub-regions resulting from the cluster analysis (spring season); assemblages representative of each cluster are described in Figure (based on Grilli et al., 2013); *Deep2* and *Sakonnet* were omitted from the analysis since they were outside the RIOSAMP area. *Contour lines* indicate bathymetry: 30 m (*dashed black*); 40 m (*dashed-dot gray*); 50 m (*solid black*)

abundance in the SAMP area). Groups of species are based on French-McCay et al. (2011), as labeled on Fig. 5 [Lobster, Squid, Demersal Fish (including Flat fishes), Baitfish, Herring, Skates, Medium and Large Gamefish, Whale, Dolphins, and Porpoise].

The five spring sub-regions can be interpreted as a result of the distance to coast (littoral versus deep water) and the dichotomy of the Block Island and Rhode Island sound areas, on the west and east of Block Island, respectively. The two sounds are opposite from an oceanographic perspective. Block Island sound receives a significant fresh water input from Long Island and shows relatively high tidal currents in the west part, creating a colder, fresher, and relatively more mixed, water column than the stratified and warmer Rhode Island Sound waters. The latter sound is crossed by the new England current, which enters the area from the east of Rhode Island sound and leaves it south of Block

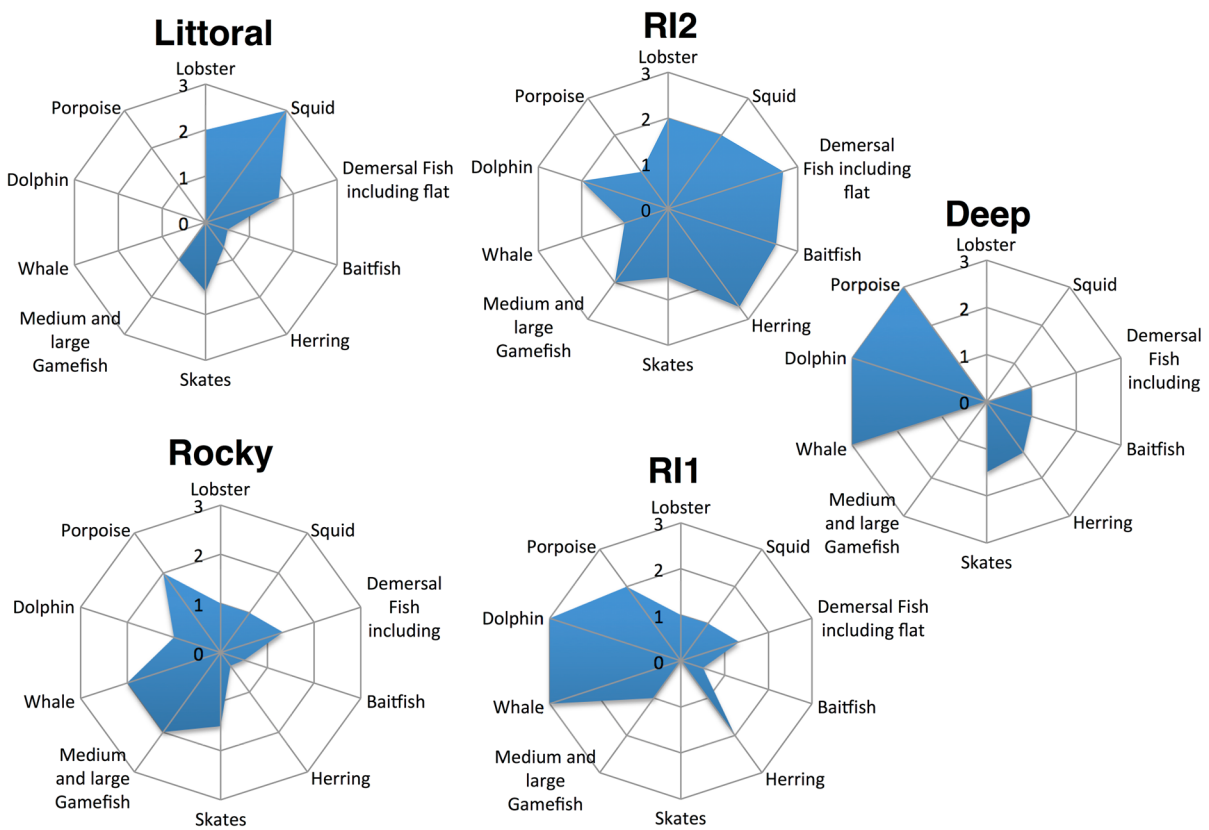

Fig. 5 Schematic representation of the ecological assemblages in each cluster of Fig. 4. Species are grouped and their relative frequency is compared to the general distribution in the entire RIOSAMP area, scaled between 0 and 3 (a score above 1 indicates a mean abundance of the group in the cluster higher than the median abundance in the RIOSAMP area; a score above 2 indicates that the mean abundance of the group in the cluster is higher than 75% of the abundance in the RIOSAMP area)

Island. The terminal moraine crossing the sound also creates alternative biotopes, altering the expected North/South–East/West pattern. These oceanographic factors are expressed through the abiotic variables selected to establish the marine landscapes, as discussed in the next section.

Keeping in mind that a score of two reflects a mean biomass higher than the average (in the SAMP area), the *Littoral* zone is mostly characterized by no mammals, a higher than average abundance of lobsters and squids, and a lower than average abundance of herrings and baitfish. Further offshore, east of Block Island, the *Rhode Island Sound 2* zone (RI2) shows the highest richness and biodiversity among all zones, regrouping more demersal, bait and herring, species, and biomass, than anywhere else in the SAMP area. On the western side, the rocky zone is defined by a lower abundance of most fish species than average, except for larger medium game fish; mammals also start showing their presence. Further offshore, the *Rhode Island Sound 1* zone (RI1) and the *Deep* zone (DEEP) are mostly characterized by a higher abundance of mammals than anywhere else. RI1 shows a high abundance not only of whales and dolphins, but also of migratory herrings; lobsters are still present. In the DEEP zone, whales and dolphins, but also porpoises, are found in great abundance; however, neither lobsters nor herrings are present. RI1 thus seems to define the southern limit of the Gulf of Main herring stock migratory loop, which is in agreement with earlier work (Sindermann, 1979; Reid et al., 1999).

Marine landscapes

The choice of variables for defining marine landscapes is inspired from the Darwinian ecological concept, which became popular in the 1970s and still inspires many habitat mapping efforts; it states that habitat characteristics are related to particular life-history strategies (Southwood, 1988). Under the life-history theory, habitat characteristics impose selective forces through abiotic factors, which affect the fitness of individuals, by modifying their growth rate, fecundity, and survivorship (Southwood, 1988; Connor et al., 2003; Kostylev & Hannah, 2007; Verfaillie et al., 2009). In this conceptual approach, characteristic variables can be regrouped into variables stimulating the scope for growth, or representing disturbances, which means, physical forces, agents, or processes, causing perturbations in an ecological system (Rykiel, 1985).

Our abiotic database includes 15 variables, which are listed in Table 1 together with data sources. This set is relatively similar to that developed in Verfaillie et al.'s (2009) protocol. It, however, differs by the inclusion of sea surface temperature, stratification, and tidal and wave drag forces, and by not including chlorophyll concentration, which results in a marine landscape typology based on abiotic variables only. Let us note that this abiotic database is also significantly expanded compared to that used in our earlier work (Grilli et al., 2013), which was restricted to six variables: water depth, sea floor slope and roughness, sediment median grain size, sea surface temperature, and water column density stratification.

Since most variables are standard and have been used in other studies (e.g., Verfaillie et al., 2009), we limit their description to the information provided in Table 1. More details on each variable are provided in Shumchenia & Grilli (2012), with the exception of the two drag forces, which were recently substituted for the initially used bottom current velocities, to better reflect the *disturbance*, or force, experienced by individuals in the water column or biotope near the seafloor. We provide below a slightly more detailed description of these two new hydrodynamic variables, as well as of the Bathymetric Position Index.

Tidal current/force The bottom current velocity was modeled on a 300-m grid in the SAMP area, for 7 days, using the regional ocean modeling system model (Shchepetkin & McWilliams, 2005; ROMS, 2009; Grilli et al., 2010). Tidal forcing along the model domain boundary was based on tidal constituents obtained using Oregon State University's Tidal Inversion Software (Egbert, 1997; OTIS, 2009). The maximum horizontal velocity was selected as a proxy for the tidal velocity near the seafloor variable, v_t. The tidal drag force, F_t, is computed proportionally to the square of the tidal velocity,

$$F_t = \rho C_d v_t^2.$$

Assuming the density, ρ, and the drag coefficient C_d are constant in the area, our proxy variable for tidal disturbance, T, can simply be reduced to,

$$T = v_t^2.$$

Wave current/force Within the realm of linear wave theory, extreme waves entering the SAMP area are

Table 1 List of variables used as descriptors in the typologies or as "features" to grow decision trees (with ND referring to non dimensional)

Variable name	Representative variable definition	Unit	Initial resolution	Data source
Tidal drag force	Square of maximum tidal velocity (density and drag coefficient assumed constant)	N	500 m	ROMS modeling (Grilli et al., 2010)
Wave drag force	Square of 95% Significant wave height in a 50-year storm event [density and drag coefficient assumed constant]	N	10 m	STWAVE modeling (Schumchenia & Grilli 2012)
Depth	Water depth	m	30 m	NOAA coastal relief model
Distance to shore	Distance from each grid cell to closest point to shore	km		(Spaulding et al., 2010)
Slope	Maximum slope between two grid cells (200 m apart)	Deg.	30 m	NGDC coastal relief model; SURFER toolbox
Roughness	Slope standard deviation in 1,000 m radius	ND	30 m	NGDC coastal relief model (LaFrance et al., 2010)
Phi median	Sediment median diameter (on a internal friction angle scale: $\Phi = -\log_2 D_{\mathrm{mm}}$)	Φ	Point data	SEABED: Atlantic coast offshore surficial sediment data. US Geological Survey (Reid et al., 2005)
Clay	Fraction of clay in sediment	%	Point data	SEABED: Atlantic coast USGS offshore surficial sediment data. (Reid et al., 2005)
SST	Mean seasonal sea surface temperature mean seasonal value (Spring/Fall)	Deg. C	0.25–2.5 km	Satellite data NASA Terra and Aqua (MODIS sensors) (Codiga & Ullman, 2010)
Stratification	Buoyancy frequency squared mean seasonal value (Spring/Fall)	s^{-2}	0.25–2.5 km	FVCOM model (Codiga & Ullman, 2010; Shchepetkin & McWilliams, 2005)
Aspect ratio	Slope directionality	Deg. (0–360)	30 m	NGDC Coastal Relief Model and GIS (Schumenia & Grilli, 2012)
BPI fine scale	Bathymetric position index fine scale	ND	30 m	NGDC Coastal Relief Model and GIS (Schumenia & Grilli, 2012)
BPI large Scale	Bathymetric position index large scale	ND	30 m	NGDC coastal relief model and GIS (Schumenia & Grilli, 2012)
North-ness	North–South component in slope	ND	30 m	NGDC Coastal Relief Model and GIS (Schumenia & Grilli, 2012)
East-ness	West–East component in slope	ND	30 m	NGDC coastal relief model and GIS (Schumenia & Grilli, 2012)

long waves, which induce a nearly uniform current over depth, proportional to the significant wave height, H_s, and inversely proportional to the square root of the water depth, h (Dean & Dalrymple, 1984). The magnitude of this current can be expressed:

$$U_w = A\sqrt{\frac{g}{h}},$$

where A is the wave root mean square amplitude, related to the significant wave height H_s as, $A = \frac{H_s}{2\sqrt{2}}$. Using a similar reasoning as for the tidal disturbance, our proxy wave disturbance variable, W, can simply be expressed as proportional to the square of the significant wave height and inversely proportional to the water depth:

$$W = \frac{H_s^2}{h}.$$

The significant wave height was modeled on a 10-m resolution grid over the SAMP area, using the Steady-State Wave (STWAVE) model from the U.S. Army Corps of Engineers, which is a steady-state spectral linear wave model (Smith et al., 2001). Wave height can be used as a proxy variable for wave disturbance,

based on the upper limit of its 95% confidence interval for extreme wave events with a 50-year return period.

Bathymetric position index (BPI) The BPI expresses the relative position of a site compared to its surroundings. A negative BPI indicates a trough in the bathymetry, whereas a positive BPI indicates a ridge in the bathymetry, and a BPI value of zero indicates a flat area (Lundblad et al., 2006). The fine-scale BPI estimates the index within a radius of 800 m, while the large-scale BPI estimates the index within a 4,000 m radius.

A typology of marine landscapes was performed by applying the principal component/cluster analyses approach to the variables in Table 1. The resulting cluster map is shown in Fig. 6, and each cluster, or specific marine landscape, is described in Table 2. The clusters, again, express the dichotomy between the two sounds: Block Island (Clusters 3 and 7) and Rhode Island (clusters 9 and 5), characterized by opposite hydrodynamic and oceanographic characteristics, as already indicated before. The glacial geology subdivides these two eastern and western zones along a SW-NE direction, in clusters 3, 8, and 9. Cluster 2 represents the very narrow littoral zones and cluster 4 the very rough areas, in particular, to the west of Block Island sound. Clusters 1 and 6 represent the two deep water areas over medium and fine sand, respectively. Further details of this typology can be found in Shumchenia and Grilli (2012).

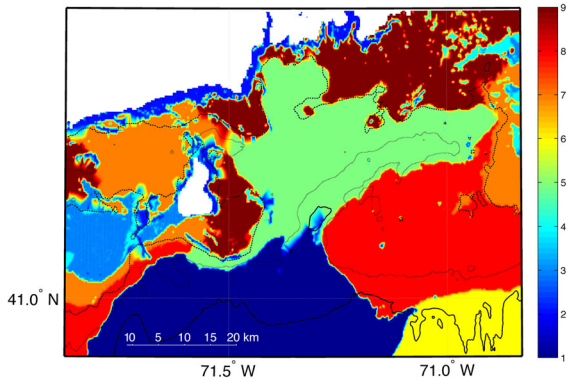

Fig. 6 Typology of marine landscape and habitat in RIOSAMP area, based on a cluster analysis. *Color scale* indicates cluster numbers. *Contour lines* indicate bathymetry: 30 m (*dashed black*); 40 m (*dashed-dot gray*); 50 m (*solid black*). (Color figure online)

Table 2 Clustering of marine landscapes or abiotic zones, offshore Rhode Island, USA (based on variables in Table 1)

Clusters	Descriptions (see Table 1 for variables definition)
1: Deep water zone 1	Stratified warmer water, in deeper offshore area on medium sand floor
2: Littoral zone	Shallow coastal waters
3: Block Island sound zone 2	High dynamic on rocky areas, fresh water input. Higher roughness and BPI index than C7; coldest water; highest bottom velocity
4: Ridge	Highest roughness and BPI index area of high bottom velocity and coldest temperature; fresh water inflow in shallow water and complex geomorphology (ridge), fine sand and clay
5: Rhode Island sound zone 2	Stratified warmer water in intermediate/deep water on similar fine sand as C9, but further away from shore
6: Deep water zone 2	Similar Stratified offshore water in the deepest area, on fine sand and clay
7: Block Island sound zone 1.	High dynamic on smooth geomorphology, fresh water input
8: Offshore shallow water (moraine)	Offshore area in relatively shallower waters; medium sand smooth geomorphology, no clay
9: Rhode Island sound zone 1	Close to shore; fine sand and clay; warmer than C7–C4

The main objective of this paper, however, is not to analyze these typologies, but rather to relate both "worlds": the abiotic marine landscapes and the ecological zones. Although the two typologies definitely have some commonalities in their patterns, they are not superimposable. Any attempt at standard correlations between both sets of patterns would show poor results and yet the two sets are related. This anticipated relationship has motivated the development and application of the proposed machine learning models, to relate the two datasets. These are discussed next.

Learning machine algorithms

Two machine learning models are proposed to predict ecological zones based on values of abiotic variables describing marine landscapes (Table 1): (i) a decision tree (DT) model; and (ii) a random forest (RF) model.

The random forest, as its name indicates, is simply based on an ensemble of random decision trees. Both models, therefore, address the same task of inferring, as a minimum, one decision tree from the data. Although the RF model is designed to better perform than a single decision tree (this is presented in more detail in the following subsections), it loses the information provided by a single tree on the most determinant variables in the classification (Breiman, 2001). The DT model therefore has advantages of simplicity and transparency (Breiman et al., 1984). For this reason, we elected to use both methods and compare their results, which will allow primarily to cross validate the two methodologies, and secondarily to compare the methods' accuracy.

Both models used here are *classifiers:* machine learning algorithms trained to recognize specific categories or classes, here, ecological zones. It should be noted that we also developed, in parallel, *regression* machine learning algorithms, trained to recognize species densities. Both of these algorithm types are adequate, and successful, for predicting zones and species density; however, for sake of conciseness, results presented here are limited to zoning predictions, using the *classifiers* machine learning algorithms.

Decision tree algorithms

The underlying strategy of growing a decision tree is non-incremental learning from a sample: the algorithm, facing a set of sites to classify, develops a decision tree guided by frequency information in the sample, beginning with the root of the tree and proceeding up to its leaves. This type of algorithm breaks down a complex decision-making process into a series of simpler decisions, with the objectives of correctly classifying as much of the training sample as possible, as well as generalizing, beyond the training sample, while meeting the optimization criterion of minimum error rate in the classification and minimum size of the tree (Quinland, 1986; Safavian & Landgrebe, 1991).

Once the model is established, it can be used to predict the ecological zone to which a given site belongs, by providing only the site's abiotic features as input. The model follows the decisions in the tree, from the root (beginning) node up to the leaf node that contains the response (zone 1, or zone 2, etc.). It is, however, generally schematized as an upside-down tree, as shown on Fig. 7.

The key steps to grow a binary decision tree are

(1) Start with all input data (N grid cells described by m variables, or *features*), and examine all possible binary splits on every feature.
(2) Select a split according to the best value of the optimization criterion, Gini's diversity index (Rao, 1982; Breiman et al., 1984).
(3) Recursively repeat for the two child nodes.
(4) Stop splitting when either the node is "pure" (i.e., it only contains observations from one zone), in the case of a fully grown tree, or when any split imposed on this node would produce "children" with fewer than the minimum leaf observations imposed, as is the case in a "pruned" tree.

DT models have the advantage of transparency; they can give good insight into relevant features to correctly classify a site. However, they have a tendency to overfit, if not pruned. Pruned trees are smaller, more general, and might have a larger training error, but a smaller testing error (Schapire, 2007).

Random forest algorithms

The RF algorithm uses many random trees rather than a single tree. The idea behind this algorithm is to induce randomness (or noise) in the process of tree growth and, therefore, stabilize the model and prevent overfitting. The algorithm combines bagging (or Bootstrap Aggregation) with an additional algorithm introducing randomness at each node of the tree, which therefore uncorrelates the trees. If N is the data set size, bagging generates k bootstrap replicas of the dataset (with N data points picked up randomly among N data points, with replacement, k times) and grows decision trees on each of these replicas. In addition to this bagging algorithm, RF uses a modified tree learning algorithm that selects, at each candidate split in the learning process, a random subset of the features. Final predictions are obtained by aggregating over the ensemble of trees.

To construct our models, we used the MATLAB statistics toolbox with 13 (of the 15 listed in Table 1) abiotic variables describing the marine landscape, constituting m features, used as input to predict the zone label at each site. The sample contains N sites or spatial cells, with $N = 21616$. Eastern-ness and Northern-ness were not included because they were found not to be significant at that scale.

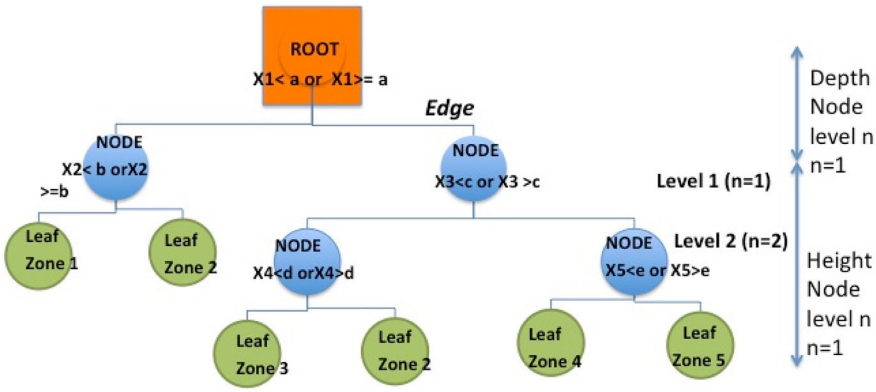

Fig. 7 Schematic view of a binary decision tree

Results

Marine landscape features

The initial question that this study should answer is whether we need all of the 13 variables or features to build a good predictive model of marine landscapes. This question is answered using the proposed models to assess the sensitivity of results to the set of features selected as input.

In total, 42 combinations of the variables were assumed as likely combinations (*sets*), and for each of those, a decision tree was grown as well as a Random Forest. The number of nodes in the decision tree and the classification error (tenfolded cross validation) of the RF model were used as quality indicators of the set. An optimal set of features would be one that minimizes the classification error, the model complexity (number of nodes), and the number of input variables.

The number of nodes and classification errors resulting from the models are plotted in Fig. 8, versus the sets featuring all possible combinations of 12 features as well as the initial 13 feature set: Tidal current ("Tide"); wave current ("Wave"); water depth ("Depth"); distance to coast ("Dist"); slope ("Slope"); roughness ("Rough"); phi median ("Phi"); fraction of clay ("Clay");Stratification ("Strat"); Sea Surface Temperature ("SST"); Aspect Ratio ("Aspect"); and bathymetric position Index ("BPI") for large and small scale ("BPIL"; "BPIS"). Figure 8 shows that (1) the error is small and relatively insensitive to omitting one of most of the 13 variables; (2) while (1) is true for most of the features/variables, it is not true for a few of these: the distance to coast, the tidal drag force, the stratification, and the phi median sediment size.

The distance to coast emerges as the key variable at the sub-regional scale; omitting it from the list of features significantly increases the classification error (in particular for the DT model) and increases the depth of the tree and, hence, the complexity of the model. Tidal drag force, Water Stratification, and the Phi median sediment size appear as secondary key variables. By contrast, the Bathymetric Position Index and the Bathymetry do not bring any relevant information that helps with the classification.

In order to reduce the number of necessary variables to the minimum, we repeated similar simulations for many sets of feature combinations, combining 4–11 features, and we similarly assessed the sensitivity of the models to the number of input features. All simulation results are shown in Table 3 and also illustrated in Fig. 9. Figure 9 plots the RF classification error (tenfolded cross validated), the size of the decision tree (number of nodes), and the number of variables used in the model. The best combinations are shown on the facing corner, in the blue zone regrouping small errors, lowest numbers of nodes (small tree), and lowest numbers of variables (shown by the colored surface). A "zoom" on this optimum area is shown in Fig. 10, which regroups cases with the smallest errors and the lesser number of nodes. Each case is labeled with a number referring to the case listed in Table 3, as well as its rank in accuracy, based on the random forest algorithm. For instance, case 1 corresponds to the highest classification accuracy, while case 42 corresponds to the least accurate model.

Combining accuracy, minimum number of variables, and smallest tree size, the optimum model

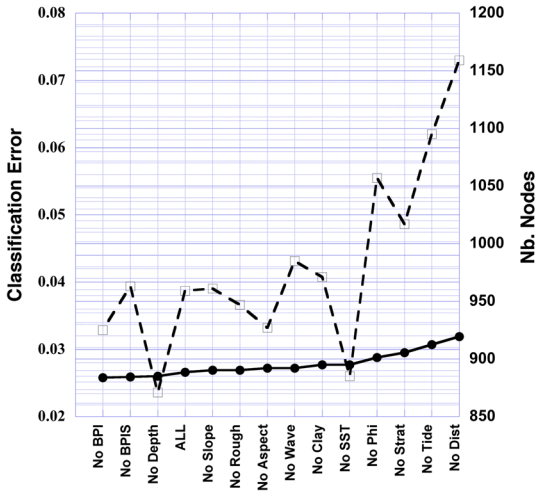

Fig. 8 Classification error (tenfolded cross validated) for the RF model (*circle*) and number of nodes (*dashed*) for the DT model, versus model type. "All" is built with the full set of 13 features, and the 13 other types are all possible 12 feature combination models, labeled as "*No one feature*"

appears to be case 3, which only requires 7 variables: Tide, Wave, Distance, Phi median, Clay, Sea Surface Temperature, and Stratification. Cases 10 and 13 also appear as good candidates, if one restricts the number of variables to 6 and 5, respectively: (tide, wave, distance, phi median, clay, and stratification) for case 10, and (tide, distance, phi median, and clay) for case 13. Both models slightly loose in accuracy and gain in complexity, but do not require the inclusion of water stratification as an input; case 13 additionally does not require the wave drag force as an input. These two models constitute an interesting alternative that reduces the effort on the input features, especially case 13, which reduces the modeling effort in input to tide velocities only.

The best feature combinations are summarized in Table 4. In the following, we use case 3 as our standard model for subsequent analyses.

Predicting ecological zones with the decision tree model

The DT model built for case 3 (Table 4) was run to predict the marine landscapes in the entire SAMP area. Results are shown in Fig. 11. At each site, the model was provided with data for 7 abiotic variables as inputs, and based on these data, it predicted the ecological zones pertaining to the site. Differences between observed (Fig. 4) and predicted zones (Fig. 11) are barely noticeable; the tenfolded cross validation indeed predicts that only 5% of samples are misclassified (Table 3). The resubstitution error is larger, estimated at 13% of the fraction of misclassified samples.

Predicting ecological zones with the random forest model

Results using the Random Forest model are shown in Fig. 12, where the model only has a 2.5% error when using the tenfolded cross-validation procedure. The classification error versus the number of trees grown is plotted in Fig. 13, where we see that one can achieve a stabilized model after about 40 tree growths, and additional trees do not further improve result accuracy.

Three measures of accuracy are shown on the figure, i.e., the standard tenfolded cross validation, the out-of-bag error, and the hold-out error. In hold-out validation tests, a specified fraction of the data (e.g., 30%) is removed, and the rest of the data is used for training. The out-of-bag validation calculates the probability of misclassification for out-of-bag observations in the training data, which are N randomly selected observations out of N total observations, with replacement. From statistical theory, drawing N out of N observations with replacement leads to omitting in average about 37% (1/e) of the observations for each decision tree. The hold-out error represents an upper bound of the model accuracy, since it is considered to overestimate the classification error; here, it stabilizes at 5%. The out-of-bag and tenfolded cross-validation classification errors converge toward 3 and 2.5%, respectively. While the tenfolded cross-validation error is standard in decision tree and ensemble classifications, the out-of-bag error is considered an unbiased estimator, and the obvious choice in RF calculations since the method is based on a bagging algorithm. Both classification errors indeed are consistent.

Sensitivity to input data

The question of the results' sensitivity to the accuracy of marine landscape data is an important issue that is briefly discussed in this section. First, let us recall that the model is based on mean seasonal variables,

Table 3 List of features used to build 42 models using each algorithm: (i) DT; and (ii) RF, and resulting classification error (cross validated—tenfolded)

Variables	Label	Number of variables	Number of nodes	Cross-validated classification error (tenfolded)	
				Decision tree	Random forest
No BP AS SL	1	9	855	0.053	0.0253
No BP SL	2	10	919	0.051	0.0256
TI WA DI PH CL SS ST	3	7	867	0.050	0.0258
No BP	4	11	925	0.052	0.0258
No BPS	5	12	963	0.053	0.0259
No DE	6	12	871	0.054	0.026
No BP AS SL RO	7	8	883	0.048	0.0261
ALL	8	13	959	0.055	0.0266
TI WA DE DI PH CL ST	9	7	867	0.049	0.0268
TI WA DI PH CL ST	10	6	909	0.049	0.0268
No RO	11	12	947	0.053	0.0269
No SL	12	12	961	0.054	0.0269
TI DI PH CL ST	13	5	917	0.047	0.027
No AS	14	12	927	0.053	0.0272
No WA	15	12	985	0.054	0.0272
No SS	16	12	885	0.052	0.0277
No CL	17	12	971	0.057	0.0277
TI WA DI PH ST SS	18	6	939	0.053	0.0281
No CL BP	19	10	971	0.059	0.0281
No PH	20	12	1057	0.060	0.0288
TI WA DI CL ST SS	21	6	943	0.056	0.0288
No ST	22	12	1017	0.059	0.0295
TI WA DI SL CL ST SS AS	23	8	941	0.057	0.0299
TI DI CL ST	24	4	971	0.053	0.03
TI WA DI SL CL ST SS	25	7	947	0.058	0.0302
TI DI SL CL ST SS AS	26	7	959	0.056	0.0303
No TI	27	12	1095	0.062	0.0307
TI WA DI SL CL ST AS	28	7	939	0.056	0.0308
TI DI PH CL	29	4	1051	0.057	0.0309
TI DI SL CL ST SS	30	6	957	0.056	0.0312
TI WA DI SL CL ST	31	6	953	0.057	0.0313
TI DI SL CL ST AS	32	6	967	0.058	0.0317
No DI	33	12	1159	0.071	0.0319
TI DI SL CL ST	34	4	959	0.054	0.0321
TI DI SL CL ST	35	5	979	0.057	0.0325
TI WA DI PH	36	4	1199	0.065	0.0364
TI WA DI CL	37	4	1137	0.064	0.0365
TI DI SL CL SS	38	5	1193	0.068	0.0392
TI DI SL CL SS AS	39	6	1179	0.071	0.0392
TI DI PH	40	3	1371	0.080	0.049
TI DI CL	41	3	1291	0.083	0.0551
WA DI CL	42	3	1899	0.134	0.0949

TD tide, *WA* wave, *DE* depth, *DI* distance, *SL* slope, *RO* roughness, *PH* phi of median sediment size, *CL* fraction of clay in sediment, *AS* aspect ratio, *ST* stratification, *SS* sea surface temperature, *BPI* bathymetric position index

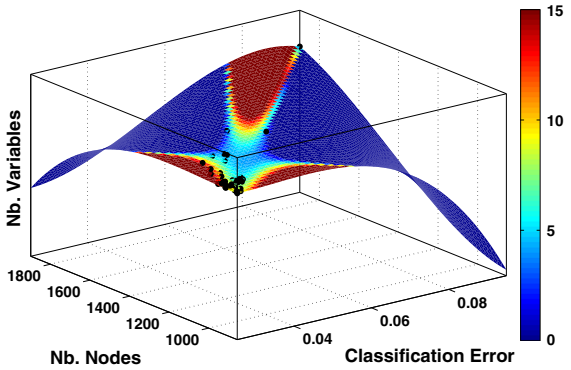

Fig. 9 RF classification error (tenfolded cross validated) as a function of the decision tree number of nodes and number of variables (a polynomial fit shows the number of variables as a color scale), for 42 feature combinations, using 4–13 features/variables. The best feature combinations are regrouped in the front corner of the *blue area*. (Color figure online)

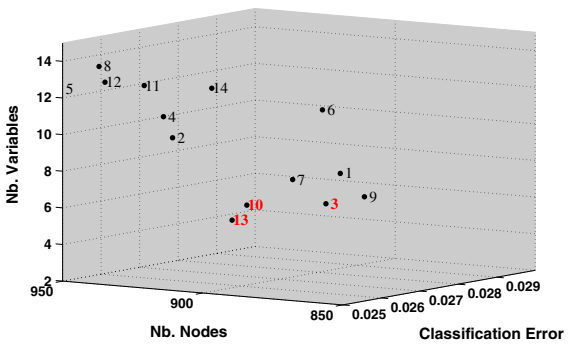

Fig. 10 Zoom on the front corner of *blue area* in Fig. 9 (the fitted surface was removed for clarity at this scale). (Color figure online)

meaning that these values are very robust and carry, by definition, a very small error. Concerning the model, one desires both robustness (i.e., an algorithm not too sensitive to errors in the input data) and a sufficient discrimination ability to accurately identify patterns. In other words, we seek an algorithm that is both unsensitive and sensitive to input data: insensitive to noise, but sensitive to a true change in pattern. This poses the dilemma of the *overfitting* concept (briefly discussed in the "Data and methods" section): the model must fit the data, but not too much. In order to have a grasp on how these models perform, we introduced randomness in the input data and checked the accuracy of the classification using the two models stochastically. Before building each tree, the algorithm is provided with maps of variables slightly deformed by a random error. The error is assumed to be Gaussian distributed, within an interval of variation proportional to the range of variation of the variable; here, arbitrarily set to 1, 2, or 3%. Results show an increase in misclassification with both models, once the random uncertainty on the variable mean values is introduced. The RF algorithm proves, however, more robust than the DT algorithm, as expected from the theory: 81% versus 78% of correct classifications for an added uncertainty of 1%, for RF and DT, respectively; a 4% versus 6% drop in accuracy for RF and DT, respectively, when the uncertainty on the mean value increases from 1 to 3%. One can conclude that besides the fact that the RF model is more robust than the DT model (a fact known from theory), misclassified samples are increased by a relatively large amount, 16–20%, when randomness is introduced (1–3%), which is both bad and good news. While classification is sensitive to uncertainty in input data, which is a little annoying if this uncertainty is an error, it is also sensitive to a true variability in the input, which is desirable. Mapping classification results (Fig. 14), however, clarify these mixed properties by showing the adequacy and power of the method: the misclassified samples are obviously located near the clusters' boundary in transitional areas; adding randomness can move some of these samples beyond the edge of one region to the adjacent region. This indeed corresponds to a reality and reminds us that the estimated *zones* are often not separated by hard, but by soft margins, or gradients, as is the transition between the Rhode Island sound 1 and 2 zones. The Rocky zone, however, stays very insensitive to any randomness in the input data.

Discussion

Our results show that the 13 abiotic variables initially selected to describe marine landscapes are not all necessary to provide an accurate model, and that using only 7 variables can provide an optimum model which (i) leads to the minimum classification error; (ii) is the simplest possible model (i.e., with a minimum number of nodes, therefore more robust); and (iii) has a minimum number of input variables. Here, the 7 input variables are tide and wave drag forces, distance to coast, median sediment size, fraction of clay in the sediment, water stratification, and sea surface

Table 4 Summary of the optimum choices of features to built a performing DT or RF model

Marine landscape descriptor	Case	TD	WA	DI	PH	CL	SS	ST
Variable number in tree		X1	X2	X3	X4	X5	X6	X7
Best 7 variable combination	3	x	x	x	x	x	x	x
Best 6 variable combination	10	x	x	x	x	x		x
Best 5 variable combination	13	x		x	x	x		x

Three possibilities are presented according to the number of variables available, 7, 6 or 4

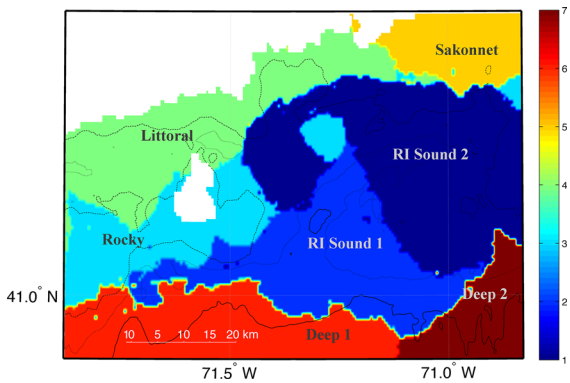

Fig. 11 Predicted ecological zones using the DT model. *Contour lines* indicate bathymetry: 30 m (*dashed black*); 40 m (*dashed-dot gray*); 50 m (*solid black*)

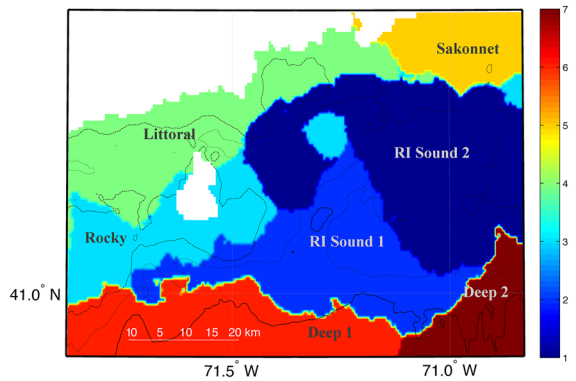

Fig. 12 Predicted ecological zones using the RF model. *Contour lines* indicate bathymetry: 30 m (*dashed black*); 40 m (*dashed-dot gray*); 50 m (*solid black*)

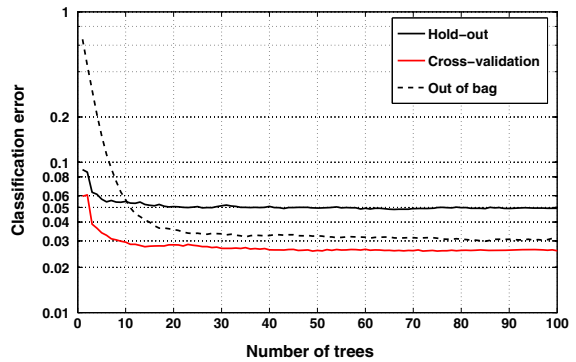

Fig. 13 Classification error for the RF model using three measures of error: (1) hold-out error when holding out 30% of the sample and training over 70% (Test); (2) cross-validation error using a tenfolded cross validation; and (3) out-of-bag error

temperature. Distance to coast was shown to be the most discriminant variable, followed by the tide drag force, and sediments characteristics (median sediment size and fraction of clay in the sediment).

At the scale of the presently defined ecological zones, the BPI is found to be irrelevant. Similar observations were made in North Sea studies (Degraer et al., 2008), in which the large-scale distribution of the major macrobenthic communities was predicted. For predictions at smaller scales, the depth-related variable may become more important, as is the case when isolating homogeneous assemblages in a tropical reef area (Malcolm et al., 2011). At larger scales, other variables, such as distance to coast and sedimentology, might have more discriminating power than bathymetry alone.

It should be pointed out that high accuracy can still be obtained when using only 6 or 5 variables as, for instance, in Cases 10 and 13. In these cases, both models slightly lose in accuracy and gain in complexity, but they are still accurate enough without the need to include the water stratification as input. Hence, these two models provide an interesting way of reducing the monitoring effort required to measure the input features; Case 13, in particular, limits the modeling effort to tide velocity and stratification.

Both DT and RF models were shown to be good predictors, with misclassification errors on the order of 5 and 3%, for each model, respectively. DT models, however, are known to be relatively unstable, meaning that they are sensitive to input data, with a tendency to overfit. The RF model includes randomness in the

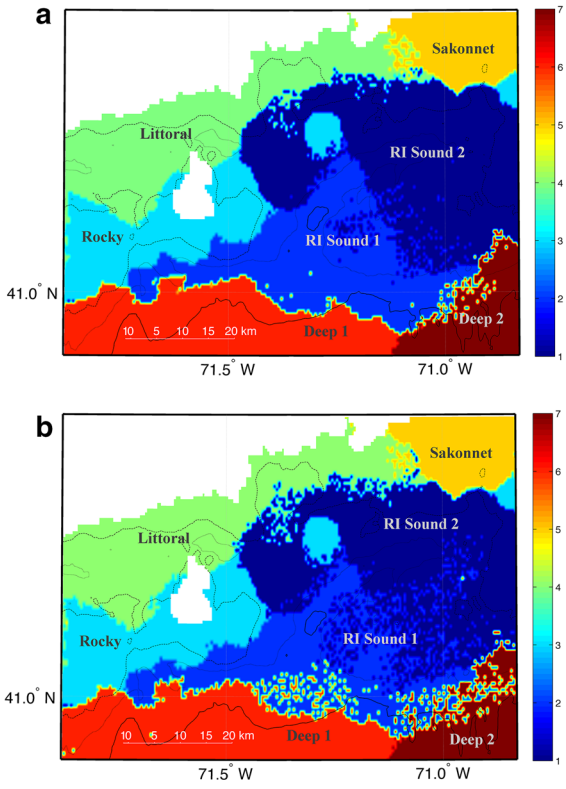

Fig. 14 Predicted ecological zones using the RF model in stochastic mode, introducing 1% (**a**) or 3% (**b**) of random Gaussian uncertainty in the abiotic data used as input in the model. *Contour lines* indicate bathymetry: 30 m (*dashed black*); 40 m (*dashed-dot gray*); 50 m (*solid black*)

process of tree growth, which solves both the stability and overfitting issues usually found in decision trees. As stated by Breiman (2001), "*Random Forests do not overfit*." Results indeed show very small cross-validation and out-of-the bag errors, which confirm the ability of the model to be an excellent predictor.

Besides this epistemic observation on the algorithms, the deliberate introduction of randomness in input data demonstrates the superior robustness of the RF algorithm, although not overwhelmingly, which still qualifies the DT algorithm as an acceptable simplified approach. The stochastic approach, however, reinforces the validity of the method by showing that some variability introduced in input data changes the classification, but in a "meaningful" way: data points, which were near the boundary of one region, fell into the adjacent region if their value was slightly changed, which is what one would expect since the cluster limits have truly soft margins. Ultimately, we expect to use such an algorithm to detect changes due to the wind farm impact.

Our ultimate objective was the development of a machine learning model, able to infer the belonging of a site to an ecological zone by providing only minimal input data, such as marine landscape information only, in an area where there is no or little ecological information. The two models tested here to this effect, DT and RF, have proved accurate, within the limits of the cross-validation error, when applied to the RIOSAMP area, and we can therefore anticipate that they can also be used outside of the limits of this area. However, these models are based on an empirical method, which as the name "machine learning" indicates that it is based on learning from a specific set of field data. Therefore, the question of the limits of validity of the inference domain stays open: how far beyond the RIOSAMP area boundary can we use the models? The quick answer would be "*anywhere,*" knowing that the algorithms will not classify any site with a different marine landscape than the RIOSAMP, since they did not learn how to recognize it. There are, however, a few restrictions to this statement, due to the learning curve of the algorithms, which are discussed in the following. To extend one model domain of applicability, more *samples* similar to those of the RIOSAMP should be added to the model input, and the model could learn how to recognize new sites from this updated dataset. Let us note that neither specific spatial sampling discretization, nor minimum sample size, nor nicely distributed data are required for such models to adequately perform.

At this stage of the models' learning curve, the models are able to recognize an environment similar to that in the RIOSAMP area, which makes it operational in parts of the Southern New England Region. Jordaan et al. (2010, 2012), using a similar multivariate analysis as that developed for the RIOSAMP (Grilli et al., 2013), have zoned the US Northeast coast at a larger scale and identified a region extending from the New York Bight to Georges Bank, as a relatively homogeneous ecological zone, very distinct from the Gulf of Maine region. One could expect that the coastal area of this region reflects the "*population*" or "*universe*" from which the RIOSAMP is a sample. This could be checked by making comparisons of the assemblages using standard statistics, if the information is available, or simply and less costly, by applying the algorithms and checking what is classified and

what is not, and what is correctly classified and what is not (only comparisons of relevant sites are required rather than a full mapping). Let us note that the domain of applicability should not extend further than the coastal section of this region, since the distance to coast is one of the discriminant variables used to develop the models, whose range of variation is limited to the range of variation represented in the sample (i.e., that of RIOSAMP). The current models can therefore be applied along the shore in a bandwidth extending up to about 80 km offshore.

Two theoretical issues remain, which could question our confidence in the accuracy of the inference process and reinforce the need to extend the sample and its cross validation.

The first one is referred to as the "Black Swan" problem of inductive learning (Hamel, 2009), which refers to the inability of a learning machine algorithm to predict a rare event, if the machine does not know that this event exists: if the sample does not have a black swan, one cannot predict its presence when one extends the domain to the whole universe. However, if 99% of the swans are white, the model is still 99% accurate. Accordingly, one can wonder whether there are alternative ecological patterns corresponding to identical landscapes that would not be present in the sample. This question can, however, be answered similarly to above: if these are not present, they are not frequent and the potential risk of not predicting them is thus minimum.

The second issue can be described from the models' logic perspective. The methods used guarantee the models' convergence, which means that at least one landscape corresponds to one ecological region; one can also say that the models are surjective. However, once outside the training sample, one could have a marine landscape associated in the training sample to a specific ecological region, say E1, associated to an ecological region that is not present in the training sample, say E2. In this case, the algorithms would naturally misclassify E2 and would label this zone as E1, since they were not trained to recognize differences between E2 and E1. If such a situation occurred, the models would have to be retrained with additional discriminant variables to capture the threshold, which would discriminate both zones E2 and E1; geographic coordinates could ultimately be these discriminant variables. Therefore, if the models are applied outside the targeted area, they must be validated with data specific to the predicted zones in order to catch such potential misclassifications.

To summarize, learning machine algorithms learn from the field, so the more exposed they are to samples, the more they learn, the more complex they become, and the more reliable the inferences they make. The key is therefore to expand the sample as much as possible, and re-train the algorithms each time when new data are acquired. Based on their training data, the current models (DT and RF) are thus operational to work with confidence, besides the targeted RISAMP area, in similar oceanographic and ecological adjacent areas, such as East of the RIOSAMP, in the offshore Massachusetts zone (keeping in mind that the distance to coast restricts the domain extent). The models are also useful for when when gaps in data/monitoring occur in a region of interest.

Conclusions

In earlier work (Grilli et al., 2013), we established an ecological typology and zoning of the "Rhode Island Ocean Special Area Management Plan" (RIOSAMP) area, aimed at optimizing the siting of future wind farms and guiding the required pre- and post-construction monitoring in this coastal area. To do so, we performed an abiotic typology, attempting to relate marine landscapes to an ecological typology, with the aim of reducing the cost of the extensive ecological surveys required in monitoring campaigns. Although clear similarities in patterns emerged that suggested a correlation between both typologies, a quantitative correlation could not be established using standard statistics. These results naturally led to the question of whether we could develop a method (or model) to correlate marine landscapes with ecological zones in a meaningful way, and with which one could thus predict ecological regions only based on abiotic variables. This question was of course targeted to the RIOSAMP area, but we were also wondering whether such a prediction could be extended outside of that area. The present study reports the follow-up research effort devoted to answering this question.

First, the marine landscape analysis was refined, expanding the initial abiotic data set from 6 to 15 variables, including new potential discriminant variables: in particular, two environmental disturbance factors, the tidal current and the wave power. Then, a

machine learning model (actually based on two algorithms, DT and RF) was established, between abiotic variables and ecological zones, to quantify the relationship between the two datasets and provide a tool for predicting the belonging of a site to a specific ecological zone, only based on abiotic information at the site.

Results of the study demonstrated that one can indeed accurately relate ecological zones and marine landscapes, using either a DT or a RF model. Between the two models, the RF model is more accurate (with about a 3% classification error), more stable, and robust. We showed that the 15 variables initially selected to describe marine landscapes are not all necessary to provide an accurate model in the case of the RIOSAMP area, but that only using 7 variables can already provide an optimum model. These variables are tide and wave drag forces, distance to coast, the internal friction angle of median sediment size, the fraction of clay in sediment, water stratification, and sea surface temperature.

The implications of these findings in terms of wind farm monitoring optimization are summarized hereafter:

1. Assuming that the algorithms have "learned" enough site-specific data, one can theoretically zone any area with a similar marine landscape as the RIOSAMP, using only the seven selected abiotic variables. Some restrictions, however, apply, as discussed above.
2. The method is ideal to fill in gaps in areas where the ecological sampling and/or zoning is incomplete.
3. This study provides additional insight on where to direct the monitoring effort. The abiotic variables are much easier to obtain than the biotic variables, and they have generally a better spatial coverage and a higher spatial discretization. While the distance to coast is straightforward to obtain, the sediment characteristics need to be accurately surveyed with enough discretization. Hydrodynamic and hydrographic variables (i.e., waves, tides, sea surface, stratification, and temperature) generally have a good spatial coverage and discretization, since they are typically obtained from state-of-the-art hydrodynamic ocean models, or from satellite data (e.g., sea surface temperature).
4. The developed models, however, do not entirely free us from ecological monitoring, which is still necessary to validate the model in the inference domain, outside of the RIOSAMP area, as discussed at length in the "Discussion" section. However, the predictive ability of the models provides a useful tool for optimizing the type of monitoring desired and its spatial deployment, in strategic areas, relative to the models' predictions.

It is important to stress that these learning algorithms do not require any particularly well-behaved data, regular sampling, or minimum sample size, as typically required in standard statistics. This strongly reduces the monitoring requirements while expanding the learning sample.

In future work, we expect to expand the models' ability to recognize additional zones and species distributions, by providing them with new "learning" data from other coastal areas open to wind farm development. The models could eventually reach a *knowledge* level such as to provide a general and robust tool to assist developers, local communities, and federal administrations with wind farm siting optimization. In work performed in parallel with the implementation and validation of the ecological *classifier* models proposed here, we are developing a broader scope wind farm siting optimization toolbox, in which *regressive* learning machine models are used to predict the spatial distribution of specific socio-economically and ecologically sensitive species, such as lobsters (*Homarus americanus*) or North Atlantic right whales (*Eubalaena glacialis*). [Effects on lobster and whale populations are among the most sensitive issues for wind farm siting in the RIOSAMP area, and likely elsewhere in New England waters in the U.S., from the perspective of local communities, for historical–cultural, social–economical as well as ecological reasons.] These regressive models could be used in a second stage to refine the initial ecological zoning performed here. Similarly to the ecological zoning models, these new models can also extrapolate results outside of the boundary of their test area or fill in gaps in the ecological monitoring within their test area. These new developments will be detailed in future publications.

References

Austin, M. P., 2002. Spatial prediction of species distribution: an interface between ecological theory and statistical modeling. Ecological Modelling 157: 101–118.

Bahn, V. & B. J. McGill, 2013. Testing the predictive performance of distribution models. Oikos 122: 321–331.

Biau, G., 2012. Analyzing of a random forest model. The Journal of Machine Learning Research 13: 1063–1095.

Bishop, C. M., 1995. Neural Network for Pattern Recognition. Clarendon Press, Oxford.

Bohaboy, E., A. Malek & J. Collie, 2010. Baseline characterization: data sources, methods, and results. Technical report for Rhode Island Ocean Special Area Management Plan, University of Rhode Island, Kingston, RI.

Breiman, L., 2001. Random forests. Machine Learning 45: 5–32.

Breiman, L., J. Friedman, C. J. Stone & R. A. Olshen, 1984. Classification and Regression Trees. CRC Press, New York.

Buddemeier, R. W., S. V. Smith, D. P. Swaney, C. J. Crossland & B. A. Maxwell, 2008. Coastal typology: an integrative "neutral" technique for coastal zone characterization and analysis. Estuarine, Coastal and Shelf Science 77: 197–205.

Codiga, D. L. & D. S. Ullman, 2010. Characterizing the physical oceanography of Coastal Waters off Rhode Island. Technical report for Rhode Island Ocean Special Area Management Plan, University of Rhode Island, Kingston, RI.

Connor, D. W., J. H. Allen, N. Golding, L. M. Lieberknecht, K. O. Northen & J. B. Reker, 2003. The National Marine Habitat Classification for Britain and Ireland. Joint Nature Conservation Committee, Peterborough.

Cowen, R. K., C. B. Paris & A. Srinivasan, 2006. Scaling of connectivity in marine populations. Science 311: 522–527.

Cutler, D. R., T. C. Edwards Jr, K. H. Beard, A. Cutler, K. T. Hess, J. Gibson & J. J. Lawler, 2007. Random forests for classification in ecology. Ecology 88: 2783–2792.

Dean, R. & A. Dalrymple, 1984. Water Wave Mechanics for Engineers and Scientists. Prentice-Hall, Publishing, Englewood Cliffs, NJ.

De'ath, G. & K. E. Fabricius, 2000. Classification and regression trees: a powerful yet simple technique for ecological data analysis. Ecology 81: 3178–3192.

Degraer, S., E. Verfailie, W. Willems, E. Adriaens, V. Van Lanker & M. Vincx, 2008. Habitat suitability modeling as a mapping tool for macrobenthic communities: an example from the Belgian part of the North Sea. Continental Shelf Research 28: 369–379.

Derous, S., E. Verfaille, V. Van Lancker, W. Cortens, E. W. M. Steinen, K. Hostens, I. Mouleurt, H. Hillewaert, J. Mees, K. Deneust, P. Deckers, D. Cuvelier, M. Vincx & S. Degraer, 2007. A Biological Valuation Map for the Belgian Part of the North Sea: BWZee Final Report. Belgian Science Policy, Brussels.

Drake, J. M., C. Randin & A. Guisan, 2006. Modelling ecological niches with support vector machines. Journal of Applied Ecology 43: 424–432.

Ewers, R. M. & R. K. Didham, 2006. Confounding factors in the detection of species responses to habitat fragmentation. Biological Reviews 81: 117–142.

Egbert, G. D., 1997. Tidal data inversion: interpolation and inference. Progress in Oceanography 40(1): 53–80.

Elith, J., C. H. Graham, R. P. Anderson, M. Dudík, S. Ferrier, A. Guisan, R. J. Hijmans, F. Huettmann & E. N. Zimmermann, 2006. Novel methods improve prediction of species' distributions from occurrence data. Ecography 29(2): 129–151.

Fahrig, L., 2003. Effects of habitat fragmentation on biodiversity. Annual Review of Ecology, Evolution, and Systematics 34: 487–515.

French-McCay, D., M. Schroeder, E. Graham, D. Reich & J. Rowe, 2011. Ecological value map (EVM) for the Rhode Island Speciall area management plan. Technical report for Rhode Island Ocean Special Area Management Plan, University of Rhode Island, Kingston, RI.

Guisan, A. & W. Thuiller, 2005. Predicting species distribution: offering more than simple habitat models. Ecology Letters 8: 993–1009.

Guisan, A. & C. Rahbek, 2011. SESAM – a new framework integrating macroecological and species distribution models for predicting spatio-temporal patterns of species assemblages. Journal of Biogeography 38: 1433–1444.

Grilli, A. R., T. Lado Insua & M. L. Spaulding, 2013. A protocol to include ecosystem services constraints in a wind farm cost model. Journal of Environmental Engineering 139: 176–186.

Grilli, S.T., J. Harris, R. Sharma, L. Decker, D. Stuebe, D. Mendelsohn, D. Crowley, & S. Decker, 2010. High resolution modeling of meteorological, hydrodynamic, wave and sediment processes in the Rhode Island Ocean SAMP study area. Technical report for Rhode Island Ocean Special Area management plan, University of Rhode Island, Kingston, RI.

Hamel, L. H., 2009. Knowledge Discovery with Support Vector Machine. Wiley, Hoboken, NJ.

Jordaan, A., 2010. Fish assemblages spatially structure along a multi-scale wave energy gradient. Environmental Biology of Fishes 87: 13–24.

Jordaan, A., Y. Chen, D. W. Townsend & S. Sherman, 2010. Identification of ecological structure and species relationships along an oceanographic gradient in the gulf of Maine using multivariate analysis with bootstrapping. Canadian Journal of Fisheries and Aquatic Sciences 67: 1–19.

Jordaan, A., M. G. Frisk, L. S. Incze, N. H. Wolff, L. Hamlin & Y. Chen, 2012. Multivariate dissemination of species relationships for use in marine spatial planning. Canadian Journal of Fisheries and Aquatic Sciences 70: 316–329.

Kenney, R. D. & K. J. Vigness-Raposa, 2010. Marine mammals and sea turtles of Narragansett Bay, Block Island Sound, Rhode Island Sound, and Nearby Waters: an analysis of existing data for the Rhode Island Ocean special area management plan. Technical report for Rhode Island Ocean special area management plan, University of Rhode Island, Kingston, RI.

Kostylev, V. E. & C. G. Hannah, 2007. Process-driven characterization and mapping of seabed habitats. In Todd, B. J. & H. G. Greene (eds), Mapping the Seafloor for Habitat Characterization: Geological Association of Canada. Geological Association of Canada, St. John's: 171–184.

LaFrance, M., E. Shumchenia, J. King, R. Pockalny, B. Oakley, P. Sheldon & J. Boothroyd, 2010. Benthic habitat distribution and subsurface geology in selected sites from the Rhode Island Ocean Special Area Management Study Area. Technical report for Rhode Island Ocean Special Area Management Plan, University of Rhode Island, Kingston, RI.

Lundblad, E., D. J. Wright, J. Miller, E. M. Larkin, R. Rinehart, S. M. Anderson, T. Battista, D. F. Naar & B. T. Donahue, 2006. A benthic terrain classification scheme for American Samoa. Marine Geodesy 29: 89–111.

Malcolm, A. H., A. Jordan & S. D. A. Smith, 2011. Testing a depth-based habitat classification system against reef-fish assemblage patterns in a subtropical marine park. Aquatic Conservation: Marine and Freshwater Ecosystems 21: 173–185.

Malek, A., J. Collie, M. LaFrance, J. Collie & J. King, 2010. Fisheries ecology and benthic habitat in Rhode Island and Block Island Sounds for the Rhode Island Ocean special area management plan. Technical report for Rhode Island Ocean Special Area Management Plan, University of Rhode Island, Kingston, RI.

McMullen, K. Y., L. J. Pope, J. F. Denny, T. A. Haupt, & J. M. Crocker, 2008. Side-scan sonar imagery and surficial geologic interpretations of the sea floor in central Rhode Island Sound. U.S. Geological Survey. Report for U.S. Department of Interior, Reston.

Needell, S. W. & R. S. Lewis, 1984. Geology of Block Island Sound, Rhode Island, and New York. U.S. Geological Survey Miscellaneous Field Studies Map MF-1621, scale 1:125,000,4 sheets.

Noss, R. F., 1990. Indicators for monitoring biodiversity: a hierarchical approach. Conservation Biology 4: 355–364.

O'Reilly C., A. R. Grilli & G. Potty, 2013. Micrositing optimization of the Block Island Wind Farm, RI, USA. In Proceedings of International Conference Ocean, Offshore and Arctic Engineering (OMAE 2013, Nantes 6/9-14/13).

OTIS, 2009 [available on internet at http://www.volkov.oce.orst.edu/tides/EC.html].

Pesch, R., G. Schmidt, W. Schroeder & I. Weustermann, 2011. Application of CART in ecological landscape mapping: two case studies. Ecological Indicators 1: 115–122.

Phillips, S. J., R. P. Anderson & R. E. Schapire, 2006. Maximum entropy modeling of species geographic distributions. Ecological Modeling 190: 231–259.

Quinland, J. R., 1986. Induction of decision tree. Machine Learning 1: 81–106.

Rao, C. R., 1982. Gini-Simpson index of diversity: a characterization, generalization and applications. Utilitas Mathematica 21: 273–282.

Reid, J.M., J.A. Reid, C.J. Jenkins, M.E. Hastings, S.J. Williams & L.J. Poppe, 2005. usSEABED: atlantic coast offshore surficial sediment data release. US Geological Survey Data Series 118, version 1.0.

Reid, R. N., L. M. Cargnelli, S. J. Griesbash, D. B. Packer, D. L. Johnson, C. A. Zetlin, W. W. Morse & P. L. Berrien, 1999. Essential fish habitat source document: Atlantic Herring, *Clupea harengus*. Life history and habitat characteristics. National Marine Fisheries Service. NOAA Technical Memorandum NMFS-NE-126.

Richmond, S. & T. Stevens, 2014. Classifying benthic biotopes on sub-tropical continental shelf reefs: how useful are abiotic surrogates ? Estuarine, Coastal and Shelf Science 138: 79–89.

Rinne, H., A. Kaskela, A. L. Downie, V. Tolvanen, M. von Numers & J. Mattila, 2014. Predicting the occurrence of rocky reefs in a heterogeneous archipelago area with limited data. Estuarine, Coastal and Shelf Science 138: 90–100.

Roff, J. C. & M. E. Taylor, 2000. National framework for marine conservation. A hierarchical approach. Aquatic Conservation: Marine and Freshwater Ecosystems 10: 209–223.

ROMS, 2009 [available on internet at https://www.myroms.org].

Rykiel, E. J., 1985. Towards a definition of ecological disturbance. Australian Journal of Ecology 3: 361–365.

SAMP, 2010 [available on internet at www.seagrant.gso.uri.edu/oceansamp/].

Safavian, S. R. & D. Landgrebe, 1991. A survey of decision tree classifier methodology. IEEE. Transactions on Systems, Man, and Cybernetics 21: 660–674.

Schapire, R. E., 2007. Lecture #5 COS 424: interacting with data [available on internet at http://www.cs.princeton.edu/courses/archive/spr07/cos424/scribe_notes/0220.pdf].

Schapire, R. E., 2013. Explaining Adaboost. Empirical inference: 37–52.

Shchepetkin, A. F. & J. C. McWilliams, 2005. Regional ocean model system: a split-explicit ocean model with a free-surface and topography-following vertical coordinate. Ocean Modelling 9: 347–404.

Shumchenia, E. J. & A. R. Grilli, 2012. Enhanced ocean landscape and ecological value characterization for the Rhode Island Ocean special area management plan study area using habitat typology and habitat template approaches. Technical report for Rhode Island Ocean Special Area Management Plan, University of Rhode Island, Kingston, RI.

Sindermann, C. J., 1979. Status of Northwest Atlantic herring stocks of concern to the United States. U.S. National Marine Fishery Service. Technical Report.

Smith, J. M. & A. R. Sherlock, D. T. Resio, 2001. STWAVE: steady-state spectral wave model user's manual for STWAVE, Version 3.0. US Army Corps of Engineers.

Southwood, T. R. E., 1988. Tactics, strategies and templates. Oikos 52: 3–18.

Spaulding, M. L., A. R. Grilli, C. Damon & G. Fugate, 2010. Application of technology development index and principal component analysis and cluster methods to ocean renewable energy facility siting. Marine Technology Society Journal 44: 8–23.

Stockwell, D., 1999. The GARP modelling system: problems and solutions to automated spatial prediction. International Journal of Geographical Information Science 13: 143–158.

Stone, B. D. & L. A. Sirkin, 1996. Geology. In Hydrogeology and Water Resources of Block Island, Rhode Island. U.S. Geological Survey Water-Resources Investigations Report. Providence, RI.

Vapnik, V., 1998. Statistical Learning Theory. Wiley, New York.

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.

Van Lancker, V. & R. Foster-Smith, 2007. How do I make a map? In MESH Guide to Habitat Mapping, MESH Project, 2007. JNCC, Peterborough, p 78 [available on internet at http://www.searchmesh.net/].

Wiley, E. O., K. M. McNyset, A. T. Peterson, C. R. Robins & A. M. Stewart, 2003. Niche modeling and geographic range predictions in the marine environment using a machine-learning algorithm. Oceanography 16: 120–127.

Zacharias, M. A. & J. C. Roff, 2000. A hierarchical ecological approach to conserving marine biodiversity. Conservation Biology 14: 1327–1334.

Zuur, A. F., E. N. Ieno & G. M. Smith, 2007. Analyzing Ecological Data. Springer, New York.