










RESEARCH ARTICLE

WILEY

Machine learning highlights the importance of primary and secondary production in determining habitat for marine fish and macroinvertebrates

Kevin D. Friedland¹  | Michelle Bachman²  | Andrew Davies³  |
 Romain Frelat⁴  | M. Conor McManus⁵  | Ryan Morse¹  |
 Bradley A. Pickens^{6,7}  | Szymon Smoliński^{8,9}  | Kisei Tanaka¹⁰ 

¹Northeast Fisheries Science Center,
Narragansett, Rhode Island, USA

²New England Fishery Management Council,
Newburyport, Massachusetts, USA

³Department of Biological Sciences, University
of Rhode Island, Kingston, Rhode Island, USA

⁴Biométris – Aquaculture and Fisheries Group,
Wageningen University, Wageningen, the
Netherlands

⁵Rhode Island Department of Environmental
Management, Division of Marine Fisheries,
Jamestown, Rhode Island, USA

⁶CSS, Fairfax, Virginia, USA

⁷National Centers for Coastal Ocean Science,
Beaufort, North Carolina, USA

⁸Demersal Fish Research Group, Institute of
Marine Research, Bergen, Norway

⁹Department of Fisheries Resources, National
Marine Fisheries Research Institute, Gdynia,
Poland

¹⁰Atmospheric and Oceanic Sciences Program,
Princeton University, Princeton, New
Jersey, USA

Correspondence

Kevin Friedland, Northeast Fisheries Science
Center, National Marine Fisheries Service,
28 Tarzwell Dr., Narragansett, RI 02882, USA.
Email: kevin.friedland@noaa.gov

Funding information

Interagency agreement #M17PG00028
between NOAA and the Bureau of Ocean
Energy Management

Abstract

1. Species distribution models for marine organisms are increasingly used for a range of applications, including spatial planning, conservation, and fisheries management. These models have been constructed using a variety of mathematical forms and drawing on both physical and biological independent variables; however, what might be called first-generation models have mainly followed the form of linear models, or smoothing splines, informed by data collected in the context of fish surveys.
2. The performance of different classes of variables were tested in a series of species occurrence models built with machine learning methods, specifically evaluating the potential contribution of lower trophic level data. Random forest models were fitted based on the classification of the absence/presence for fish and macroinvertebrates surveyed on the US Northeast Continental Shelf.
3. The potential variables included physical, primary production, secondary production, and terrain variables. For accepted model fits, six variable importance measures were computed, which collectively showed that physical and secondary production variables make the greatest contribution across all models. In contrast, terrain variables made the least contribution to these models.
4. Multivariable analyses that account for all performance measures reinforce the role of water depth and temperature in defining species presence and absence; however, chlorophyll concentration and some specific zooplankton taxa, such as *Metridia lucens* and *Paracalanus parvus*, also make important contributions with strong seasonal variations.
5. Our results suggest that lower trophic level variables, if available, are valuable in the creation of species distribution models for marine organisms.

KEYWORDS

habitat, lower trophic level, random forest, species distribution model

1 | INTRODUCTION

Species distribution models (SDMs) are increasingly being used to describe the distribution and ecological relationships of species. Long-term monitoring data from concurrent, standardized sampling of marine species, and their environment, have provided the necessary data to build SDMs that can identify a species niche and predict their abundances and distributions (Marshall, Glegg & Howell, 2014). Hindcast predictions from such models have provided insight into the ecological or evolutionary changes of a species or ecosystem (Elith & Leathwick, 2009), and combined with predictions of future environmental conditions, SDMs have allowed fisheries scientists to gauge how the distributions of marine species may change (Hobday et al., 2019). Applications for marine fish SDMs include defining essential fish habitat (Laman et al., 2018), constructing fisheries-independent abundance indices for use in stock assessments (Thorson et al., 2015; Cao, Chen & Richards, 2017), designating areas of significance to support marine spatial planning discussions (Robinson et al., 2011), and identifying areas where spatially explicit management would promote stock protection (Chu et al., 2019). Spatial management of resource species may include the need to differentiate habitat use and distribution by life stage and sex, and will often require seasonal habitat characterization (Gruss et al., 2017). Given these important applications, the improvement of these models, both in terms of data inputs and statistical framework, will influence the future of fisheries science and management.

The distribution of a species in time and space is jointly determined by multiple abiotic and biotic processes occurring across different spatial and temporal scales. Abiotic (Grinnelian niche) variables such as temperature and salinity can directly influence the physiological processes and geographic distribution of a species (Kearney & Porter, 2009; Austin & Van Niel, 2011). Whereas biological (Eltonian niche) variables, such as food resources and species interactions, can provide information on what constitutes habitat for a species at a finer spatial and temporal resolution (Sheppard, Lawler & Marsh, 2007; Coops, Wulder & Iwanicka, 2018). One of the primary motivations for species distribution modellers is to incorporate a broader spectrum of ecological and environmental variables to achieve better predictive performance. The ecological and environmental requirements of most commercial stocks are well studied; as such, their occurrence or abundance can be related to environmental, demographic, climatic, and satellite-derived predictor variables. For example, the use of remotely sensed data in SDMs has become common practice in order to provide additional data on ecosystem productivity and seasonality (Leitao, Moreira & Osborne, 2010; Cord et al., 2013).

Most species habitat modelling efforts have evaluated the significance of abiotic factors in describing preferred habitats and predicting their distribution. Physical oceanographic data, principally temperature, are most often used to describe the habitat of species given the direct or indirect impacts that temperature has on the biological rates and behaviours of species, including metabolism, growth, reproduction, migration, and survival (Hare, 2014). As defined by

Hutchinson's fundamental niche (Hutchinson, 1957), temperature has been considered to be part of the niche for marine species and is essential in describing their abundance and distribution. The effect of temperature on the physiology and behaviours of species is often nonlinear, with optimal conditions at given temperatures and suboptimal effects above and below the optimum level. As species distribution modelling in marine ecosystems has accelerated, the potential to add complexity, such as predator-prey relationships, has been recognized (Robinson et al., 2011).

Often, prey availability or lower trophic level production have not been included in modelling efforts because of the limited availability of suitable prey data (Torres, Read & Halpin, 2008); however, including such information is expected to improve model predictions and better describe the available habitat, given that species are supported by both biotic and abiotic features of their environment (Vezza et al., 2015). Prey species have been included in the modelling efforts for predators at various life stages and trophic levels (McManus et al., 2018; Xue et al., 2018). Similarly, although less explicitly explored than oceanographic variables, the inclusion of benthic and depth habitat characteristics for spatio-temporal modelling of species distributions has been shown to improve estimates of abundance (Johnson, Thorson & Punt, 2019; McHenry et al., 2019).

We used a machine learning approach for fitting species distribution models to evaluate a relatively large and comprehensive set of ecological predictors. From these model fits, we analysed a suite of variable metrics to gauge the importance of different classes of variables and individual variables on SDM fitting. We analysed these performance metrics as distinct variables and evaluated them using multivariate statistics. In the process of building species distribution models for fish and macroinvertebrates sampled in a resource-monitoring programme, we recognized that our model procedures had the potential to shed light on the factors affecting the distributional ecology of species and on the operational use of habitat predictors in these types of models. Our goal is to provide insights into species distributional responses relevant to the subject study system and for aquatic ecosystems elsewhere.

2 | METHODS

2.1 | Fish and macroinvertebrate distribution

This study is based on a series of SDMs incorporating habitat features for taxa captured in the Northeast Fisheries Science Center (NEFSC) fishery-independent bottom trawl survey conducted in the US Northeast Shelf (NES) ecosystem, a well-studied continental shelf marine system located along the western boundary of the North Atlantic Ocean (Figure 1). The bottom trawl survey has been conducted each year since 1963 in the autumn, and since 1968 in the spring, using over 300 stations during each season, and is based on a random stratified design. Catches were standardized for various correction factors related to the vessels and gears used in the time series (Miller et al., 2010). The survey data are publicly available at <https://inport.nmfs.noaa.gov/inport/>.

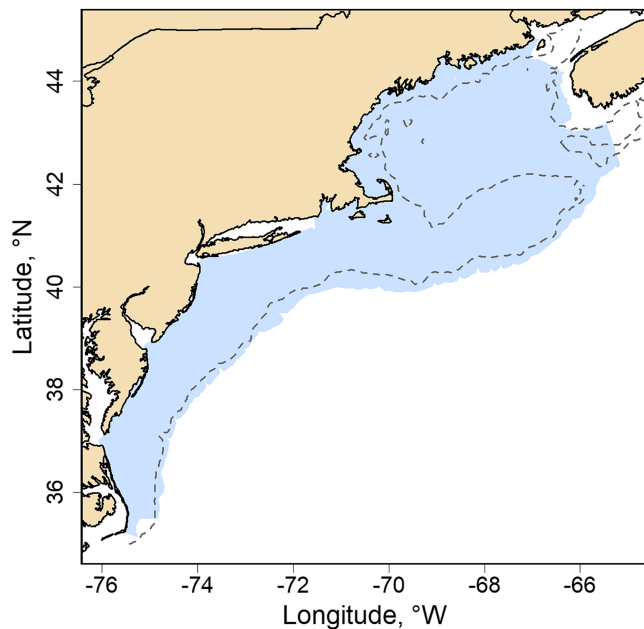


FIGURE 1 Map of the study system, the US Northeast Shelf ecosystem, showing the approximate extent of the survey used to develop habitat models (shaded area). The dashed line is the 100-m depth contour

The models were constructed using random forest (RF) methods (Breiman, 2001) for consistently abundant taxa from the surveys, defined as species that occurred in at least 150 trawl tows, which numbered 96 species. Modelling for each taxon was based on the categorical response of absence or presence of the species in a trawl sample against an initial candidate list of 91 independent covariates that can be organized into four categories: physical environment variables, terrain descriptors, primary production variables, and secondary production (zooplankton) variables (Table 1). Although many of the datasets had longer time series, the modelling was restricted to a dataset of the years 1992–2016, given that salinity measurements have only been available with electronic instrumentation over this time period. In addition, this recent period is also more representative of current environmental conditions and is therefore more relevant to current resource management efforts.

2.2 | Environmental data

Physical and biological environmental data included dynamic variables that changed annually with recurring sampling and static variables that were held constant over the years. Dynamic physical environmental variables, including surface and bottom water temperatures and salinity, represented observations made contemporaneously with survey trawl samples. These parameters were measured using conductivity/temperature/depth (CTD) instruments. The depth of the survey station (in metres) was measured with the vessel depth sounder and is considered a static variable as the depth at a location would not be expected to change from year to year. Benthic terrain descriptors

TABLE 1 Summary of predictor variables used in the development of spring and autumn occupancy habitat models

Predictor variable categories	Description	No.
Physical environment variables	Physical and oceanographic variables including depth, surface and bottom temperature, and surface and bottom salinity, derived from point surveys	5
Benthic terrain descriptors	A series of variables that characterize the structure of benthic habitats, most of which are based on bathymetry data (for details, see Table 2)	19
Secondary production variables	Abundance of zooplankton taxa and a zooplankton biomass index (settled biovolume) composed mostly of copepod species. Some taxa only identified to family or other general category (for details, see Table 3)	19
Remote-sensing primary production variables	Remotely sensed measurements of monthly chlorophyll concentration and gradient magnitude or frontal data for the same monthly fields	24
Remote-sensing physical environment variables	Remotely sensed measurements of monthly sea surface temperature and gradient magnitude or frontal data for the same monthly fields	24

No., number of variables.

included a series of static variables that characterize the shape and complexity of the substrate. Most benthic terrain variables were derived from the depth measurements, such as vector ruggedness, rugosity, and slope (Table 2). Other variables described the substrate itself, such as benthic sediment grain size. The vorticity of benthic currents was also considered as a benthic terrain variable. In addition to the dynamic and instantaneous station temperature variables, monthly sea surface temperature (SST) fields were used to derive static variables based on data from the Moderate Resolution Imaging Spectroradiometer (MODIS) Terra sensor (<https://oceancolor.gsfc.nasa.gov/data/terra/>). Monthly mean SST data and monthly gradient magnitude, or frontal fields of the SST, were assigned to each station. There are many methods used to identify fronts (Belkin & O'Reilly, 2009) in oceanographic data that usually use some focal filter to reduce noise and then identify gradient magnitude with a Sobel filter. Calculations were performed in R using RASTER (2.6-7) by applying a three-by-three mean focal filter and a Sobel filter to generate x and y derivatives, which were then used to calculate gradient magnitudes.

TABLE 2 Summary of benthic terrain predictor variables used in the development of spring and autumn occupancy habitat models

Variable	Notes	References
Complexity: terrain ruggedness index	The difference in elevation values from a central cell and the eight cells immediately surrounding it. Each of the difference values are squared to make them all positive and are then averaged. The index is the square-root of this average	(Riley, DeGloria & Elliot, 1999)
NAMERA BPI	BPI is a second-order derivative of the surface depth using the TNC Northwest Atlantic Marine Ecoregional Assessment (NAMERA) data, with an inner radius = 5 m and outer radius = 50 m	(Lundblad et al., 2006)
NAMERA VRM	Vector ruggedness measure (VRM) measures terrain ruggedness as the variation in three-dimensional orientation of grid cells within a neighbourhood based on NAMERA data	(Hobson, 1972; Sappington, Longshore & Thompson, 2007)
Prcurv: 2, 10, and 20 km;	Benthic profile curvature at spatial scales of 2, 10, and 20 km derived from depth data	(Winship, 2018)
Rugosity	A measure of small-scale variations of amplitude in the height of a surface, the ratio of the real to the geometric surface area	(Friedman et al., 2012)
seabedforms	Seabed topography as measured by a combination of seabed position and slope	http://www.northeastoceandata.org/
Slp: 2, 10, and 20 km	Benthic slope at spatial scales of 2, 10, and 20 km	(Winship, 2018)
Slpslp: 2, 10, and 20 km	Benthic slope of slope at spatial scales of 2, 10, and 20 km	(Winship, 2018)
soft_sed	Soft sediments based on grain size distribution from USGS usSEABED: Atlantic coast offshore surficial sediment data	http://www.northeastoceandata.org/
Vort: fall (fa), spring (sp), summer (su), and winter (wi)	Benthic current vorticity at a spatial scale of 1/6 degree (approx. 19 km)	(Kinlan, 2016)

Biological covariates included predictor variables representing lower trophic level primary and secondary production. Primary production potential was represented by monthly chlorophyll concentration static variables extracted from remote-sensing data sources. The chlorophyll concentration data included measurements made with the Sea-viewing Wide Field of View Sensor (SeaWiFS), Moderate Resolution Imaging Spectroradiometer (MODIS) on the Aqua satellite, Medium Resolution Imaging Spectrometer (MERIS), and Visible and Infrared Imaging/Radiometer Suite (VIIRS) sensors during the period 1997–2016. These data were merged using the Garver–Siegel–Maritorena (GSM) algorithm (Maritorena et al., 2010) obtained from GlobColour (Hermes, hermes.acri.fr/index.php). As these were provided with the remote-sensing SST data, the monthly gradient magnitude was also calculated for the chlorophyll data, providing chlorophyll frontal fields.

Secondary production variables were based on zooplankton abundances measured by the Ecosystem Monitoring Program (EcoMon), which conducts shelf-wide bimonthly surveys of the Northeast US Shelf ecosystem (Kane, 2007). Zooplankton are collected obliquely through the water column to a maximum depth of 200 m using paired 61-cm Bongo samplers equipped with 333-micron mesh nets. Sample location in this survey is based on a randomized strata design, with strata defined by bathymetry and along-shelf location. Plankton taxa are sorted and identified to the lowest possible

TABLE 3 Summary of zooplankton predictor variables used in the development of spring and autumn occupancy habitat models

Variable name	Full name
acarspp	<i>Acartia</i> spp.
calfin	<i>Calanus finmarchicus</i>
chaeto	Chaetognatha
cham	<i>Centropages hamatus</i>
cirr	Cirripedia
ctyp	<i>Centropages typicus</i>
echino	Echinodermata
evadnespp	<i>Evadne</i> spp.
gas	Gastropoda
hyper	Hyperidea
larvaceans	Appendicularians
mlucens	<i>Metridia lucens</i>
oithspp	<i>Oithona</i> spp.
para	<i>Paracalanus parvus</i>
penilia	<i>Penilia</i> spp.
pseudo	<i>Pseudocalanus</i> spp.
salps	Salpa
tlong	<i>Temora longicornis</i>
volume	Plankton biovolume

taxonomic rank. We used the density estimates (number per 100 m³) of the 18 most abundant taxonomic categories and a biomass indicator (settled biovolume) as potential predictor variables (Table 3). The zooplankton time series has some values missing, and this was ameliorated by summing data over 5-year time steps for each seasonal period and interpolating a complete field using ordinary kriging. Thus, for example, the data for spring 2000 would include the available data from tows made during the period 1998–2002. These data are available to the public from the National Center for Environmental Information (<https://accession.nodc.noaa.gov/0187513>).

2.3 | Random forest models and variable importance analysis

Distribution models were developed using RF machine learning (Cutler et al., 2007), which were fitted using the R package RANDOMFOREST 4.6-14. RF models have been demonstrated to achieve comparable predictive power to other statistical methods (Smoliński &

Radtke, 2017). Prior to fitting the model, the independent variable set was tested for multicollinearity among the predictors, and variables were eliminated using the R package RFUTILITIES 2.1-5. From this reduced set of predictors (Table 1), the final model variables were selected using the model selection criteria of Murphy, Evans & Storfer (2010), as implemented in RFUTILITIES. This procedure repeatedly fits an RF model and removes the poorest performing variable until the minimum set of variables that produces the best fit is found. The accuracy of models was evaluated based on out-of-bag classification using the area under the curve (AUC) of the receiver operating characteristic (ROC) using the R package METRICS 0.1.4. The AUC has been widely used for testing the predictive ability of species distribution models (Guisan & Zimmermann, 2000). Values of AUC range from 0 to 1.0, with 0.50 considered to be random classification (Manel, Williams & Ormerod, 2001). Models with a minimum AUC of 0.65 were included in the analysis. To evaluate the importance of a variable in the occupancy models, six performance measures were considered: the number of times a variable was the root variable (i.e. the variable associated with the root node); the mean minimum node depth for the

TABLE 4 Random forest occupancy models, and area under the curve (AUC) statistic, for species captured in the Northeast Fisheries Science Center (NEFSC) spring bottom trawl survey

Species	FG	AUC	Species	FG	AUC
<i>Alosa aestivalis</i>	p	0.66	<i>Malacoraja senta</i>	b	0.77
<i>Alosa pseudoharengus</i>	p	0.79	<i>Melanogrammus aeglefinus</i>	b	0.81
<i>Amblyraja radiata</i>	dp	0.66	<i>Menidia menidia</i>	p	0.67
<i>Anchoa mitchilli</i>	p	0.72	<i>Merluccius albidus</i>	dp	0.87
<i>Cancer irroratus</i>	b	0.68	<i>Merluccius bilinearis</i>	dp	0.80
<i>Centropristis striata</i>	b	0.70	<i>Mustelus canis</i>	b	0.76
<i>Chlorophthalmus agassizi</i>	b	0.71	<i>Myoxocephalus octodecemspinosus</i>	b	0.90
<i>Citharichthys arctifrons</i>	b	0.80	<i>Paralichthys dentatus</i>	dp	0.82
<i>Clupea harengus</i>	p	0.72	<i>Paralichthys oblongus</i>	dp	0.84
<i>Dipturus laevis</i>	b	0.71	<i>Peprilus triacanthus</i>	p	0.84
<i>Enchelyopus cimbrius</i>	b	0.72	<i>Placopecten magellanicus</i>	b	0.80
<i>Gadus morhua</i>	dp	0.76	<i>Prionotus carolinus</i>	b	0.75
<i>Glyptocephalus cynoglossus</i>	b	0.84	<i>Pseudopleuronectes americanus</i>	b	0.85
<i>Helicolenus dactylopterus</i>	p	0.79	<i>Raja eglanteria</i>	b	0.75
<i>Hemirhamphys americanus</i>	dp	0.75	<i>Scomber scombrus</i>	p	0.69
<i>Hippoglossoides platessoides</i>	b	0.91	<i>Scophthalmus aquosus</i>	b	0.77
<i>Homarus americanus</i>	b	0.81	<i>Scyliorhinus retifer</i>	b	0.84
<i>Illex illecebrosus</i>	pp	0.83	<i>Sebastes fasciatus</i>	p	0.88
<i>Leucoraja erinacea</i>	b	0.84	<i>Squalus acanthias</i>	dp	0.81
<i>Leucoraja garmani</i>	b	0.65	<i>Stenotomus chrysops</i>	p	0.68
<i>Leucoraja ocellata</i>	dp	0.77	<i>Urophycis chesteri</i>	dp	0.65
<i>Limanda ferruginea</i>	b	0.82	<i>Urophycis chuss</i>	dp	0.81
<i>Loligo pealeii</i>	pp	0.87	<i>Urophycis regia</i>	dp	0.84
<i>Lophius americanus</i>	dp	0.74	<i>Urophycis tenuis</i>	dp	0.86
<i>Macrozoarces americanus</i>	b	0.72			

Note: FG are species functional groups: b, benthivores; dp, demersal piscivores; p, planktivores; pp, pelagic piscivores. Only models with an AUC of ≥ 0.65 are included in the analyses.

variable; a decrease in the Gini index of node impurity; a decrease in prediction accuracy; the proportion of models in which the variable was included; and whether the variable was among the 10 highest ranked variables. The first four of these indices were computed using the R package *RANDOMFORESTEXPLAINER* 0.10.0: as is usual, the number of times that a variable was a root was plotted against the mean minimum node depth for a variable and the decrease in Gini index was plotted against the decrease in accuracy. In addition, the performance of variables across functional group categories of benthivores, demersal piscivores, pelagic piscivores, and planktivores was also considered. With the number of comparisons made and the usefulness of plotting two variable importance measures together, these results were examined with a visualization of confidence intervals among variable classes.

Principal component analysis (PCA) was used to combine data from the individual importance measures into a single variable performance metric. The PCA did not include the proportion of models when the variable was included but did include all measures of the influence of a variable: number of times a root; mean minimum node depth; Gini index decrease; prediction accuracy decrease; and whether the variable was among the 10 highest ranked variables (scored as 0 or 1). The mean minimum node depth metric was multiplied by -1 , so that all measures could be interpreted as greater numbers being more influential in the models. All variables were standardized with a mean of zero and a standard deviation of one; the R packages *ADE4* 1.7-13 and *FACTOEXTRA* 1.0.5 were used for the PCA analysis. The first dimension of the PCA was initially negative, so it was multiplied by -1 to give the interpretation that greater values

TABLE 5 Random forest occupancy models, and area under the curve (AUC) statistic, for species captured in the Northeast Fisheries Science Center (NEFSC) autumn bottom trawl survey. FG are the species functional groups benthivores (b), demersal piscivores (dp), pelagic piscivores (pp), and planktivores (p); only those models with an AUC of ≥ 0.65 are included in the analyses

Species	FG	AUC	Species	FG	AUC
<i>Alosa aestivalis</i>	p	0.72	<i>Malacoraja senta</i>	b	0.74
<i>Alosa pseudoharengus</i>	p	0.80	<i>Melanogrammus aeglefinus</i>	b	0.82
<i>Amblyraja radiata</i>	dp	0.68	<i>Merluccius albidus</i>	dp	0.80
<i>Anchoa hepsetus</i>	p	0.81	<i>Merluccius bilinearis</i>	dp	0.80
<i>Anchoa mitchilli</i>	p	0.80	<i>Micropogonias undulatus</i>	b	0.87
<i>Cancer irroratus</i>	b	0.68	<i>Mustelus canis</i>	b	0.82
<i>Centropristis striata</i>	b	0.74	<i>Myoxocephalus octodecemspinosus</i>	b	0.86
<i>Chlorophthalmus agassizi</i>	b	0.73	<i>Ovalipes ocellatus</i>	b	0.67
<i>Citharichthys arctifrons</i>	b	0.79	<i>Paralichthys dentatus</i>	dp	0.88
<i>Clupea harengus</i>	p	0.90	<i>Paralichthys oblongus</i>	dp	0.82
<i>Cynoscion regalis</i>	pp	0.88	<i>Peprilus triacanthus</i>	p	0.75
<i>Dipturus laevis</i>	b	0.67	<i>Placoepecten magellanicus</i>	b	0.84
<i>Enchelyopus cimbrius</i>	b	0.69	<i>Pollachius virens</i>	dp	0.65
<i>Gadus morhua</i>	dp	0.79	<i>Pomatomus saltatrix</i>	pp	0.75
<i>Glyptocephalus cynoglossus</i>	b	0.87	<i>Prionotus carolinus</i>	b	0.78
<i>Helicolenus dactylopterus</i>	p	0.77	<i>Prionotus evolans</i>	dp	0.78
<i>Hemitripterus americanus</i>	dp	0.72	<i>Pseudopleuronectes americanus</i>	b	0.87
<i>Hippoglossoides platessoides</i>	b	0.91	<i>Raja eglanteria</i>	b	0.76
<i>Homarus americanus</i>	b	0.77	<i>Scomber scombrus</i>	p	0.65
<i>Illex illecebrosus</i>	pp	0.81	<i>Scophthalmus aquosus</i>	b	0.84
<i>Leiostomus xanthurus</i>	b	0.86	<i>Scylliorhinus retifer</i>	b	0.82
<i>Lepophidium profundorum</i>	b	0.72	<i>Sebastes fasciatus</i>	p	0.93
<i>Leucoraja erinacea</i>	b	0.85	<i>Squalus acanthias</i>	dp	0.81
<i>Leucoraja garmani</i>	b	0.77	<i>Stenotomus chrysops</i>	p	0.87
<i>Leucoraja ocellata</i>	dp	0.86	<i>Urophycis chesteri</i>	dp	0.70
<i>Limanda ferruginea</i>	b	0.81	<i>Urophycis chuss</i>	dp	0.83
<i>Loligo pealeii</i>	pp	0.85	<i>Urophycis regia</i>	dp	0.83
<i>Lophius americanus</i>	dp	0.75	<i>Urophycis tenuis</i>	dp	0.88
<i>Macrozoarces americanus</i>	b	0.65	<i>Zenopsis conchifera</i>	p	0.70

Note: FG are species functional groups: b, benthivores; dp, demersal piscivores; p, planktivores; pp, pelagic piscivores. Only models with an AUC of ≥ 0.65 are included in the analyses.

represent more influential variables. The first dimension resulting from the PCA was used as a dependent variable in linear mixed models for the spring and autumn seasons using the R package *NMLE* 3.1-142. For the mixed model, the random effects were specified as species nested within functional groups and the fixed effect was the variable class (physical, primary production, secondary production, or benthic terrain). Models were initially evaluated with a likelihood ratio test, then all contrasts were tested with an adjusted Tukey multiple comparisons test to determine differences ($\alpha = 0.05$) using the R package *LSMEANS* 2.30-0.

3 | RESULTS

3.1 | Model fits

For the models based on spring data, 49 fish and macroinvertebrate occupancy models met the minimum standard of an AUC score of at least 0.65 for consideration as successfully fitted (Table 4). Overall, these models averaged an AUC score of 0.78. These species included finfish and macroinvertebrates from various functional groups, including benthivores, demersal piscivores, pelagic piscivores, and

planktivores. For the autumn data, 58 taxa were found to have sufficient classification performance for inclusion in the analysis (Table 5). These models averaged an AUC score of 0.79. A total of 48 taxa models were common to both seasons, whereas one taxa was unique to spring and 10 taxa were unique to autumn.

3.2 | Importance of variable classes

The different variable classes contributed to occupancy models in a hierarchical fashion when considering the performance measures of number of times a root, mean minimum node depth, Gini index decrease, and accuracy decrease. Physical and biological variables had a larger influence on the model fits than benthic terrain variables (Figure 2). To provide context, variable classes with a higher average number of times a root and lower mean minimum node depth score (in the upper left quadrant of a plot panel, Figure 2) are indicative of variables that are more important. Likewise, variable classes with large decreases in Gini index and accuracy values (in the upper right quadrant of a plot panel, Figure 2) are indicative of variables that are more important for explaining the variation in occupancy of a species. The spring models suggest that physical, primary production, and

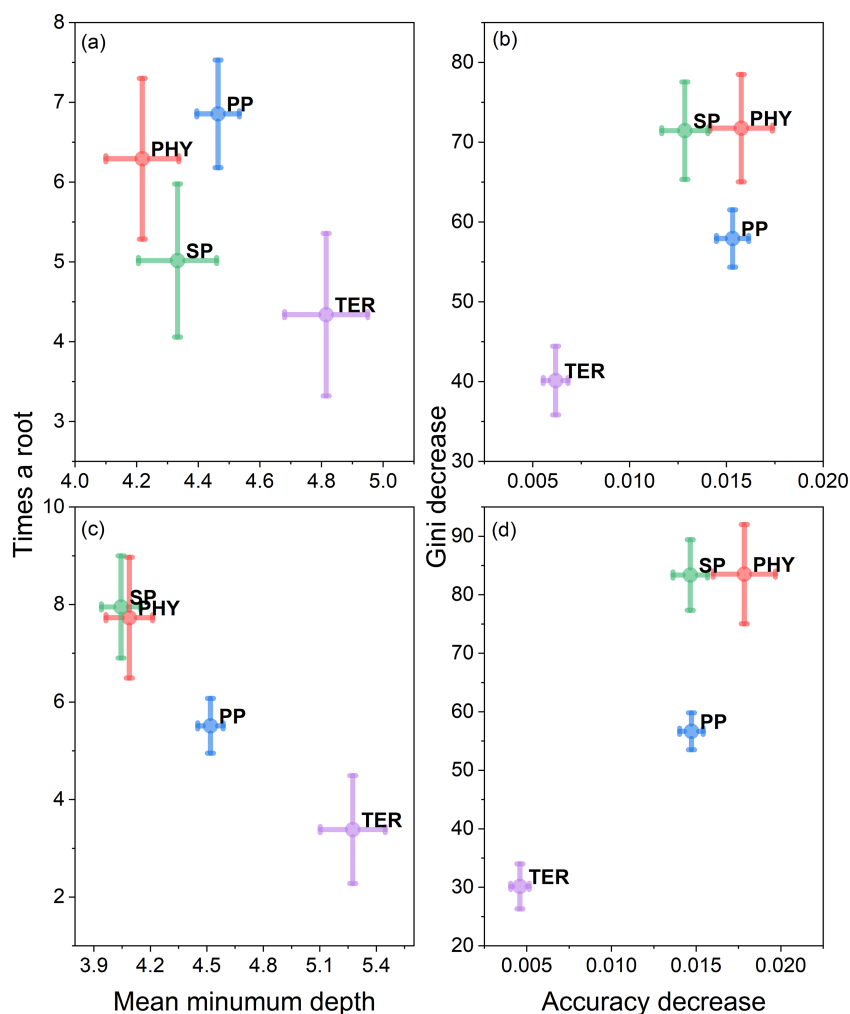
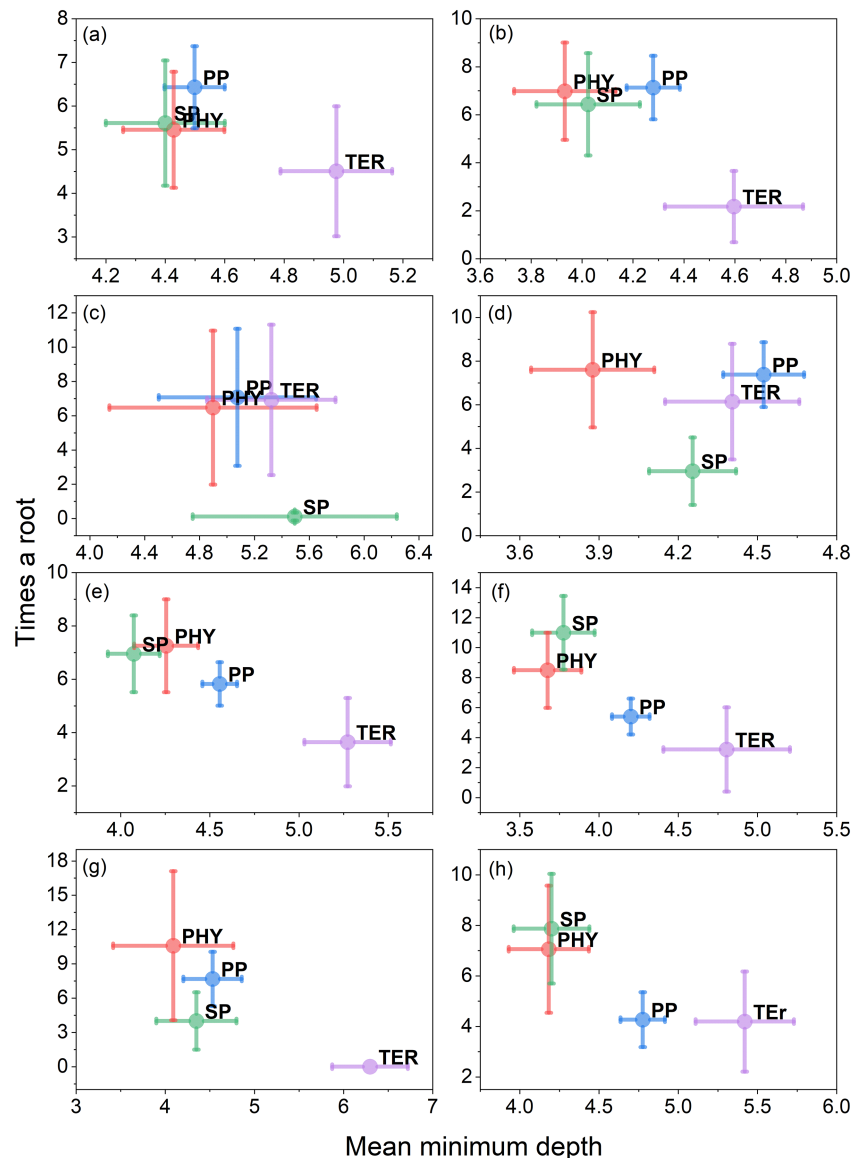


FIGURE 2 The mean number of times a variable was the root node variable versus the mean minimum node depth of a variable in a tree for spring (a) and autumn (c) models. Variables are grouped by their class: PHY, physical; PP, primary production; SP, secondary production; and TER, benthic terrain complexity. The mean decrease in the Gini index of node impurity versus the mean accuracy decrease if a variable were to be removed in spring (b) and autumn models (d), presented by variable class. Error bars are 95% confidence intervals. Physical, primary production, secondary production, and terrain variables are represented in red, blue, green, and purple, respectively

FIGURE 3 The mean number of times a variable was the root node variable versus the mean minimal depth of a variable in a tree for spring models for benthivores (a), demersal piscivores (b), pelagic piscivores (c), and planktivores (d), by variable classes: PHY, physical; PP, primary production; SP, secondary production; and TER, benthic terrain complexity. Data for autumn models shown in panels (e)–(h) for the same taxonomic groups, respectively. Error bars are 95% confidence intervals. Physical, primary production, secondary production, and terrain variables are represented in red, blue, green, and purple, respectively



secondary production variable classes were more important than benthic habitat variables (Figure 2a,b). For the autumn models, however, physical and secondary production variables were the most important variables, suggesting a less important role for primary production variables than in the spring (Figure 2c,d). For both sets of seasonal models, the terrain variables made the lowest contribution.

The contribution of variable classes varied with the functional grouping of species models by season and by performance measure. For spring, benthivores and demersal piscivore models were influenced by physical and primary and secondary production variables, and less so by benthic terrain variables (Figures 3a,b and 4a,b). The variable hierarchy for pelagic piscivores was slightly different, with secondary production variables appearing to play less of a role (Figures 3c and 4c). The variable contribution for planktivores was distinctly different from the other functional groups, however, with the dominant variable class appearing to be physical variables (Figures 3d and 4d). For autumn, benthivores and demersal piscivore models were also shaped by the contribution of physical and primary

and secondary production variables, and less so by terrain variables (Figures 3e,f and 4e,f). The variable hierarchy for pelagic piscivores in autumn was different from the spring variable contribution, with secondary production variables appearing to play more of a role (Figures 3g and 4g). The variable contribution for autumn planktivores was dominated by physical and secondary production variables (Figures 3f and 4f).

The inclusion of variables differed by variable class and were similar between seasons. The proportion of models that included a primary production variable was approximately 70% in both spring and autumn models (Figure 5a). Physical and secondary production variables appeared in approximately 50–60% of models in both seasons; however, terrain variables only appeared in approximately 25% of the models. The top 10 most prevalent variables across all models revealed a slightly different pattern. Individual physical and secondary production variables tended to occur in models as a top-10 variable at rates exceeding 20% of species- and season-specific models, whereas individual primary production variables were among

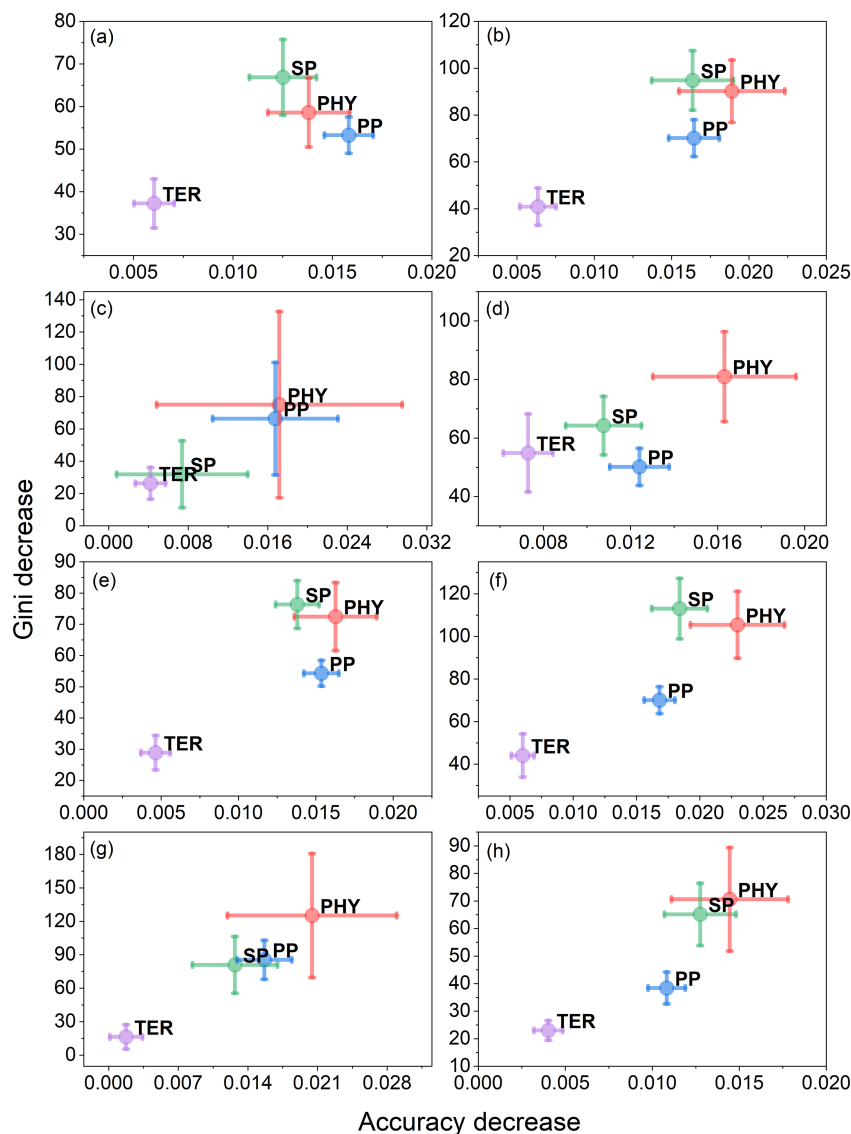


FIGURE 4 The mean decrease in Gini index of node impurity versus the mean accuracy decrease if a variable were to be removed in spring models for benthivores (a), demersal piscivores (b), pelagic piscivores (c), and planktivores (d), by variable classes: PHY, physical; PP, primary production; SP, secondary production; and TER, benthic terrain complexity. Data for autumn models are shown in panels (e)–(h) for the same taxonomic groups, respectively. Error bars are 95% confidence intervals. Physical, primary production, secondary production, and terrain variables are represented in red, blue, green, and purple, respectively

the top 10 variables approximately 15% of the time (Figure 5b). Individual terrain variables were among the top 10 variables only 3–6% of the time. The pattern in the occurrence of variables in the models is also reflected in the inclusion of variables among functional groups of taxa. In both spring and autumn models, the proportion of models in which a variable occurred across variable and functional groups (Figure 6a,b) reflected the overall patterns seen for all species (Figure 5a). Likewise, the pattern in the occurrence of a variable as a top 10 variable across variable classes and functional groups in both spring and autumn models (Figure 7a,b) reflected the overall patterns seen for all species (Figure 5b).

3.3 | Importance of individual variables

Some individual variables emerge repeatedly among the most important variables for given performance metrics. In spring, chlorophyll concentration, water depth, and the abundance of the zooplankton taxon *Paracalanus parvus* were the top-performing variables in one or

more metrics (Figure 8). Across all indicators, 43% of the variables were primary production variables, 33% were physical variables, and 23% were secondary production variables. The highest frequency of occurrence as a root variable was chlorophyll concentration in July (Figure 8a); this variable was also among the most frequently occurring variable across all models (Figure 8e). Water depth was the top-performing variable (Figure 8b–d) among mean minimum node depth, Gini index decrease, and accuracy decrease indices, and was also among the most frequently occurring variables across all models. *Paracalanus parvus* and water depth were the two most frequently occurring variables among the top 10 (Figure 8f). In autumn, water depth was the top-performing variable in five of the six indices (Figure 9). Across all indices, 20% of the variables were primary production variables, 28% were physical variables, and 52% were secondary production variables, which was in contrast to the proportion of variable classes in the spring models. Bottom temperature was the most frequent variable among the top 10 variables (Figure 9f) and among the top three variables over all indices. The most important zooplankton variable was the copepod *Metridia lucens*, which

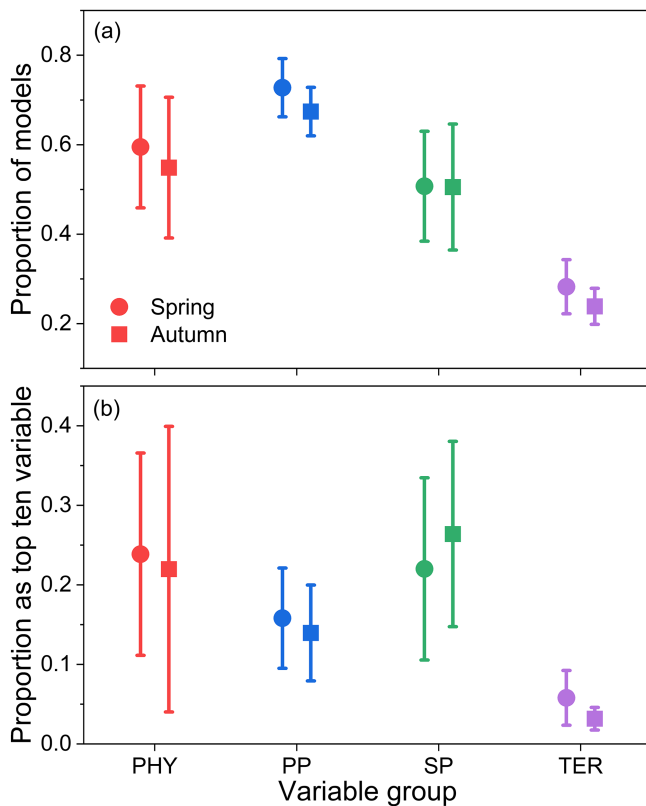


FIGURE 5 The proportion of spring and autumn models in which a variable was present (a) from variable classes: PHY, physical; PP, primary production; SP, secondary production; and TER, benthic terrain complexity. The proportion of spring and autumn models in which a variable was among the top 10 model variables present (b) from the same variable classes. Error bars are 95% confidence intervals. Physical, primary production, secondary production, and terrain variables are represented in red, blue, green, and purple, respectively

appeared in nearly all of the models (Figure 9e) and was among the top four variables in all indices.

3.4 | Variable importance across performance measures

The first dimension of the PCA (PCA1) had an eigenvalue of 3.51 and explained 70.2% of the variance in variable importance indices. The second dimension had an eigenvalue of 0.64 and explained 12.8% of the variance; this dimension was not considered further, as eigenvalues of <1 are generally not informative. The loadings on PCA1 were similar across variables: number of times a root (17.4%); mean minimum node depth (22.2%); Gini index decrease (22.4%); prediction accuracy decrease (21.2%); and whether the variable was among the 10 highest ranked variables (16.9%). The linear mixed model showed the combined measure of variable importance (PCA1) differed by variable class in the autumn ($F_{3, 1752} = 45.7, P < 0.0001$) and spring ($F_{3, 1568} = 10.2, P < 0.0001$). The contrasts showed consistent results for the spring and autumn (Tables 6 and 7). Most notably,

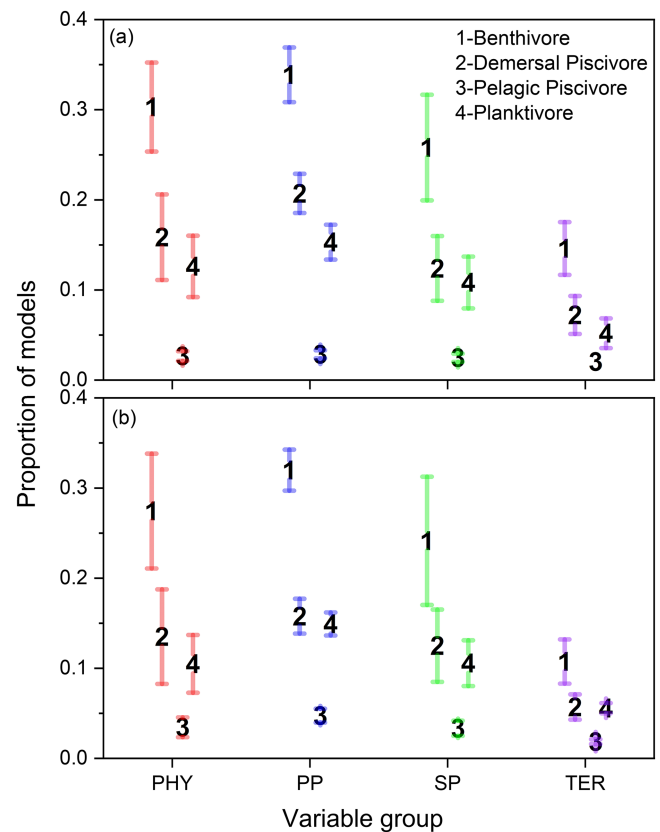


FIGURE 6 The proportion of spring (a) and autumn (b) models in which a variable was present from variable classes physical (PHY), primary production (PP), secondary production (SP), and terrain complexity (TER), by functional group. Error bars are 95% confidence intervals. Physical, primary production, secondary production, and terrain variables are represented in red, blue, green, and purple, respectively

secondary production variables had more influence than primary production and terrain variables. Physical variables were also more influential than primary production and terrain variables. No statistical difference was observed between the influence of secondary production and physical variables. The PCA scores of the individual variables showed that the physical variables of water depth and bottom temperature were among the most influential in both seasons (Figure 10). The secondary production variables of zooplankton species abundance were commonly among the most influential variables, but individual species of importance often differed between the spring and autumn. To summarize, our analysis suggests that two physical environment variables, bottom temperature and water depth, play a pre-eminent role across most SDMs; however, many SDMs were also informed by the primary and secondary production variables.

4 | DISCUSSION

Species distribution models (SDMs) provide useful information about the spatial ecology of resource and protected species, the value of

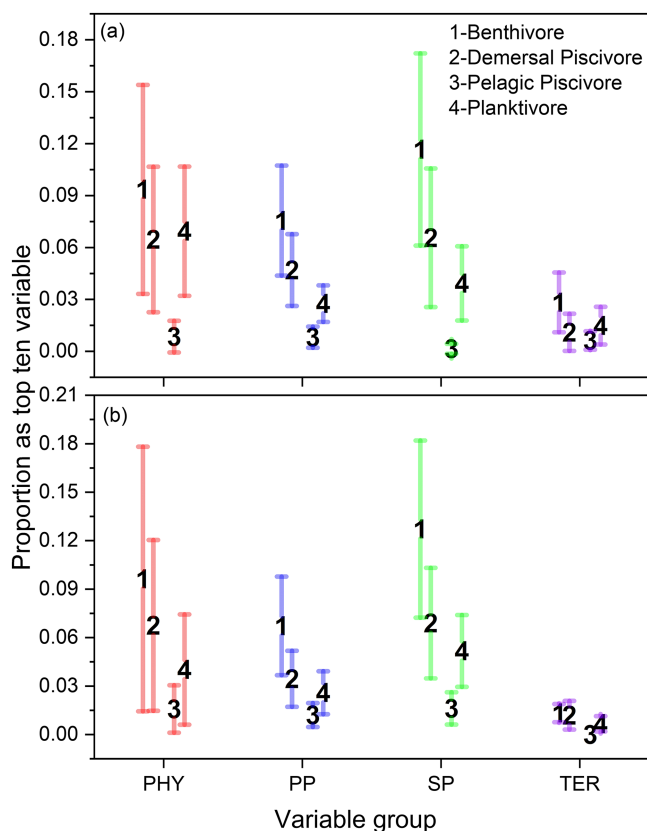


FIGURE 7 The proportion of spring (a) and autumn (b) models in which a variable was among the top 10 model variables present from variable classes physical (PHY), primary production (PP), secondary production (SP), and terrain complexity (TER), by functional group. Error bars are 95% confidence intervals. Physical, primary production, secondary production, and terrain variables are represented in red, blue, green, and purple, respectively

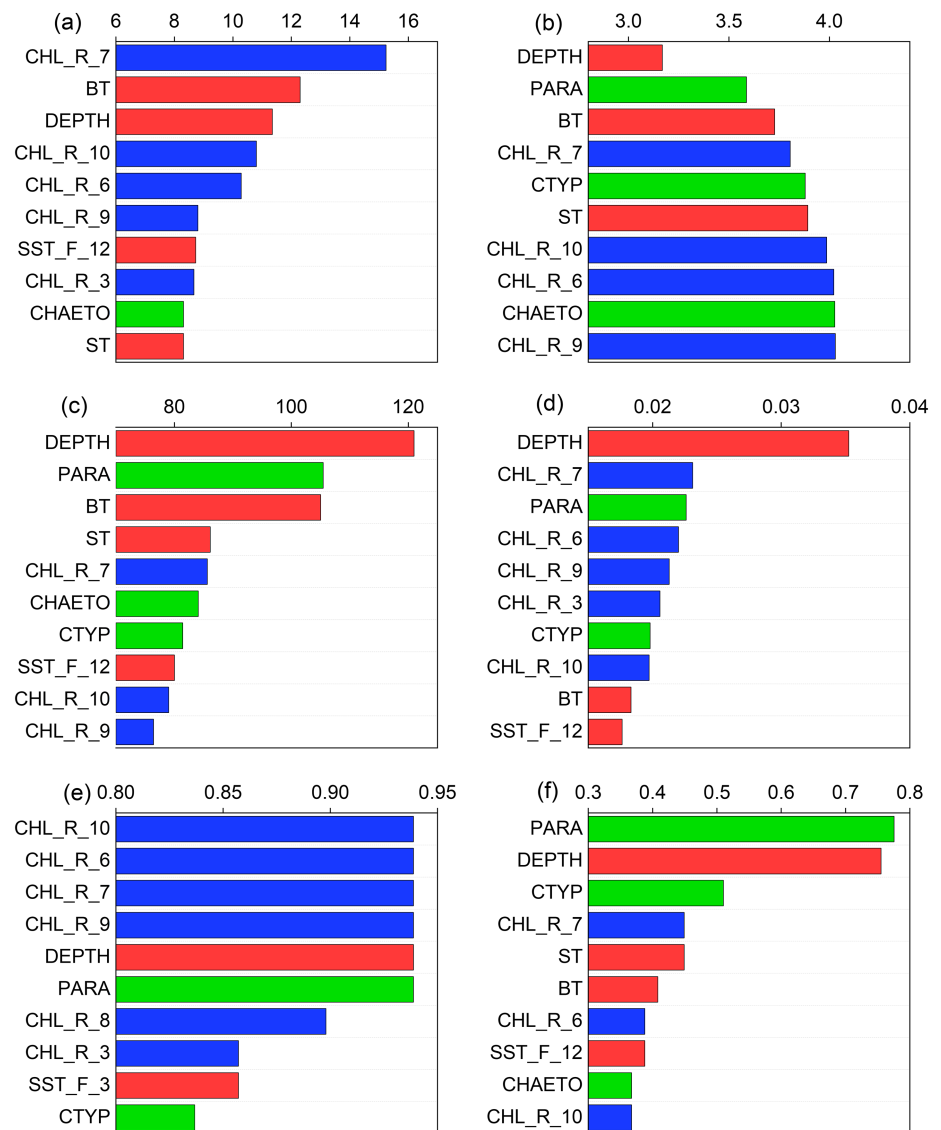
which is enhanced when we consider the processes underlying their construction. Machine learning provides the means to evaluate a wide range of factors that potentially define the niche of a species. We concentrated on this aspect of SDM development to ask which variable classes and individual variables made the most substantial contributions to the SDMs for a range of marine fish and macroinvertebrates. Although SDMs are correlative models, important ecological insights can be gained from comparing a broad suite of predictor variables across species. We found that variables related to the physical environment and secondary productivity played key roles in defining the distributions of species, with a lesser role for primary productivity and benthic terrain variables. The most important physical variables were water depth and bottom temperature, both of which are key in defining the biophysical habitats occupied by marine organisms (McGowan, Horne & Rogers, 2019; Rubec et al., 2019). The role of temperature in the growth and reproduction of marine species is manifold (Loisel, Isla & Daufresne, 2019), and is reinforced by the effect of temperature to define distribution limits based on thermal tolerances (Norin et al., 2019). The inclusion of temperature in habitat modelling for NES shelf species has become increasingly important

given the rapid, recent increases in temperatures experienced in this ecosystem (Pershing et al., 2015) and its projected temperature change into the future (Saba et al., 2016). Depth is most likely functioning as a proxy variable as it is related to clines in temperature, salinity, light penetration, and other factors. Depth and other abiotic or geographic variables have been incorporated into modelling efforts to support the prediction of thermal distribution and document species shifts within the region (Kleisner et al., 2016; Kleisner et al., 2017; Johnson, Thorson & Punt, 2019).

Variables associated with the distribution of secondary productivity were of high importance in model construction. The zooplankton species differed in importance by season, highlighting their differing seasonal phenology. The individual taxa *Paracalanus parvus* and *Metridia lucens* played important roles in spring and autumn models, respectively. The critical role of zooplankton in the life history of fish and macroinvertebrates may be based on direct dependence through predation, but it may also be related to what zooplankton represent in terms of the distribution of primary production and energy flow (Druon et al., 2019). Zooplankton are known to concentrate in areas of high primary production (i.e. chlorophyll concentration) and along water column features like fronts (Labat et al., 2009), both of which are associated with a more active flux of energy to the benthos (Olli et al., 2002). Our results are consistent with cross-system comparisons suggesting that bottom-up processes are dominating the mechanisms controlling fish productivity (Iverson, 1990). For example, strong bottom-up trophic linkages were observed in the north-east Pacific, suggesting that enhanced primary production is channelled up the food web to the resident fish community (Ware & Thomson, 2005). For this reason, zooplankton occurrence can serve as an indicator of the biologically productive zones that are often located between frontal and stratified areas (Maravelias & Reid, 1997). Aggregations of zooplankton, micronekton, and fish in marine systems are frequently observed in such areas of high hydrographic activity and strong vertical mixing (Genin, 2004). Few of the fish and macroinvertebrates represented by the models in our study directly prey on phytoplankton, and therefore zooplankton could simply be a superior predictor because it is a better metric for the productivity of the ecosystem. The associations observed among zooplankton and fish species suggest that more effort is warranted to understand zooplankton distribution and populations.

The low number of influential terrain variables in this set of models should be interpreted in the context of the spatial scale of the response variable and resultant niche estimates. The relationship between marine species and the nature of the substrate has significant foundation. Demersal fish are known to have preferences related to surficial sediments (Rau et al., 2019) and may be associated with heterogeneous depths (Manderson et al., 2011; Baker et al., 2019). The bottom trawl survey integrates captures over a trawl path that can exceed 1 km, which may cover diverse microhabitats often associated with multiple habitat preferences (Amorim et al., 2018). There is also a lack of shelf-wide sediment data at a fine-scale resolution that may be required to tease out meaningful benthic relationships. Thus, the lack of influential terrain variables is also likely to represent a

FIGURE 8 The top 10 variables for spring models selected by performance metrics: number of times a root (a); mean minimum node depth (b); Gini index decrease (c); accuracy decrease (d); proportion of models (e); and proportion as top-10 variable (f). Physical, primary production, and secondary production variables are represented in red, blue, and green, respectively



deficiency in the environmental data available. Moreover, to detect the effect of a predictor variable, it is important that observations cover the entire range of environmental conditions (Florin, Sundblad & Bergstrom, 2009; Wenger & Olden, 2012). Bottom trawl surveys are typically limited to trawlable substrates; habitats with highly complex bottom structure can be permanently under-represented in the bottom trawl data sets (Smoliński & Radtke, 2017). Therefore, the full range of environmental variables describing habitat complexity may not be captured in the training data, causing biased results and underestimating the importance of habitat complexity in the predictive models developed. At these spatial resolutions terrain variables appear to be less effective predictors than physical and lower trophic level variables. The use of a different modelling approach and data at different spatial and temporal resolutions could yield different results with respect to the relative importance of variables. The general hierarchy of variable importance extended to most of the results was disaggregated by functional group, with the exception of pelagic piscivores. By some measures, terrain variables were of greater importance for pelagic piscivores and, overall, zooplankton variables

appeared to play a lesser role compared with models for the other functional groups. We do not ascribe much significance to these differences as there were relatively few pelagic piscivores among the species modelled. In part, this is related to the low efficiency of the benthic trawl with regards to pelagic piscivores.

There are many resource and environmental laws and regulations in force in the NES ecosystem that would be informed by SDM information. The Magnuson-Stevens Fishery Conservation and Management Act (MSA) mandates that essential fish habitat (EFH) be described, identified, and mapped for the nearly 1,000 species managed under federal fishery management plans. The EFH provisions cover all life-history stages from eggs and larvae to settled juveniles and adult stages, and the National Marine Fisheries Service (NMFS) is mandated to work with other agencies to determine what actions (in addition to fishing) may impact designated EFH. Some contemporary NES examples of activities where EFH may be impacted include wind farm construction and operations, offshore oil and gas exploration and production, proposed changes to shipping lanes and vessel traffic, the installation of submarine cables, and sand mining. SDMs

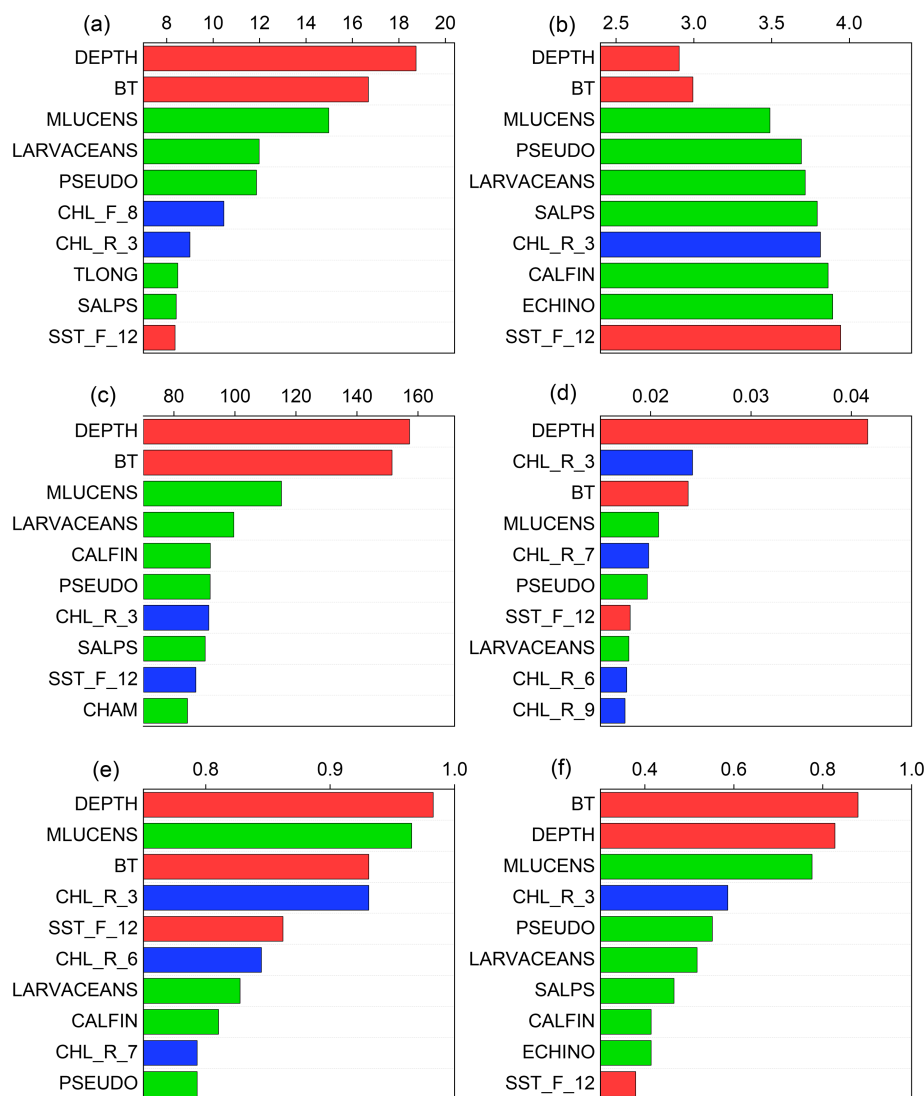


FIGURE 9 The top 10 variables for autumn models selected by performance metrics: number of times a root (a); mean minimum node depth (b); Gini index decrease (c); accuracy decrease (d); proportion of models (e); and proportion as top-10 variable (f). Physical, primary production, and secondary production variables are represented in red, blue, and green, respectively

TABLE 6 By variable type, mean ± 1 SE of the first axis of the principal components analysis describing five measures of variable influence

Season	Variable type	Mean PCA1	\pm SE of PCA1
Spring	PHY	0.66	0.19
Spring	SP	0.51	0.2
Spring	PP	0.25	0.19
Spring	TER	0.11	0.21
Autumn	PHY	1.08	0.19
Autumn	SP	0.94	0.19
Autumn	PP	0.25	0.18
Autumn	TER	0.03	0.21

Note: Variables types: PHY, physical; PP, primary production; SP, secondary production; TER, benthic terrain complexity. Higher values represent more influential variables.

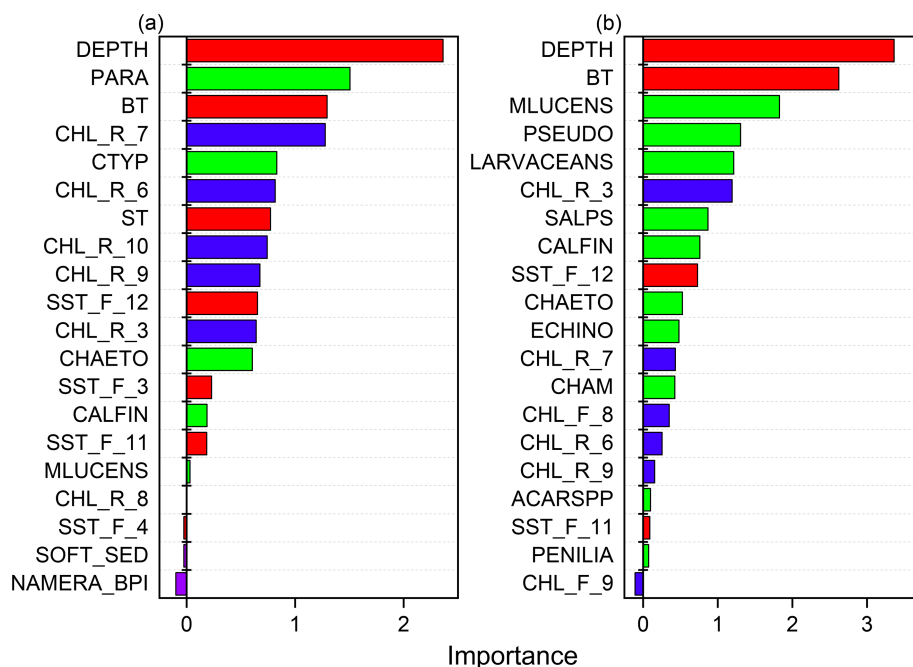
have been used to inform and update EFH designations, and there is interest in increasing the role of SDMs in EFH designation work, including in the NES. In Alaska, Laman et al. (2018) used maximum

entropy and generalized additive models (GAMs) to fit SDMs to a variety of physical variables as well as ocean colour data. SDMs were used to understand the relationship between adult and nursery habitats for *Lutjanus campechanus* (red snapper) in the Gulf of Mexico (Dance & Rooker, 2019), and SDMs have also been used to parameterize ecosystem models (Gruss et al., 2018). Although many of these examples highlight the importance of the physical environment contributing to the distribution of species, our results suggest that ecosystem-level processes such as primary and secondary production may factor more heavily into where species occur than would be suggested by much of the literature. In particular, the effects of thermohaline and tidal fronts may act to aggregate zooplankton and enhance the flux of energy to the benthos, leading to increased production of the species modelled in this study. By holistically incorporating the various habitat components essential to the life histories of these species, the resulting SDMs may provide more advanced predictions of species abundance and distribution than models including only traditional covariates. In an ecosystem-based management context, these SDMs may serve as the ideal tool for best conserving NES fish and invertebrate species in a system where

TABLE 7 Post-hoc Tukey test results from a linear mixed model of variable influence represented as the first dimension of a principal components analysis

Season	Contrasts	β	SE	df	T ratio	P value	Interpretation
Spring	PHY vs PP	0.41	0.09	1,568	4.6	<0.0001	PHY > PP
	PHY vs SP	0.15	0.10	1,568	1.4	0.5	PHY = SP
	PHY vs TER	0.55	0.13	1,568	4.2	<0.0001	PHY > TER
	PP vs SP	-0.26	0.09	1,568	-2.9	0.023	SP > PP
	PP vs TER	0.14	0.12	1,568	1.2	0.65	PP = TER
	SP vs TER	0.40	0.13	1,568	3.1	0.012	SP > TER
Autumn	PHY vs PP	0.83	0.09	1752	9.1	<0.0001	PHY > PP
	PHY vs SP	0.14	0.10	1752	1.3	0.53	PHY = SP
	PHY vs TER	1.05	0.14	1752	7.7	<0.0001	PHY > TER
	PP vs SP	-0.70	0.09	1752	-8.1	<0.0001	SP > PP
	PP vs TER	0.22	0.13	1752	1.7	0.32	PP = TER
	SP vs TER	0.91	0.13	1752	6.8	<0.0001	SP > TER

Note: Contrasts indicate differences in influence among predictor variable types. Variables types: PHY, physical; PP, primary production; SP, secondary production; TER, benthic terrain complexity.

FIGURE 10 The top 10 variables across all spring (a) and autumn (b) models based on PC1 variable scores as an index of importance. Physical, primary production, secondary production, and terrain variables are represented in red, blue, green, and purple, respectively

anthropogenic activities are increasing and coexistence with the ecosystem is imperative.

There are spatial and temporal strengths and weaknesses in the approach that we have used. As these data are based on a specific region and period, models informed with different data will likely yield different results. The biological observations used in this study were constrained to the spring and autumn, potentially limiting the generalizability of the results over the course of the full year (Yates et al., 2018). Variable rankings may not fully express the importance of environmental parameters during seasonal migrations or other times of the year. Moreover, the importance of the predictors may be expected to differ at different spatial scales, grid sizes, or extents of

the study (Sandman et al., 2013; Bennett, 2014). Potentially, the predictive power of variables can also differ between different areas of a species' range as a result of differing environmental conditions (Pickens & King, 2014) as well as local differences in phenotypic plasticity, local modifications of a species' realized niche, or existence of ecotypes (Randin et al., 2006). Further studies involving spatial cross-validation (Brenning, 2012) and parallel assessment of the variable predictive power for different regions within the north-east US continental shelf (e.g. Gulf of Maine and Georges Bank) could be of great importance for the assessment of the transferability of the model (Wenger & Olden, 2012) and a deeper understanding of the mechanisms driving species distribution. A synthesis of the SDM

literature demonstrated consistency in variable importance for a given species in different ranges by analysing results for species that were modelled in multiple regions or at different times, however (Bradie & Leung, 2017). Therefore, the utility of SDMs to provide information about the important predictors of species distributions across diverse regions and times is well supported.










Standard methods for computing variable importance measures using out-of-bag estimates are implemented in the `RANDOMFOREST` package and provide an efficient tool for screening variables based on predictive ability (Smoliński, 2018). These estimates of importance can be biased towards correlated variables (Strobl et al., 2008) or can be unreliable when variables vary in their scale of measurement or number of categories, however (Strobl et al., 2007). We believe that this is not the case here, because our dataset combines continuous variables only, and no substantial multicollinearity between variables was indicated during the initial analysis. Moreover, six various variable importance measures were calculated using the `RANDOMFORESTEXPLAINER` package to make the results more independent of the criteria applied. The high similarity of the results provided by different measures investigated in the initial phase of the analysis and a high proportion of variance explained by the first dimension of the PCA indicated a reasonable consistency of the importance estimates and the rankings of the variables in the specific models.

Although species distribution models have been developed for many NES taxa, machine learning methods are new for this system and may provide a better understanding of the influence of variables on the niche definitions for multiple species. Our predictor dataset for the NES is exceptionally rich and allows us to test approaches to optimize the use of a wider range of variables in marine SDMs. Clearly, this produces a dilemma for practitioners, as they must consider whether to accept models where the mechanistic relationship between predictor and response is unknown; however, exploratory approaches have the potential to provide novel insights into species biology and ecology that may be overlooked otherwise. This may be particularly challenging with variables such as chlorophyll concentration or zooplankton abundance, which may be functioning primarily as indicators. It would be a false argument to suggest that simpler models based on variables such as temperature and depth have any truer mechanistic foundation, however, especially in understudied species. Although there are certainly undescribed underlying dynamics that we are not able to characterize, this study and the associated results for individual species, can serve as a point of departure for further field research and modelling studies to elucidate the mechanistic relationships between variables and marine species.

ACKNOWLEDGEMENTS

We thank the North Atlantic Regional Team (NART) for supporting the discussion that led to the completion of this work and J. Thorson for useful comments. B. Pickens was supported by interagency agreement #M17PG00028 between NOAA and the Bureau of Ocean Energy Management. The views expressed in this article are those of the authors and do not necessarily represent the views of their agencies.

ORCID

Kevin D. Friedland  <https://orcid.org/0000-0003-3887-0186>
 Michelle Bachman  <https://orcid.org/0000-0001-7593-0331>
 Andrew Davies  <https://orcid.org/0000-0002-2087-0885>
 Romain Frelat  <https://orcid.org/0000-0002-8631-4398>
 M. Conor McManus  <https://orcid.org/0000-0003-3504-0371>
 Ryan Morse  <https://orcid.org/0000-0002-0854-2723>
 Bradley A. Pickens  <https://orcid.org/0000-0003-3643-6343>
 Szymon Smoliński  <https://orcid.org/0000-0003-2715-984X>
 Kisei Tanaka  <https://orcid.org/0000-0002-1901-6972>

REFERENCES

- Amorim, E., Ramos, S., Elliott, M. & Bordalo, A.A. (2018). Dynamic habitat use of an estuarine nursery seascape: Ontogenetic shifts in habitat suitability of the European flounder (*Platichthys flesus*). *Journal of Experimental Marine Biology and Ecology*, 506, 49–60. <https://doi.org/10.1016/j.jembe.2018.05.011>
- Austin, M.P. & Van Niel, K.P. (2011). Improving species distribution models for climate change studies: Variable selection and scale. *Journal of Biogeography*, 38(1), 1–8. <https://doi.org/10.1111/j.1365-2699.2010.02416.x>
- Baker, M.R., Pålsson, W., Zimmermann, M. & Rooper, C.N. (2019). Model of trawlable area using benthic terrain and oceanographic variables-Informing survey design and habitat maps in the Gulf of Alaska. *Fisheries Oceanography*, 28(6), 629–657. <https://doi.org/10.1111/fog.12442>
- Belkin, I.M. & O'Reilly, J.E. (2009). An algorithm for oceanic front detection in chlorophyll and SST satellite imagery. *Journal of Marine Systems*, 78(3), 319–326. <https://doi.org/10.1016/j.jmarsys.2008.11.018>
- Bennett, J.R. (2014). Comparison of native and exotic distribution and richness models across scales reveals essential conservation lessons. *Ecography*, 37(2), 120–129. <https://doi.org/10.1111/j.1600-0587.2013.00393.x>
- Bradie, J. & Leung, B. (2017). A quantitative synthesis of the importance of variables used in MaxEnt species distribution models. *Journal of Biogeography*, 44(6), 1344–1361. <https://doi.org/10.1111/jbi.12894>
- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32. <https://doi.org/10.1023/A:1010933404324>
- Brenning, A. (2012). Spatial Cross-Validation and Bootstrap for the Assessment of Prediction Rules in Remote Sensing: The R Package *Sprrorest*. *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, 2012, 5372–5375. <https://doi.org/10.1109/IGARSS.2012.6352393>
- Cao, J., Chen, Y. & Richards, R.A. (2017). Improving assessment of *Pandalus* stocks using a seasonal, size-structured assessment model with environmental variables. Part II: Model evaluation and simulation. *Canadian Journal of Fisheries and Aquatic Sciences*, 74(3), 363–376. <https://doi.org/10.1139/cjfas-2016-0021>
- Chu, J.W.F., Nephin, J., Georgian, S., Knudby, A., Rooper, C. & Gales, K.S.P. (2019). Modelling the environmental niche space and distributions of cold-water corals and sponges in the Canadian Northeast Pacific Ocean. *Deep-Sea Research Part I: Oceanographic Research Papers*, 151. <https://doi.org/10.1016/j.dsr.2019.06.009>
- Coops, N.C., Wulder, M.A. & Iwanicka, D. (2018). Demonstration of a satellite-based index to monitor habitat at continental-scales. *Ecological Indicators*, 9(5), 948–958. <https://doi.org/10.1016/j.ecolind.2008.11.003>
- Cord, A.F., Meentemeyer, R.K., Leitao, P.J. & Vaclavik, T. (2013). Modelling species distributions with remote sensing data: Bridging disciplinary perspectives. *Journal of Biogeography*, 40(12), 2226–2227. <https://doi.org/10.1111/jbi.12199>

- Dance, M.A. & Rooker, J.R. (2019). Cross-shelf habitat shifts by red snapper (*Lutjanus campechanus*) in the Gulf of Mexico. *PLoS ONE*, 14(3), e0213506. <https://doi.org/10.1371/journal.pone.0213506>
- Druon, J.N., Helaouet, P., Beaugrand, G., Fromentin, J.M., Palialexis, A. & Hoepffner, N. (2019). Satellite-based indicator of zooplankton distribution for global monitoring. *Scientific Reports*, 9, 4732. <https://doi.org/10.1038/s41598-019-41212-2>
- Elith, J. & Leathwick, J.R. (2009). Species Distribution Models: Ecological Explanation and Prediction Across Space and Time. *Annual Review of Ecology Evolution and Systematics*, 40, 677–697. <https://doi.org/10.1146/annurev.ecolsys.110308.120159>
- Florin, A.B., Sundblad, G. & Bergstrom, U. (2009). Characterisation of juvenile flatfish habitats in the Baltic Sea. *Estuarine Coastal and Shelf Science*, 82(2), 294–300. <https://doi.org/10.1016/j.ecss.2009.01.012>
- Friedman, A., Pizarro, O., Williams, S.B. & Johnson-Roberson, M. (2012). Multi-Scale Measures of Rugosity, Slope and Aspect from Benthic Stereo Image Reconstructions. *Plos ONE*, 7(12). <https://doi.org/10.1371/journal.pone.0050440>
- Genin, A. (2004). Bio-physical coupling in the formation of zooplankton and fish aggregations over abrupt topographies. *Journal of Marine Systems*, 50(1–2), 3–20. <https://doi.org/10.1016/j.jmarsys.2003.10.008>
- Gruss, A., Thorson, J.T., Babcock, E.A. & Tarnecki, J.H. (2018). Producing distribution maps for informing ecosystem-based fisheries management using a comprehensive survey database and spatio-temporal models. *ICES Journal of Marine Science*, 75(1), 158–177. <https://doi.org/10.1093/icesjms/fsx120>
- Gruss, A., Thorson, J.T., Sagarese, S.R., Babcock, E.A., Karnauskas, M., Walter, J.F. et al. (2017). Ontogenetic spatial distributions of red grouper (*Epinephelus mono*) and gag grouper (*Mycteroperca microlepis*) in the US Gulf of Mexico. *Fisheries Research*, 193, 129–142. <https://doi.org/10.1016/j.fishres.2017.04.006>
- Guisan, A. & Zimmermann, N.E. (2000). Predictive habitat distribution models in ecology. *Ecological Modelling*, 135(2–3), 147–186. [https://doi.org/10.1016/S0304-3800\(00\)00354-9](https://doi.org/10.1016/S0304-3800(00)00354-9)
- Hare, J.A. (2014). The future of fisheries oceanography lies in the pursuit of multiple hypotheses. *ICES Journal of Marine Science*, 71(8), 2343–2356. <https://doi.org/10.1093/icesjms/fsu018>
- Hobday, A.J., Hartog, J.R., Manderson, J.P., Mills, K.E., Oliver, M.J., Pershing, A.J. et al. (2019). Ethical considerations and unanticipated consequences associated with ecological forecasting for marine resources. *ICES Journal of Marine Science*, 76(5), 1244–1256. <https://doi.org/10.1093/icesjms/fsy210>
- Hobson, R.D. (1972). Surface roughness in topography: quantitative approach. In: R.J. Chorley (Ed.), *Spatial analysis in geomorphology*. New York, NY, USA: Harper and Row, pp. 221–245.
- Hutchinson, G.E. (1957). Concluding Remarks. *Cold Spring Harbour Symposium on Quantitative Biology*, 22, 415–427. <https://doi.org/10.1101/SQB.1957.022.01.039>
- Iverson, R.L. (1990). Control of Marine Fish Production. *Limnology and Oceanography*, 35(7), 1593–1604. <https://doi.org/10.4319/lo.1990.35.7.1593>
- Johnson, K.F., Thorson, J.T. & Punt, A.E. (2019). Investigating the value of including depth during spatiotemporal index standardization. *Fisheries Research*, 216, 126–137. <https://doi.org/10.1016/j.fishres.2019.04.004>
- Kane, J. (2007). Zooplankton abundance trends on Georges Bank, 1977–2004. *ICES Journal of Marine Science*, 64(5), 909–919. <https://doi.org/10.1093/icesjms/fsm066>
- Kearney, M. & Porter, W. (2009). Mechanistic niche modelling: Combining physiological and spatial data to predict species' ranges. *Ecology Letters*, 12(4), 334–350. <https://doi.org/10.1111/j.1461-0248.2008.01277.x>
- Kinlan, B.P. (2016). Modeling At-Sea Occurrence and Abundance of Marine Birds to Support Atlantic Marine Renewable Energy Planning: Phase I Report. U.S. Department of the Interior, Bureau of Ocean Energy Management, Office of Renewable Energy Programs, Sterling, VA. OCS Study BOEM 2016-039. xvii+113 pp.
- Kleisner, K.M., Fogarty, M.J., McGee, S., Barnette, A., Fratanoni, P.M., Greene, J. et al. (2016). The Effects of Sub-Regional Climate Velocity on the Distribution and Spatial Extent of Marine Species Assemblages. *PLoS ONE*, 11(2), e0149220. <https://doi.org/10.1371/journal.pone.0149220>
- Kleisner, K.M., Fogarty, M.J., McGee, S., Hare, J.A., Moret, S., Perretti, C.T. et al. (2017). Marine species distribution shifts on the US Northeast Continental Shelf under continued ocean warming. *Progress in Oceanography*, 153, 24–36. <https://doi.org/10.1016/j.pocean.2017.04.001>
- Labat, J.P., Gasparini, S., Mousseau, L., Prieur, L., Boutoute, M. & Mayzaud, P. (2009). Mesoscale distribution of zooplankton biomass in the northeast Atlantic Ocean determined with an Optical Plankton Counter: Relationships with environmental structures. *Deep-Sea Research Part I: Oceanographic Research Papers*, 56(10), 1742–1756. <https://doi.org/10.1016/j.dsr.2009.05.013>
- Laman, E.A., Rooper, C.N., Turner, K., Rooney, S., Cooper, D.W. & Zimmermann, M. (2018). Using species distribution models to describe essential fish habitat in Alaska. *Canadian Journal of Fisheries and Aquatic Sciences*, 75(8), 1230–1255. <https://doi.org/10.1139/cjfas-2017-0181>
- Leitao, P.J., Moreira, F. & Osborne, P.E. (2010). Breeding Habitat Selection by Steppe Birds in Castro Verde: A Remote Sensing and Advanced Statistics Approach. *Ardeola*, 57, 93–116.
- Loisel, A., Isla, A. & Daufresne, M. (2019). Variation of thermal plasticity in growth and reproduction patterns: Importance of ancestral and developmental temperatures. *Journal of Thermal Biology*, 84, 460–468. <https://doi.org/10.1016/j.jtherbio.2019.07.029>
- Lundblad, E.R., Wright, D.J., Miller, J., Larkin, E.M., Rinehart, R., Naar, D.F. et al. (2006). A Benthic Terrain Classification Scheme for American Samoa. *Marine Geodesy*, 29(2), 89–111. <https://doi.org/10.1080/01490410600738021>
- Manderson, J., Palamara, L., Kohut, J. & Oliver, M.J. (2011). Ocean observatory data are useful for regional habitat modeling of species with different vertical habitat preferences. *Marine Ecology Progress Series*, 438, 1–17. <https://doi.org/10.3354/meps09308>
- Manel, S., Williams, H.C. & Ormerod, S.J. (2001). Evaluating presence-absence models in ecology: The need to account for prevalence. *Journal of Applied Ecology*, 38(5), 921–931. <https://doi.org/10.1046/j.1365-2664.2001.00647.x>
- Maravelias, C.D. & Reid, D.G. (1997). Identifying the effects of oceanographic features and zooplankton on prespawning herring abundance using generalized additive models. *Marine Ecology Progress Series*, 147 (1–3), 1–9. <https://doi.org/10.3354/meps147001>
- Maritorena, S., d'Andon, O.H.F., Mangin, A. & Siegel, D.A. (2010). Merged satellite ocean color data products using a bio-optical model: Characteristics, benefits and issues. *Remote Sensing of Environment*, 114(8), 1791–1804. <https://doi.org/10.1016/j.rse.2010.04.002>
- Marshall, C.E., Glegg, G.A. & Howell, K.L. (2014). Species distribution modelling to support marine conservation planning: The next steps. *Marine Policy*, 45, 330–332. <https://doi.org/10.1016/j.marpol.2013.09.003>
- McGowan, D.W., Horne, J.K. & Rogers, L.A. (2019). Effects of temperature on the distribution and density of capelin in the Gulf of Alaska. *Marine Ecology Progress Series*, 620, 119–138. <https://doi.org/10.3354/meps12966>
- McHenry, J., Welch, H., Lester, S.E. & Saba, V. (2019). Projecting marine species range shifts from only temperature can mask climate vulnerability. *Global Change Biology*, 25(12), 4208–4221. <https://doi.org/10.1111/gcb.14828>
- McManus, M.C., Hare, J.A., Richardson, D.E. & Collier, J.S. (2018). Tracking shifts in Atlantic mackerel (*Scomber scombrus*) larval habitat suitability on the Northeast US Continental Shelf. *Fisheries. Oceanography*, 27(1), 49–62. <https://doi.org/10.1111/fog.12233>

- Miller, T.J., Das, C., Politis, P.J., Miller, A.S., Lucey, S.M., Legault, C.M. et al. (2010). Estimation of Albatross IV to Henry B. Bigelow calibration factors. *NEFSC Ref. Doc.*, 10–05.
- Murphy, M.A., Evans, J.S. & Storfer, A. (2010). Quantifying *Bufo boreas* connectivity in Yellowstone National Park with landscape genetics. *Ecology*, 91(1), 252–261. <https://doi.org/10.1890/08-0879.1>
- Norin, T., Canada, P., Bailey, J.A. & Gamperl, A.K. (2019). Thermal biology and swimming performance of Atlantic cod (*Gadus morhua*) and haddock (*Melanogrammus aeglefinus*). *PeerJ*, 7, e7784. <https://doi.org/10.7717/peerj.7784>
- Olli, K., Riser, C.W., Wassmann, P., Arashkevich, E. & Pasternak, A. (2002). Seasonal variation in vertical flux of biogenic matter in the marginal ice zone and the central Barents Sea. *Journal of Marine Systems*, 38(1–2), 189–204. [https://doi.org/10.1016/S0924-7963\(02\)00177-X](https://doi.org/10.1016/S0924-7963(02)00177-X)
- Pershing, A.J., Alexander, M.A., Hernadez, C.M., Kerr, L.A., Le Bris, A., Mills, K.E. et al. (2015). Slow adaptation in the face of rapid warming leads to collapse of the Gulf of Maine cod fishery. *Science*, 350(6262), 809–812. <https://doi.org/10.1126/science.aac9819>
- Pickens, B.A. & King, S.L. (2014). Linking multi-temporal satellite imagery to coastal wetland dynamics and bird distribution. *Ecological Modelling*, 285, 1–12. <https://doi.org/10.1016/j.ecolmodel.2014.04.013>
- Randin, C.F., Dirnbock, T., Dullinger, S., Zimmermann, N.E., Zappa, M. & Guisan, A. (2006). Are niche-based species distribution models transferable in space? *Journal of Biogeography*, 33(10), 1689–1703. <https://doi.org/10.1111/j.1365-2699.2006.01466.x>
- Rau, A., Lewin, W.C., Zettler, M.L., Gogina, M. & von Dorrien, C. (2019). Abiotic and biotic drivers of flatfish abundance within distinct demersal fish assemblages in a brackish ecosystem (western Baltic Sea). *Estuarine Coastal and Shelf Science*, 220, 38–47. <https://doi.org/10.1016/j.ecss.2019.02.035>
- Riley, S.J., DeGloria, S.D. & Elliot, R. (1999). A terrain ruggedness index that quantifies topographic heterogeneity. *Intermountain Journal of Sciences*, 5(1–4).
- Robinson, L.M., Elith, J., Hobday, A.J., Pearson, R.G., Kendall, B.E., Possingham, H.P. et al. (2011). Pushing the limits in marine species distribution modelling: Lessons from the land present challenges and opportunities. *Global Ecology and Biogeography*, 20(6), 789–802. <https://doi.org/10.1111/j.1466-8238.2010.00636.x>
- Rubec, P.J., Santi, C., Ghile, Y. & Chen, X.J. (2019). Modeling and Mapping to Assess Spatial Distributions and Population Numbers of Fish and Invertebrate Species in the Lower Peace River and Charlotte Harbor, Florida. *Marine and Coastal Fisheries*, 11(4), 328–350. <https://doi.org/10.1002/mcf2.10086>
- Saba, V.S., Griffies, S.M., Anderson, W.G., Winton, M., Alexander, M.A., Delworth, T.L. et al. (2016). Enhanced warming of the Northwest Atlantic Ocean under climate change. *Journal of Geophysical Research-Oceans*, 121(1), 118–132. <https://doi.org/10.1002/2015jc011346>
- Sandman, A.N., Wikstrom, S.A.A., Blomqvist, M., Kautsky, H. & Isaeus, M. (2013). Scale-dependent influence of environmental variables on species distribution: A case study on five coastal benthic species in the Baltic Sea. *Ecography*, 36(3), 354–363. <https://doi.org/10.1111/j.1600-0587.2012.07053.x>
- Sappington, J.M., Longshore, K.M. & Thompson, D.B. (2007). Quantifying landscape ruggedness for animal habitat analysis: A case study using bighorn sheep in the Mojave Desert. *Journal of Wildlife Management*, 71(5), 1419–1426. <https://doi.org/10.2193/2005-723>
- Sheppard, J.K., Lawler, I.R. & Marsh, H. (2007). Seagrass as pasture for seacows: Landscape-level dugong habitat evaluation. *Estuarine Coastal and Shelf Science*, 71(1–2), 117–132. <https://doi.org/10.1016/j.ecss.2006.07.006>
- Smoliński, S. (2018). Incorporation of optimal environmental signals in the prediction of fish recruitment using random forest algorithms. *Canadian Journal of Fisheries and Aquatic Sciences*, 76(1), 15–27. <https://doi.org/10.1139/cjfas-2017-0554>
- Smoliński, S. & Radtke, K. (2017). Spatial prediction of demersal fish diversity in the Baltic Sea: Comparison of machine learning and regression-based techniques. *ICES Journal of Marine Science*, 74(1), 102–111. <https://doi.org/10.1093/icesjms/fsw136>
- Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T. & Zeileis, A. (2008). Conditional variable importance for random forests. *BMC Bioinformatics*, 9(1), 307. <https://doi.org/10.1186/1471-2105-9-307>
- Strobl, C., Boulesteix, A.-L., Zeileis, A. & Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations. *Sources and a Solution. BMC Bioinformatics*, 8(1), 25. <https://doi.org/10.1186/1471-2105-8-25>
- Thorson, J.T., Shelton, A.O., Ward, E.J. & Skaug, H.J. (2015). Geostatistical delta-generalized linear mixed models improve precision for estimated abundance indices for West Coast groundfishes. *ICES Journal of Marine Science*, 72(5), 1297–1310. <https://doi.org/10.1093/icesjms/fsu243>
- Torres, L.G., Read, A.J. & Halpin, P. (2008). Fine-scale habitat modeling of a top marine predator: Do prey data improve predictive capacity? *Ecological Applications*, 18(7), 1702–1717. <https://doi.org/10.1890/07-1455.1>
- Veza, P., Munoz-Mas, R., Martinez-Capel, F. & Mouton, A. (2015). Random forests to evaluate biotic interactions in fish distribution models. *Environmental Modelling & Software*, 67, 173–183. <https://doi.org/10.1016/j.envsoft.2015.01.005>
- Ware, D.M. & Thomson, R.E. (2005). Bottom-up ecosystem trophic dynamics determine fish production in the northeast Pacific. *Science*, 308(5726), 1280–1284. <https://doi.org/10.1126/science.1109049>
- Wenger, S.J. & Olden, J.D. (2012). Assessing transferability of ecological models: An underappreciated aspect of statistical validation. *Methods in Ecology and Evolution*, 3(2), 260–267. <https://doi.org/10.1111/j.2041-210X.2011.00170.x>
- Winship, A.J. (2018). 'Modeling At-Sea Density of Marine Birds to Support Atlantic Marine Renewable Energy Planning: Final Report. U.S. Department of the Interior, Bureau of Ocean Energy Management, Office of Renewable Energy Programs, Sterling, VA. OCS Study BOEM 2018-010. x+67 pp.
- Xue, Y., Tanaka, K.S., Yu, H.M., Chen, Y., Guan, L.S., Li, Z.G. et al. (2018). Using a new framework of two-phase generalized additive models to incorporate prey abundance in spatial distribution models of juvenile slender lizardfish in Haizhou Bay. *China. Marine Biology Research*, 14(5), 508–523. <https://doi.org/10.1080/17451000.2018.1447673>
- Yates, K.L., Bouchet, P.J., Caley, M.J., Mengersen, K., Randin, C.F., Parnell, S. et al. (2018). Outstanding Challenges in the Transferability of Ecological Models. *Trends in Ecology & Evolution*, 33(10), 790–802. <https://doi.org/10.1016/j.tree.2018.08.001>

How to cite this article: Friedland KD, Bachman M, Davies A, et al. Machine learning highlights the importance of primary and secondary production in determining habitat for marine fish and macroinvertebrates. *Aquatic Conserv: Mar Freshw Ecosyst*. 2021;31:1482–1498. <https://doi.org/10.1002/aqc.3527>