NRHA Modeling Approach and Methods Overview

Updated Sept 2, 2022

Basic Rationale/Considerations

Our goal was to develop a comprehensive habitat modeling framework that could 1) be used to assess historical patterns of habitat use for marine species on the Northeast Shelf, and 2) be adapted to generate long-term projections of habitat use based on projected future climate scenarios.

While traditional species distribution models (SDMs) explain patterns of habitat use as a function of environmental predictors, other factors such as biotic interactions can give rise to correlations in the occurrence or abundance of species that are not explained by the environment. By modeling species environmental responses as well as their "residual" covariances in space and time, spatiotemporal joint SDMs (JSDMs) may offer benefits over traditional, single species models that do not attempt to control for these factors. These gains can include less biased estimates of species environmental responses and/or the uncertainty around them, the pooling of information across species to improve parameter estimation (i.e., borrowing strength) and the resolution of underlying or "latent" gradients, and the option to predict joint occurrences and/or condition upon the occurrence states or abundances of other species, producing more realistic predictions of species assemblages. Finally, the residual correlations estimated by JSDMs (and/or the partial correlations derived from them) may provide insights on potentially important ecological processes, such as biotic interactions or unmeasured "missing" predictors.

Because basic ecological requirements, species interactions, and habitat use patterns can vary over ontogeny, we chose to model adults and juveniles of each species as distinct groups, based on length at maturity (when data was available).

Due to the dynamic nature of the marine environment, whenever possible we used time-varying measurements for covariates. In addition to considering water depth as a covariate, we also sought to explore the role of other correlates of depth, such as gradients in hydrodynamic stress or in the quality of underwater light, which may be linked more mechanistically to organismal function.

Community-level basis function model (CBFM) – a novel approach to joint SDMs

Most existing implementations of joint SDMs represent adaptations of the latent variable model (LVM; Warton et al. 2015, Hui 2016, Ovaskainen et al. 2016) and employ a Bayesian framework that depends on computationally expensive Markhov chain Monte Carlo (MCMC) sampling for parameter estimation. This computational burden is compounded in a spatio-temporal context, where the estimation of spatially and/or temporally structured latent fields means that model complexity scales rapidly with the number of observational units. As such, fitting spatiotemporal JSDMs to large datasets (such as ours) can involve processing times that make this approach largely impractical.

We (with collaborator Dr. Francis Hui at Australian National University and others) developed a novel approach to fitting JSDMs, the community-level basis function model (CBFM). In lieu of spatially structured latent variables, CBFM employs a pre-specified set of fixed spatial (and/or temporal) basis functions that are common (i.e., shared) across all species. Species' covariances in space, time, and with each other are then modeled via their respective basis function coefficients (i.e., weights or loadings), which are treated as random slopes drawn from a common distribution. This approach can be seen as a parallel of LVM but offers several advantages, including better scaling for datasets with many observations, because the "randomness" is integrated at the species, instead of observation, level. Additionally, the basis function approach relaxes assumptions about stationarity, so that the strength of correlations between two points in spacetime is not just a function of their distance from one another, but also their specific locations. Moreover, because the basis function approach is closely related to generalized additive models (GAMs), CBFM can model species responses to covariates as smooth terms, whereas most existing JSDM frameworks are limited to linear or quadratic/polynomial terms. This was particularly convenient in the context of NRHA, where the flexibility and data-driven nature of GAMS had made them a method of choice for single-species models, and permitted straightforward comparisons of the two. *REFER TO CBFM MS for additional information……*

Biological response data

We examined abundance data from the NMFS Spring & Fall Bottom trawl surveys for the time period between 2000-2019. This limited timeframe was selected because: (1) recent patterns of habitat use are likely the most informative/relevant for identifying habitat suitability at present or in the future, and, (2) prior to this period, data for many satellite-derived covariates are not available. For assessing out-of-sample predictive performance, the dataset was split, with 15 years used for training (2000-2014) and 5 years held out for testing (2015-2019). The model(s) included response data for NRHA species of interest, as well as other dominant community constituents (i.e., those with high rates of occurrence) and taxa thought to constitute important prey for species of interest, based on the literature.

While calibration factors have been estimated (i.e., Miller et al. 2010) to account for the 2009 change in vessel and sampling gear from the RFV Albatross IV to the RFV Henry Bigelow, they do not exist for all species and are rarely stage specific. Moreover, in the case of presence-absence models, the calibrations can induce false “absences” in samples where species were in fact present. To circumvent these issues, we used raw (un-calibrated) count data, estimating a species-specific (and stage-specific where appropriate) “VESSEL” effect to attempt to control for the gear change.

Stage-specific counts were generated using length information collected during the surveys along with estimated length at maturity data (L50s) collected from the literature. For a given tow, each observed length-class bin was classified as either “adult” or “juvenile”, and then the counts for each applicable length-class bin were summed for each life stage. Finally, the proportion of measured individuals comprising each life stage was multiplied by the total abundance of each species to obtain estimates of stage-specific abundance.

When maturity information was not available for a given species, all individuals were treated as a single group. Likewise, if either life stage of a species had fewer than 500 nonzero counts, the two stages were pooled into a single group. Species with fewer than 500 nonzero counts altogether (i.e., across stages) were not considered in the model. Species were also assigned to functional groups based on water column use (i.e., demersal, pelagic, or benthic in the case of epifaunal or infaunal invertebrates).

Basic modeling framework

Because some species overlap in their use of benthic/demersal and pelagic habitats, members of all three water column use groups are being combined in a single model. However, discrete models for benthic/demersal and pelagic communities are also being developed, permitting covariate sets to be more custom-tailored for each group (as the number of covariates that can be considered simultaneously in the model is limited). At present, the model includes 97 different species-stages. Benthic invertebrates (primarily molluscs and crustaceans) were not included in the presented model fits, but will be included in model runs that are currently in preparation.

We also opted to model the Spring and Fall surveys together, pooling the tow-level observations into a single model, instead of fitting to the data from each season separately. This was based on the reasoning that species' niches (i.e., responses to covariates) would be more completely represented by observations spanning a broader range of environmental conditions. In the case of temperatures, this might be particularly important for long-term projections based on climate model outputs. Moreover, preliminary comparisons indicate that the combined-season model performs comparably to (i.e., typically as well as or better than) single-season models for out-of-sample prediction.

Models were fitted to both binary presence-absence and abundance (count) response data. The overdispersed nature of counts and high proportion of zeros for many species necessitated (as indicated by model residuals) a more flexible mean-variance relationship than that of the negative binomial error distribution. To accommodate this, we are presently comparing two different approaches: (1) a stepwise "hurdle" model, first fitting to presence-absence data with binomial error and then fitting to counts conditional on presence using a zero-truncated negative binomial error distribution; and (2) a covariate-dependent zero-inflation model, fitting to counts (including "true" zeros) with negative binomial error, while the probability of observing "false" zeros is likewise modeled as a function of the covariates. Accordingly, the former (hurdle) approach assumes that detectability/catchability is somewhat invariant (with the exception of estimated vessel effects), while the latter (zero-inflation) assumes that detection efficiency can vary with environmental factors. (At present, results are only available for the hurdle model).

Species responses to continuous predictor variables were modeled as smooth terms using thin-plate regression splines, while the effect of vessel was modeled as a parametric term. In the presented models, the vessel effect is estimated for each species independently as a fixed effect; however, we are exploring the potential for drawing vessel effects randomly from a shared (i.e., species-common) distribution.

Spatiotemporal covariance was modeled using a tensor product of 50 multi-resolution thin-plate spline (MRTS) spatial basis functions and 3 temporal (gaussian) basis functions associated with month of the year (i.e., from 1-12; akin to using a smooth on "month"). This structure allows the effect of space to evolve over the year, with observations falling closer together in space-time (within the course of a year) more closely correlated, while also helping to account for seasonal patterns that may not be explained by measured covariates. Meanwhile, longer-term (i.e., interannual) temporal variability was modeled via a gaussian process smooth for "year" (with exponential dependency), drawn from a species-common distribution. This approach was selected after exploring several other alternatives, including a random "year" effect, an additional set of species-common temporal basis functions, and a tensor product (interaction) of spatial and long-term temporal (i.e. year) basis functions. However, the random effect tended to estimate unrealistically large year-to-year variation for some species, while the latter (more complex) approaches tended to lead to overfitting and poorer out-of-sample prediction in both CBFM and single-species GAMs.

As a point of comparison, we also fitted "stacked" single-species GAMs using the same covariate set, as well as stacked GAMS with an analogous spatiotemporal structure (which included a tensor-product smooth for latitude, longitude, and month, with a separate gaussian process smooth for year). Again, more complex spatiotemporal structures did not improve out-of-sample prediction performance for single-species GAMs, and often hindered it.

Uncertainty in species responses to covariates is measured via 95% confidence intervals, with corresponding uncertainty in predictions quantified via 95% prediction intervals. If estimates of uncertainty are available for predictor variables, additional model runs that incorporate this uncertainty (by including the upper/lower bounds of predictors and iterating over different combinations thereof) can be conducted.

Covariates & Covariate Selection

Physicochemical covariates included the following:
Surface temperature (monthly mean)
Bottom temperature (monthly mean)
Surface salinity (monthly mean)
Bottom salinity (monthly mean)
Annual minimum surface temperature (i.e., for the 12 months prior to an observation)
Annual maximum surface temperature (i.e., for the 12 months prior to an observation)
Annual minimum bottom temperature (i.e., for the 12 months prior to an observation)
Annual maximum bottom temperature (i.e., for the 12 months prior to an observation)
Sea surface height anomaly (monthly mean)
Bottom stress (95th quantile, static)
PAR (at 0.5* depth - monthly mean, modeled as a tensor product with hue angle)

Hue angle (at 0.5* depth - monthly mean, modeled as a tensor product with PAR)

Bathymetric Position Index (or BPI; broad scale, static)

Topographic complexity (Standard deviation of fine-scale BPI)

Mean sediment grain size (in phi units)

Temperatures, salinities, and sea surface height were obtained from the GLORYS 12v1 reanalysis (Jean-michel et al. 2021), which provides spatially and temporally continuous data at a spatial resolution of 1/12 degree (~ 9km) daily and corresponds closely to measured observations on the NE shelf (Chen et al. 2021). In addition to monthly mean surface and bottom temperatures and salinities, we also included long-term (annual) temperature extremes (i.e., min and max), which can be important drivers for many taxa (Morley et al. 2018). Sea surface height variations related to circulatory features such as fronts and eddies are often associated with productivity, making them a valuable predictor for many species as well (McHenry et al 2019). If desired, data from other circulation models (i.e., ROMS/HYCOM) could be substituted here, and it may be worth performing some type of bias correction for these predictors, based on the instantaneous point measurements taken during NMFS surveys.

Sea Bottom Stress (95th quantile, annual) was sourced from the USGS Sea Floor Stress and Sediment Mobility database (Dalyander et al., 2012), with spatial resolutions ranging from 3.5 to 5 km. As time-varying data were not available, this was treated as a static variable. A measure of the strength of hydrodynamic forcing due to waves and currents at the seabed, bottom stress is a close correlate of water depth that has direct physical implications for locomotion, resource acquisition, and energetic costs of marine organisms, and may indirectly reflect other aspects of the benthic environment (e.g., epifaunal or infaunal community composition). Rather than a mean or median, we opted to use 95th quantile values to capture the magnitude of more extreme events, which are often more ecologically relevant determinants of habitat use (Denny et al. 2009).

Water column optical characteristics were estimated from remote sensing data with a horizontal resolution of 4km, following the methods of Lee et al. (2021) and Lee et al. (2005) for hue angle and photosynthetically-active radiation (PAR) at depth, respectively (in collaboration with Dr. ZhongPing Lee at UMass Boston). PAR measures the intensity of light (largely without regard for the spectral distribution), with high values indicating greater levels of illumination, such as those that would be experienced in clearer waters or shallower depths where attenuation by the water column is limited. Alternatively, hue angle quantifies the spectral distribution (i.e., the “color”) of light, ranging from roughly 40 deg in shallow, “red” estuarine waters with high levels of dissolved and suspended substances, to 120 deg in oceanic surface waters, and up to 240 deg at the lower end of the photic zone in clear, deep “blue” oceanic waters.

Together, these two variables describe the basic quality of underwater illumination, which is closely related to both water depth and distance to the coastline, but may be a more proximal driver of habitat selection given it can be directly observed or sensed by (most) organisms occupying the photic zone (i.e., most of the modeled species). Moreover, because vision is the primary sensory mechanism through which many of these organisms perceive and navigate their surroundings, identify resources, capture prey, and detect and avoid predators, the surrounding optical environment may have more direct ecological relevance than water depth; indeed, the ocular systems of fish and invertebrates exhibit physiological specializations adapted to the intensity and spectral composition of light in the habitats they occupy (de Busserolles et al. 2017, Cortesi et al 2020). To capture the suitability of this “optical habitat”, we modeled the interaction between PAR and hue angle via a tensor product smooth, producing a 2-D response surface.

Because most spectral attenuation (and consequently the change in light quality) occurs in the upper levels of the water column, near-surface and bottom values of hue angle, and to a lesser extent PAR, are correlated. To limit the number and collinearity of predictor variables, while also accommodating the fact that our model includes both demersal and pelagic functional groups, we estimated “generalized” optical parameters at the midpoint of the water column (i.e., 0.5 * depth). In the case of discrete demersal and pelagic models (where covariates could be more tailored to the functional group of interest), we estimated these parameters at some fixed near-surface depth (10m for pelagic species) or at the seabed (for demersals).

Hue angle is closely correlated with remotely sensed Chlorophyll A (CHLA) concentration (a covariate often taken to be representative of productivity), however our exploratory analyses indicated that the former was more informative, and thus CHLA was omitted from model fits that included optical parameters. However, for the sake of completeness and comparison, we are fitting alternative models that include water depth and mean monthly CHLA concentration in lieu of optical parameters and seabed stress (i.e., correlates of depth).

Benthic habitat characteristics related to topography (bathymetric position or complexity) and to substrates (e.g., sediment type or grain size) can be important environmental predictors for demersal species. Bathymetric data for the study area were derived in combination from the USGS Coastal Relief Model (CRM90), the Nature Conservancy's NAMERA dataset, and NOAA's BlueTopo source bathymetry. We attempted to capture spatial variation in the nature/character of seafloor habitats through three different variables. Bathymetric Position Index (BPI) reflects the elevation of a given locale relative to the surrounding seascape, with more positive values indicating locally raised areas, more negative values indicating local depressions, and values near 0 signifying regions of more uniform slope. To capture broad-scale topographic features (i.e., seamounts, ridges, etc.) we estimated BPI from 100 meter resolution bathymetric data using NOAA's Benthic Terrain Modeler in ARCGIS with an inner radius of 900 m and an outer radius of 9000 m (following previous work). Topographic complexity related to finer-scale seafloor features was quantified as the standard deviation of fine scale BPI (estimated with an inner radius of 300 m and outer radius of 1500 m), and then averaged within a 10 km radius of each 100 m grid cel (again following previous work). Finally, the character of benthic sediments was quantified through mean sediment grain size, extracted from the TNC's NAMERA soft-sediment map layer.

Covariates with heavily skewed distributions were log-transformed, and all were standardized (centered and scaled) for numerical stability. Covariate selection is carried out during the model-fitting process, wherein an additional penalty associated with each smooth term serves to shrink the effect of any non-informative covariates to zero, effectively removing them from the model.

To manage model complexity and reduce the potential for multicollinearity, we have limited the predictor variables considered to those that are broadly influential across the overall species pool. Likewise, we have (at present) limited explanatory variables to those that are measured (or estimated) at relatively fine spatio-temporal scales, and as such, we have not included region-wide indices or coarsely interpolated (i.e. over several years) spatial fields. However, we will continue to explore the utility of additional explanatory variables in the model.

*Due to time constraints, the current model runs did not include the additional shrinkage penalty (which extends model run times considerably), however prior experimentation shows this has little to no effect on prediction.

Model Assessment

Out-of-sample prediction performance was assessed by training on 15 years of data (2000-2014) and extrapolating to 5 years (2015-2019). For presence-absence models, classification and discrimination performance were quantified using AUC and Tjur R^2, while predictive deviance and RMSE were used to assess error/precision. For count models, we used pseudo R^2 (the spearman correlation between predicted and observed counts) and RMSE.

PRELIMINARY RESULTS:

Model Checking

Residual checks indicate that with the exception of a few extreme values (~4 SDs), overall distributional assumptions were met. There is evidence of a strong pattern for one species in the residual vs fitted values plot; we are presently diagnosing this issue. Refer to PLOTS/MODEL_CHECKING to view.

Predictive Performance

Considered across the species pool, the CBFM presence-absence fit had somewhat greater classification (AUC) and notably better discrimation (Tjur R^2) performance than single-species spatiotemporal GAMS (and much more so than GAMS that did not consider space and time), with comparable levels of error (RMSE). The median AUC was 0.93 (ranging from 0.78 - 0.99), the median Tjur R2 was 0.50 (0.1 - 0.75), and median RMSE was 0.28 (0.09 - 0.42). Refer to PLOTS/MODEL_PERFORMANCE to view.

COUNT MODEL PERFORMANCE to be included

For example species (Summer flounder and winter flounder):

Summer Flounder: AUC = 0.94 and 0.93, Tjur R2 = 0.62 and 0.30, RMSE = 0.34 and 0.22, for adults and juveniles, respectively.

Winter Flounder: AUC = 0.95 and 0.96, Tjur R2 = 0.66 and 0.65, RMSE = 0.30 and 0.26, for adults and juveniles, respectively.

COMMUNITY-LEVEL METRIC PERFORMANCE to be included

Predictor Significance

Across the two models (P/A and count), no covariate was significant for fewer than 31 spp. In both models, optical parameters were significant for the greatest number of spp (92 and 75 spp, for P/A and count, respectively), followed by bottom shear stress (75 and 61 spp, respectively). Every species had at least 2 significant predictors. Refer to PLOTS/ALL_SPECIES/PREDICTOR_SIGNIFICANCE to view

Species Response to Predictors

Vessel effects were significant for 47 spp in the P/A model and 54 spp in the count model, and were reasonably similar (i.e.,correlated) across the two count models (Spearman's R = 0.75). The work of Miller et al 2010 indicate that the Henry Bigelow (HB) is generally more efficient than the Albatross (AL). Consistent with this, the estimated effect of vessel (with AL being the baseline) was positive for the vast majority of species. There were, however, a few exceptions to this that require additional exploration. We are looking into the possibility of drawing vessel effects from a species-common, non-zero mean distribution which may enhance their estimation (they are presently estimated independently at the species level). To provide an additional point of comparison, we may also run equivalent fits omitting the vessel effect and using the pre-calibrated count data.

For smooth terms, the majority of estimated smooths appear reasonable and resemble the characteristic "niche" model (however there are some particularly "wiggly" smooths that deserve further exploration, and may require adjustments to wiggliness penalties, etc). Note that the tensor product for optical parameters (hue angle and PAR) is plotted as a 2-dimensional response surface, where PAR is on the Y axis and Hue angle is on the X axis, with the color gradient fill reflecting the overall effect magnitude. Higher values of PAR correspond with greater levels of illumination. Low hue angles typically indicate an optical environment on the "redder" (more estuarine or coastal) end of the spectrum, whie higher angles correspond to a "bluer" (more offshore) environment.

In future model outputs, covariate values/plot axes will be back-transformed to their original (un-standardized) units.

Refer to PLOTS/ALL_SPECIES/SPECIES_RESPONSE to view by predictor type (note the y-axis scale varies by sp. to exaggerate the shape of the response).

Refer to PLOTS/EXAMPLE_SPECIES/SPECIES_RESPONSE to view by species (note the y-axis scale is fixed so that effect magnitudes are comparable across predictors).

Variance Partitioning

Variance partitioning plots show the proportion of variance explained by each of the environmental predictor variables, by species.

Refer to PLOTS/ALL_SPECIES/VARIANCE_PARTITIONING to view plots for the entire community or PLOTS/EXAMPLE_SPECIES/VARIANCE_PARTITIONING to view subsets for summer and winter flounder.

Predictions

Predictions of presence/absence and of abundance were generally consistent with observations across the 20-year period, and appeared to resolve seasonal differences well. In the case of hurdle count models, some spuriously high predicted counts occur (for a few species) at the extremes of covariate space (most often for depth and/or optical parameters), usually near the shelf break where observed data were very sparse. These erroneous predictions are arising when the presence/absence component of the model estimates mean probabilities of occurrence to be very low (but not exactly zero), yet the count model estimates very high mean counts. We are in the process of addressing this issue.

Refer to PLOTS/EXAMPLE_SPECIES/PREDICTIONS to view

ADD PLOTS TO PLOTS/ALL_SPECIES/PREDICTIONS WHEN READY

Correlations

Residual correlations reflect the estimated residual covariance between species, or correlations in presence/absence or abundance that are not explained by species responses to the predictor variables. These may reflect the effects of missing predictor variables, dispersal processes, or biotic interactions. Partial correlations are obtained by inversion of the residual correlation matrix and control for indirect effects (e.g., if two species are positively correlated due to their shared negative correlations with another species) and therefore are considered to be a better indicator of “direct” biotic interactions. Still, correlations should be interpreted with caution/skepticism.

Across the community, the most noticeable general pattern is a tendency for strong positive correlations between adults and juveniles of the same species, which may reflect their common responses to unmeasured environmental variability but may also be indicative of dispersal processes/limitations.

Refer to PLOTS/ALL_SPECIES/CORRELATIONS to view the full matrices for the entire community

Refer to PLOTS/EXAMPLE_SPECIES/CORRELATIONS to view subsets for summer and winter flounder