Vol. 653: 105–119, 2020
https://doi.org/10.3354/meps13484

MARINE ECOLOGY PROGRESS SERIES
Mar Ecol Prog Ser

Published October 29[§]

Comparative species abundance modeling of Capitellidae (Annelida) in Tampa Bay, Florida, USA

Justin Hilliard[1,*], David Karlen[2], Thomas Dix[2], Sara Markham[2], Anja Schulze[1]

[1]Texas A&M University Galveston Campus, Galveston, Texas 77553, USA
[2]Environmental Protection Commission of Hillsborough County, Tampa, Florida 33619, USA

ABSTRACT: Capitellid polychaetes are ubiquitous throughout the world's oceans and are often encountered in high abundance. We used an extensive dataset of species abundance and distribution records of the *Capitella capitata* complex, *C. aciculata*, *C. jonesi*, *Heteromastus filiformis*, *Mediomastus ambiseta*, and *M. californiensis* from Tampa Bay, Florida, USA, as a model system of closely related species filling a similar ecological niche. We sought to (1) characterize the spatial distribution of each species, (2) determine if a single species abundance modeling strategy could be applied to them all, and (3) assess environmental drivers of species distribution and abundance. We found that all species had a zero-inflated abundance distribution and there was spatial autocorrelation by bay regions. Lorenz curves were an effective tool to assess spatial patterns of species abundance across large areas. Bay segment, depth, and dissolved oxygen were the most important environmental drivers. Modeling was accomplished by comparing 6 different approaches: 4 generalized additive models (GAMs: Poisson, negative binomial, Tweedie, and zero-inflated Poisson distributions), hurdle models, and boosted regression trees. There was no single model with top performance for every species. However, GAM-Tweedie and hurdle models performed well overall and may be useful for studies of other benthic marine invertebrates.

KEY WORDS: Polychaete · Ecology · Generalized additive model · Machine learning

1. INTRODUCTION

Species abundance modeling is frequently used in ecology for understanding biogeographic patterns and predicting impacts of climate change. These models are not to be confused with those commonly termed species distribution models and ecological niche models (Elith & Leathwick 2009, Sillero 2011, Peterson & Soberón 2012). While some of these methods use abundance data, they often use presence-only/-absence data, with the prediction of species distributions, and often maps, as an end goal. Research using species abundance models for marine organisms has been focused on theoretical ecology, conservation planning, and climate change. Taxonomic representation has been varied but biased toward vertebrates, with nearly 50% of studies focused on fish, birds, and mammals (Robinson et al. 2017).

Generalized linear models (GLMs) (Nelder & Wedderburn 1972, McCullagh & Nelder 1983) and generalized additive models (GAMs) (Hastie & Tibshirani 1986, 1990) are regression techniques frequently used for modeling species abundance data. GLMs and GAMs are similar in that they both allow for non-Gaussian response distributions and use a monotonic function, often logarithmic, to link the response and predictors. The difference is that GAMs utilize smoothing functions on the predictors to determine their individual relationships with the response (Guisan et al. 2002, Zuur et al. 2009). When there is overdispersion due to zero-inflation, GLMs can be ex-

*Corresponding author: jstnhllrd@gmail.com

[§]References to the supplementary material mounted on Zenodo (p. 109) were added after publication. This corrected version: November 4, 2020

tended to 2-part, or hurdle, models (Cragg 1971). This approach first fits the abundance data as presence/absence with a binary response and then truncates the data and fits non-zero abundance with a Poisson or negative binomial response (Zuur et al. 2009).

A newer method being applied to abundance modeling uses boosted regression trees (BRTs). This approach combines classification and regression trees (Breiman et al. 1984) with boosting algorithms (Freund & Schapire 1996). Some advantages of BRTs include their ability to accommodate nonparametric datasets and fit complex interactions (De'ath 2007, Elith et al. 2008). For marine species abundance modeling, GLMs and GAMs were used 18% of the time while BRTs were used in only 4.2% of papers reviewed by Robinson et al. (2017). Hegel et al. (2010) provided a general overview of various other modeling strategies. Several comparisons of modeling strategies have been made, and while some have targeted marine organisms (Connolly et al. 2009, Shelton et al. 2014), they are often focused on terrestrial organisms (e.g. Potts & Elith 2006, Baldridge et al. 2016) and vertebrates (Oppel et al. 2012). Such a comparative study of modeling strategies has not been completed for capitellid polychaetes.

Capitellids occur ubiquitously throughout the world's oceans. They have been reported from river mouths, estuaries, sea grass beds, deep sea sediments, and even wood and bones from whale falls in the deep sea (Judge & Barry 2016, Silva et al. 2016). This is especially the case for the best-known genus of the family, *Capitella*. Cryptic species of *C. capitata* (Fabricius 1780) were initially reported off the coast of Massachusetts (USA) primarily on the basis of life history characteristics and allozyme data (Grassle & Grassle 1976). Since then, 50+ putative species have been described worldwide on the basis of life history alone (Méndez et al. 2000). Recent efforts have aimed to understand this species complex using the mitochondrial cytochrome *c* oxidase subunit I (COI) gene from the coasts of Brazil, Japan, Korea, Italy, and the Gulf of Mexico (Hilliard et al. 2016, Tomioka et al. 2016, Livi et al. 2017, Man-Ki et al. 2017, Silva et al. 2017). Some DNA barcoding with COI has been done on other genera (Carr et al. 2011, Lobo et al. 2016), and a phylogeny of the family indicates monophyly only for *Capitella* and a need to revise other genera (Tomioka et al. 2018).

We recognize that our analyses may be confounded by presence of cryptic species. Little is known about how many species comprise the *C. capitata* complex in Tampa Bay (Florida, USA), but recent work indicates at least 3 distinct genetic lineages (J. Hilliard et al. unpubl.). Preliminary work on *Heteromastus filiformis* (Claparède 1864) in the Gulf of Mexico indicates the presence of distinct genetic lineages worldwide and likely the presence of another species complex in Capitellidae (J. Hilliard pers. obs.). While there has been no work on genetic lineages of *Mediomastus ambiseta* (Hartman 1947) and *M. californiensis* (Hartman 1944), it can be hypothesized that they are also species complexes due to their large geographic range (Blake 2008) and the emerging patterns in *Capitella*. Unfortunately, this is a factor that we cannot control or account for at this time.

These capitellids have shown utility as bioindicators (reviewed by Dean 2008), and understanding ecological drivers of their abundance is a necessary step to effectively use them as indicators. All of these species, as well as *C. aciculata* (Hartman 1959) and *C. jonesi* (Hartman 1959), occur throughout Tampa Bay. Tampa Bay lies on the west-central Florida coast (27° 27′–28° 3′ N; 82° 20′–82° 44′ W), in a biogeographic transition zone between the Northern Gulf of Mexico and Floridian ecoregions, creating a very diverse system (Spalding et al. 2007, Yates & Greening 2011 and references therein). Tampa Bay has an average depth of 4 m and a surface area of nearly 1036 km^2 (Morrison & Yates 2011). The shorelines are characterized by tidal flats and mangroves (Glick & Clough 2006).

The Environmental Protection Commission of Hillsborough County (EPCHC) has been continuously surveying the benthos of Tampa Bay since 1993. This dataset provides a unique opportunity for spatial modeling of benthic organisms in an estuarine system. We sought to conduct a meta-analysis using EPCHC data on the *C. capitata* complex, *C. aciculata*, *C. jonesi*, *H. filiformis*, *M. ambiseta*, and *M. californiensis* (Fig. S1 in the Supplement at www.int-res.com/articles/suppl/m653p105_supp.pdf). One goal of this study was to explore the data to understand spatial patterns inherent to each species. A second goal was to model environmental drivers of species abundance and ask whether there is one modeling strategy that works well for all species. This was accomplished by comparing 6 different approaches: 4 GAMs (Poisson, negative binomial, Tweedie, and zero-inflated Poisson distributions), hurdle models, and BRTs. The third goal was to use random forest models (Breiman 2001), another classification and regression method, to evaluate environmental drivers. These results can be used to inform future studies of benthic invertebrate spatial ecology in general as well as population genetics, speciation, phylogeography, and toxicology of capitellids in the Gulf of Mexico.

2. MATERIALS AND METHODS

2.1. Data collection

The EPCHC has been collecting benthic samples throughout the bay since 1993 following a program designed by the Tampa Bay Estuary Program to monitor large changes throughout the bay using robust randomized sampling (Squires et al. 1993). Tampa Bay is divided into 7 segments: Hillsborough Bay (HB), Boca Ciega Bay (BCB), Terra Ceia Bay (TCB), Manatee River (MR), Lower Tampa Bay (LTB), Middle Tampa Bay (MTB), and Old Tampa Bay (OTB) (Fig. 1). Hexagon grids are overlaid to further divide regions. Smaller hexagons are used in smaller regions (HB, BCB, TCB, and MR) to increase the number of samples. A number of hexagons are randomly selected for sampling each year (July–October) and a random point is generated within each hexagon. The same general strategy has been followed even though some aspects of the design (number of samples, reporting period, etc.) have changed over time.

Prior to 2007 there was a very large sampling effort with up to 134 samples collected in 1995 (Table S1). However, the effort was not consistent, with as few as 78 samples collected in 2006 (Table S1). Substantially fewer samples have been collected per year since 2007, but the sampling effort has become more consistent by bay segment and overall, with about 44 samples collected per year (Table S1). This is also evidenced by the sample decimal ratio, which increases and becomes more consistent from 2007 onward (Table S1).

From 2007 onward, MR+TCB and LTB+MTB were treated as single reporting units by EPCHC for the random sample selection (Table S1). We did not combine these bay regions for modeling and kept them as separate bay segment categories for consistency and to avoid added complexity. Despite this sampling change, the number of samples for each bay segment remained relatively constant (Table S1).

Infaunal samples were collected with a single benthic grab using a Young-modified Van Veen grab (0.04 m^2), sifted through a 500 micron mesh sieve, and bulk preserved. A 10% formalin solution was used for preservation prior to 2012 and NOTOXhisto™ (Scientific Device Laboratory) has been used since. After 72+ h, samples were washed and transferred to 70% isopropanol for storage and identification. Surface and bottom water quality (pH, temperature, dissolved oxygen, and salinity) and depth were recorded at the time of collection. A sample was also collected for calculation of the silt/clay fraction. More sampling design details, including a map of the hexagon grids, are available in the EPCHC Benthic Report (Karlen et al. 2015).

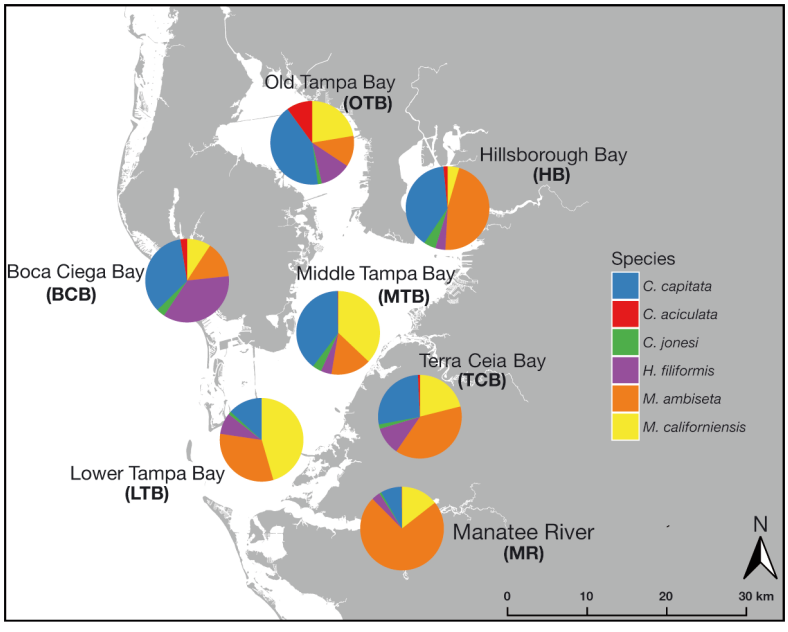

Fig. 1. Tampa Bay, Florida (USA). The 7 bay segments are indicated by the pie charts showing dominance of the 6 species (*Capitella*, *Heteromastus*, and *Mediomastus* spp.) over all 23 years. Note that the only instance of no species occurrence is *C. aciculata* in Lower Tampa Bay. Map created using QGIS Desktop and Inkscape. The Tampa Bay shapefile was sourced from the Florida Geographic Data Library

2.2. Database acquisition and filtering

The EPCHC maintains a Microsoft Access database on an FTP site (ftp://ftp.epchc.org/EPC_ERM_FTP/Benthic_Monitoring/). 'EPC DataSubmittals.zip' is the relevant file and contains all data from 1993 through to the present. Data used in this study span 1993 to 2015 and were downloaded in December 2017. Species identifications were performed by different agencies using the most current identification literature available at the time. The EPCHC is continually verifying identifications and posting data updates. Therefore, minor changes to our dataset have occurred since we completed the analyses and will continue to occur as specimens are re-examined.

The 2 relevant tables within the database are 'Biology,' with records of species abundance data, and 'DataSpreadsheet,' with all details related to field work and the measured environmental variables. These tables were joined and filtered for bottom measurements, year, hexagon ID, station number ID, latitude, longitude, and absolute abundance (number per 0.04 m^2) of the *Capitella capitata* complex (hereafter *C. capitata*), *C. aciculata*, *C. jonesi*, *Heteromastus filiformis*, *Mediomastus ambiseta*, *M. californiensis*, and all other Capitellidae entries. The final data frame for modeling consisted of 1788 observations of the abundance data for the 6 listed species, latitude, longitude, bay segment, year, temperature (°C), salinity (psu), pH, dissolved oxygen (mg l^{-1}), silt clay fraction, and depth (m).

2.3. Spatial statistics

Unless otherwise stated, all analyses were performed with R, version 3.4.4 (R Core Team 2019), and all graphics were generated with base-R graphics and the package 'ggplot2' (Wickham 2016). Inkscape was used to further modify figures for final presentation.

Spatial patterns were explored for data description and to inform statistical model construction. Species constancy (presence in samples) and dominance (proportion of total abundance) were calculated (Carmo et al. 2013). Constancy (C, %) is calculated as: $C = (p \times 100)/N$, where p is the number of samples in which the species was present, and N is the total number of samples. Categories include C ≥ 50%, constant; C = 25–50%, accessory; and C ≤ 25%, rare. Dominance (D, %) is calculated as: $D = (i/t) \times 100$, where i is the abundance of the species of interest, and t is total abundance. Categories include D ≥ 10%, eudominant; D = 5–10%, dominant; D = 2–5%, subdominant; D = 1–2%, recessive; and D ≤ 1%, rare. Pie charts of species dominance were plotted and overlaid onto a map of Tampa Bay to illustrate relative species dominance in the 7 regions (Fig. 1). To assess abundance as a function of space, Lorenz curves (Lorenz 1905, Burt et al. 2009), a graphical way to assess equality, were manually fit and overlaid with violin plots for each species as a function of bay segment. Abundance was averaged by bay segment for Lorenz curves to account for unequal sample sizes. A table was generated to assess variation in sampling spatially and temporally. Local indicator of spatial association (LISA), or the local Moran's I (Anselin 1995), was calculated and plotted using the software GeoDa (Anselin et al. 2006). The K-nearest neighbors method was used to define neighborhoods with 5 neighbors (6 total including the sample being considered). A random seed of 37 and 999 permutations was used to assess LISA significance, set at a pseudo p-value of 0.05. GeoDa was also used to generate bubble plots of species abundance. All of these data were exported as ESRI shapefiles and processed in QGIS Desktop for visualization.

Species abundance structures were assessed for overdispersion by comparing their mean and sample variances (Potts & Elith 2006). When variance equals mean, a Poisson model is appropriate. Overdispersion is present when the variance is larger than the mean and indicates that another distribution may be more appropriate. This was also assessed by a likelihood-ratio test between Poisson and negative binomial models using the r-package 'lmtest,' version 0.9-36 (Zeileis & Hothorn 2002). Zero-inflation of species abundance was assessed by visualization with violin plots using the r-package 'vioplot,' version 0.3.4 (Adler 2005).

2.4. Species abundance modeling

GAMs were fit for each species using all terms as covariates with the R package 'mgcv,' version 1.8-23 (Wood 2011, 2017, Wood et al. 2016). To not assume a linear relationship between each predictor and the response, a smoothing function was applied to all continuous covariates to generate a data-driven structure. Instead of using year as a categorical term, the covariate total samples yr^{-1} was calculated and used to determine if species abundance is a function of sampling effort. The models took the form of: Species ~ s(Temperature) + s(Salinity) + s(pH) + s(Dissolved oxygen) + s(Depth) + s(Silt clay fraction) + s(Total samples yr^{-1}) + Bay segment. All GAMs were fit using the 'REML' smoothing method and a logarithmic link function. The only exception is that the zero-inflated Poisson distribution models had to be fit with an identity link function.

The hurdle model was fit using the R package 'pscl,' version 1.5.2 (Zeileis et al. 2008, Jackman 2017), with all covariates in the binomial and negative binomial parts of the model. Link functions used were logarithmic for the negative binomial model and logit for the binomial model. BRTs were fit by first using the 'gbm.step' function (available in the supplementary materials of Elith et al. 2008, used for this study, and the R package 'dismo') to determine an optimal number of trees. A seed of 37 was set to

permit reproducibility, a Poisson distribution was used for the response, tree complexity (or interaction depth) was set to 1 to not allow any interactions, bag fraction was 0.5, and learning rate (or shrinkage) was started at 0.01 and adjusted until an optimal tree count of at least 1000 was reached (Elith et al. 2008). An exception was that *C. aciculata* was fit with a bag fraction of 1.0 due to algorithm convergence problems.

All models had static structure; all terms were used in every model with no term selection procedure or term interactions. This is because the goal was to assess model specification 'out-of-the-box' and not refine any particular model to optimize its fit. Predictor significance was assessed at alpha of 0.05 for all models except for BRTs, for which the relative influence of each term is estimated (Friedman 2001, Friedman & Meulman 2003). Significance of bay segment in hurdle models was assessed with a likelihood ratio test (R package 'lmtest'), resulting in a single significance value for the entire model.

Model evaluation and selection of a 'best' model was completed by testing internal predictive performance, comparing the model's predicted values against the observed values. Statistics used were the Pearson correlation coefficient, Spearman rank correlation, root mean square error (RMSE), average (mean absolute) error (AVE), slope, and intercept (Potts & Elith 2006). We also followed the methods of Potts & Elith (2006) to correct the calibration statistics by estimating bias using the 0.632+ bootstrap method (Efron & Tibshirani 1997, Steyerberg et al. 2001) with 200 iterations. R-code was sourced from the online supplementary material of Zuur et al. (2009).

A dataset with as many zeros as *C. aciculata* (16 records of presence, data not shown), especially an entire bay segment with all zeros (LTB), presented unique problems. Algorithm convergence failures and matrix singularities were routinely encountered and required more exploration of model parameterization to resolve this issue. A consequence is that we could not use the 0.632+ bootstrap corrections and had to remove bay segment. Instead of removing bay segment from the analyses, removing LTB samples would also have worked. We instead chose to keep LTB samples for their information in other cofactors. Another consequence is that stochasticity could not be included in the BRT, and bag fraction had to be fixed at 1.0. These problems are not unrelated, as bootstrapping for estimate correction and introducing stochasticity into the BRT both require subsampling the data frame. Limited presence records for *C. aciculata* results in a higher (1.05×10^{-7}) probability of a data frame with all zeros being built. In comparison, *C. capitata* has a 2.92×10^{-141} probability of this happening.

2.5. Environmental factors

Analyses for this section were performed with R version 3.6.0 (R Core Team 2019). Random forest models were built to assess environmental drivers of species abundance independent of the model specification comparisons. We allowed interactions in the random forest, as our goal was not to compare this to the other model specifications but to use it to understand environmental driver importance, and inclusion of interactions can aid this. As total samples yr^{-1} is not an environmental term, it was not included in this analysis. We specifically used conditional random forests because they reduce variable-selection bias due to differences in variable types and structures (Strobl et al. 2009). Models were fit with the R package 'party,' version 1.3-3 (Hothorn et al. 2006, Strobl et al. 2007, 2008). Hyperparameters were set at an mtry (number of input variables randomly sampled at each node) of 3 and the number of trees tuned until variable importance ranks stabilized when random seeds of 37 and 72 were compared (Strobl et al. 2008). Results are provided in a Zenodo repository (http://doi.org/10.5281/zenodo.4212321). A plot of variable importance was created.

In the interest of space, we chose to only assess the relationship between each species and its most important variable. Partial dependence plots (PDPs) (Friedman 2001) were generated using the R package 'pdp,' version 0.7.0 (Greenwell 2017). PDPs average the effects of a factor on model predictions while holding all other factors constant. See Molnar (2019) for further reading on PDPs.

PDPs were constructed on forests with hyperparameters tuned to an mtry of 3 and 3000 trees, the optimal tree number for some species, in the interest of reducing computation time. Using 3000 trees was enough to stabilize the rank of at least the top 2 or 3 predictors for all species. This was checked for *C. aciculata* with a PDP for a forest with 7000 trees (optimal setting) and there was no discernable difference. Results are provided in the Zenodo repository (see above). Additionally, since these data were collected for a large observational study and not to test any particular hypothesis, LOESS smoothers were plotted on continuous factor PDPs to emphasize the trends in relationships. An R-Script and the data frame are provided in the Zenodo repository (see above) for reproduction of all analyses and figures.

3. RESULTS

3.1. Spatial statistics

Constancy and dominance of the 6 species were variable throughout all bay segments. *Capitella capitata* was most dominant overall (D = 33.34%) (Table 1). *C. aciculata* was the least dominant overall at 2.81% (Table 1) and was not found at all in LTB. Its peak dominance was in OTB, where it comprised 10.02% of abundance (Fig. 1, Table 1). *C. jonesi* also occurred in a small portion of the samples (C = 5.31%) but was found throughout the bay with dominance ranging from 0.75–4.65% (Fig. 1, Table 1). *C. capitata* and *Mediomastus* spp. were the most constant throughout the bay and had dominance values that ranged from 4.54–73.15% of species abundance (Fig. 1, Table 1). *Heteromastus filiformis* had dominance throughout the bay at 3.89–11.93% but was most dominant in BCB (D = 38.20%) (Fig. 1, Table 1).

Species abundance distributions indicated that there is zero-inflation for all species (Fig. 2). There was also evidence of statistical overdispersion, indicating that distributions other than Poisson were appropriate (Table 2). Comparing GAM-Poisson and GAM-negative binomial models with a likelihood ratio test indicates that a negative binomial distribution described all species better (Table 3). These results led to use of a negative binomial distribution for the hurdle model.

The Lorenz curves illustrated zero inflation and spatial autocorrelation (Fig. 3). The violin plots for every species had a similar shape to those in Fig. 2, indicating zero inflation. Spatial autocorrelation was indicated by steep and changing slopes in the Lorenz curves. For example, the violin plot for *H. filiformis* (Fig. 3d) at BCB showed a very large abundance, and the steep slope between LTB and BCB indicated that BCB had a large portion, or unequal share, of all *H. filiformis* abundance in Tampa Bay. A contrasting example is *M. californiensis* (Fig. 3f), whose violin plots appeared more equal and Lorenz curve was closer to the line of equal distribution.

Table 1. Abundance (N), constancy (C, presence in samples, %), and dominance (D, proportion of total abundance, %) overall and by bay segment for *Capitella, Heteromastus,* and *Mediomastus* spp. See Fig. 1 for bay segment locations and abbreviations

	TOTAL			HB		
	N	C	D	N	C	D
C. capitata	2760	16.55	33.34	1082	20.54	38.99
C. aciculata	233	0.89	2.81	44	0.74	1.59
C. jonesi	250	5.31	3.02	129	6.44	4.65
H. filiformis	949	8.45	11.46	108	7.43	3.89
M. ambiseta	2706	13.20	32.69	1286	12.62	46.34
M. californiensis	1381	11.80	16.68	126	3.22	4.54
	OTB			MTB		
	N	C	D	N	C	D
C. capitata	618	18.01	42.13	392	9.50	39.60
C. aciculata	147	1.47	10.02	1	0.30	0.10
C. jonesi	23	5.88	1.57	35	5.34	3.54
H. filiformis	175	11.03	11.93	41	5.93	4.14
M. ambiseta	176	12.87	12.00	154	6.23	15.56
M. californiensis	328	9.56	22.36	367	13.35	37.07
	LTB			MR		
	N	C	D	N	C	D
C. capitata	66	10.05	13.23	65	9.47	8.16
C. aciculata	0	0.00	0.00	1	0.59	0.13
C. jonesi	6	2.28	1.20	6	1.18	0.75
H. filiformis	41	3.20	8.22	27	7.10	3.39
M. ambiseta	159	10.05	31.86	583	23.67	73.15
M. californiensis	227	23.74	45.49	115	11.24	14.43
	TCB			BCB		
	N	C	D	N	C	D
C. capitata	111	18.48	27.21	426	26.10	31.72
C. aciculata	3	1.09	0.74	37	2.03	2.76
C. jonesi	7	5.43	1.72	44	7.80	3.28
H. filiformis	44	9.78	10.78	513	14.58	38.20
M. ambiseta	157	20.65	38.48	191	16.27	14.22
M. californiensis	86	18.48	21.08	132	13.22	9.83

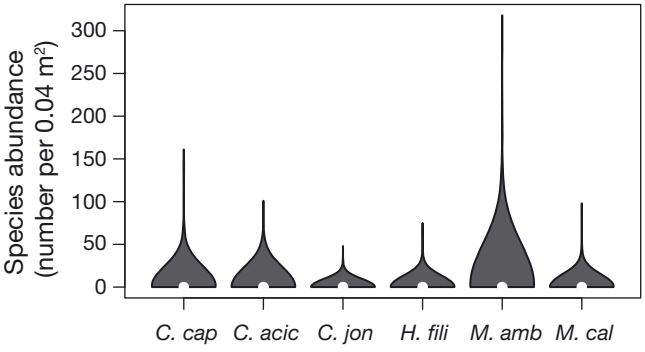

Fig. 2. Capitellid species abundance in Tampa Bay, Florida. The white points represent the median of each range, zero for all species. *C. cap* = *Capitella capitata*; *C. acic* = *C. aciculata*; *C. jon* = *C. jonesi*; *H. fili* = *Heteromastus filiformis*; *M. amb* = *Mediomastus ambiseta*; *M. cal* = *M. californiensis*

Table 2. Variance [(number per 0.04 m^2)^2] and means (number per 0.04 m^2) of species distributions for *Capitella*, *Heteromastus*, and *Mediomastus* spp.

	Variance	Mean
C. capitata	65.10	1.54
C. aciculata	7.37	0.13
C. jonesi	2.15	0.14
H. filiformis	13.00	0.53
M. ambiseta	126.38	1.51
M. californiensis	21.46	0.77

Table 3. Likelihood ratio test results from comparing the generalized additive model (GAM)-Poisson and GAM-negative binomial models for *Capitella*, *Heteromastus*, and *Mediomastus* spp. Significance at $p < 0.05$ indicates a better fit of the GAM-negative binomial model

	χ^2	df	p
C. capitata	5785.4	29.228	<0.0001
C. aciculata	172.84	27.359	<0.0001
C. jonesi	448.81	26.337	<0.0001
H. filiformis	1246.4	28.485	<0.0001
M. ambiseta	5390.5	19.511	<0.0001
M. californiensis	3414.9	31.856	<0.0001

However, all species showed some degree of spatial autocorrelation.

Looking at bubble plots of abundance and LISA plots (Fig. S2), there was evidence of spatial autocorrelation. For example, LISA plots showed that areas of species presence often resulted in significantly autocorrelated neighborhoods (Fig. S2). There was also a pattern of species presence near shore, with few occurrences in the open-water areas of the bay (Fig. S2).

3.2. Species abundance modeling

The 0.632+ bootstrap to correct model optimism was applied to every species except *C. aciculata*. A problem with matrix singularity was encountered during the 0.632+ bootstrap processes for *C. aciculata*, and apparent (non-adjusted) statistics of the model fits are presented for this species. Model calibration can be assessed with the slope and intercept (observed count ~ predicted count) (Fig. 4, Table S2) (Potts & Elith 2006). It is clear that there was a lot of variation within and between species. All models had a bias (intercept) within ±1.0 except BRT, *C. aciculata*; GAM-zero-inflated Poisson, *C. capitata*; and hurdle, *M. ambiseta*. The consistency/spread (slope) was more variable (Fig. 4, Table S2). The models considered best calibrated for each species were: hurdle, *C. capitata* (m = 0.95, b = 0.04); GAM-Poisson, *C. aciculata* (m = 1.01, b = 0.00); GAM-Tweedie, *C. jonesi* (m = 0.97, b = 0.01); hurdle, *Heteromastus filiformis* (m = 1.01, b = −0.01); GAM-negative binomial, *M. ambiseta* (m = 0.98, b = 0.26); and GAM-Tweedie, *M. californiensis* (m = 0.94, b = 0.07) (Fig. 4, Table S2).This calibration was reflected in both correlation values, with the best calibrated models generally having the highest, or near highest, values (Fig. S3, Table S2). It was also corroborated by the RMSE and AVE values, with selected models generally having the lowest or relatively low values (Fig. S3, Table S2).

3.3. Environmental factors

Significance was assessed at alpha = 0.05 for all models (Fig. 5) except for BRTs, for which the relative influence of each term was estimated (Friedman 2001, Friedman & Meulman 2003) (Fig. S4). GAM-Poisson and GAM-zero-inflated Poisson models found all terms significant for every species except for salinity in one *Heteromastus filiformis* model (Fig. 5). Bay segment was significant for every species/model combination. Depth was significant for most species/model combinations and was significant for every best-calibrated model. Total samples had the next overall significance with every species/best-calibrated model combination returning it as significant except for *C. jonesi* (Fig. 5). This was generally corroborated by the BRTs for which depth or bay segment were within the top 2 influential terms for most species. The exceptions were *C. aciculata*, which had pH as the only term that influenced abundance, and *M. californiensis*, which had depth third but still of strong influence (Fig. S4).

Random forest models indicated that bay segment and/or depth had high importance overall, evidenced by at least one of those terms ranking in the top 2 predictors for every species except *C. aciculata* (Fig. 6). Bay segment was most important for *C. jonesi*, while depth was most important for *C. capitata*, *H. filiformis*, and *M. ambiseta* (Fig. 6). Dissolved oxygen was most important for *M. californiensis* and pH for *C. aciculata* (Fig. 6). PDPs for *C. capitata*, *H. filiformis*, and *M. ambiseta* indicated a negative relationship between species abundance and depth; higher species abundances were found at shallower depths of 5 m or less (Fig. S5). The partial dependence of *C. jonesi* with bay segment (Fig. S5) reflected patterns of spatial autocorrelation observed in Lorenz curves (Fig. 3). For example, the

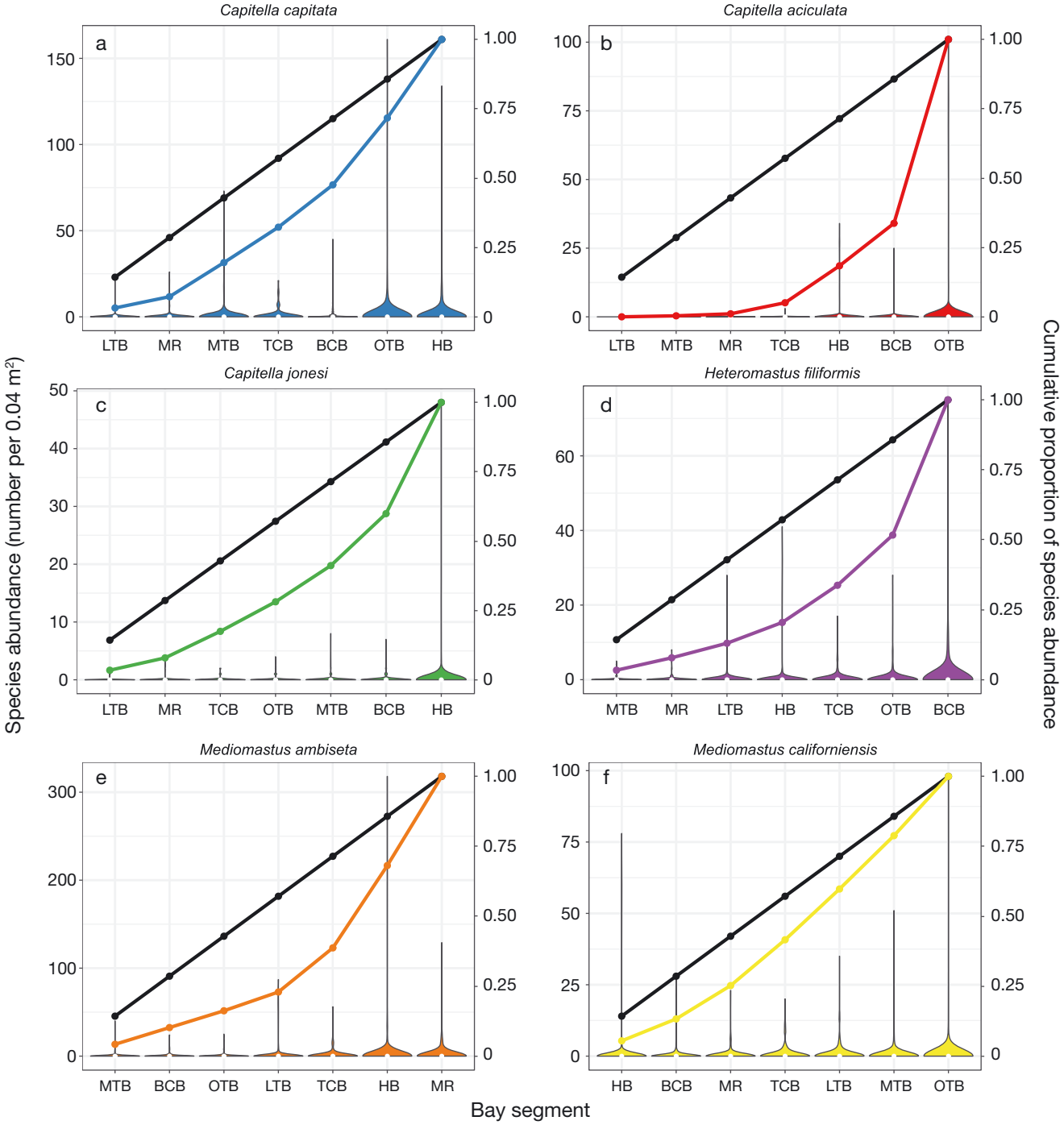

Fig. 3. Species abundance by bay segment (see Fig. 1 for locations and abbreviation), ordered by increasing average abundance. Lorenz curves of the cumulative proportion of averaged species abundance are overlaid. The different colored lines correspond to the unique distribution of each species, and the black lines are a representation of a species with abundance equally distributed among bay segments. Note that the *y*-axis scale varies among graphs

Lorenz curve for *C. jonesi* indicated a large portion of species abundance in HB, and the PDP showed a relatively large effect of HB. *M. californiensis* had a somewhat sigmoidal relationship with dissolved oxygen, with a large effect on abundance between ~4.4 and 7.5 mg l^{-1} (Fig. S5). *C. aciculata* also had a sigmoidal-like relationship with pH, with an increased effect on abundance between pH ~7.5 and 8.5 (Fig. S5).

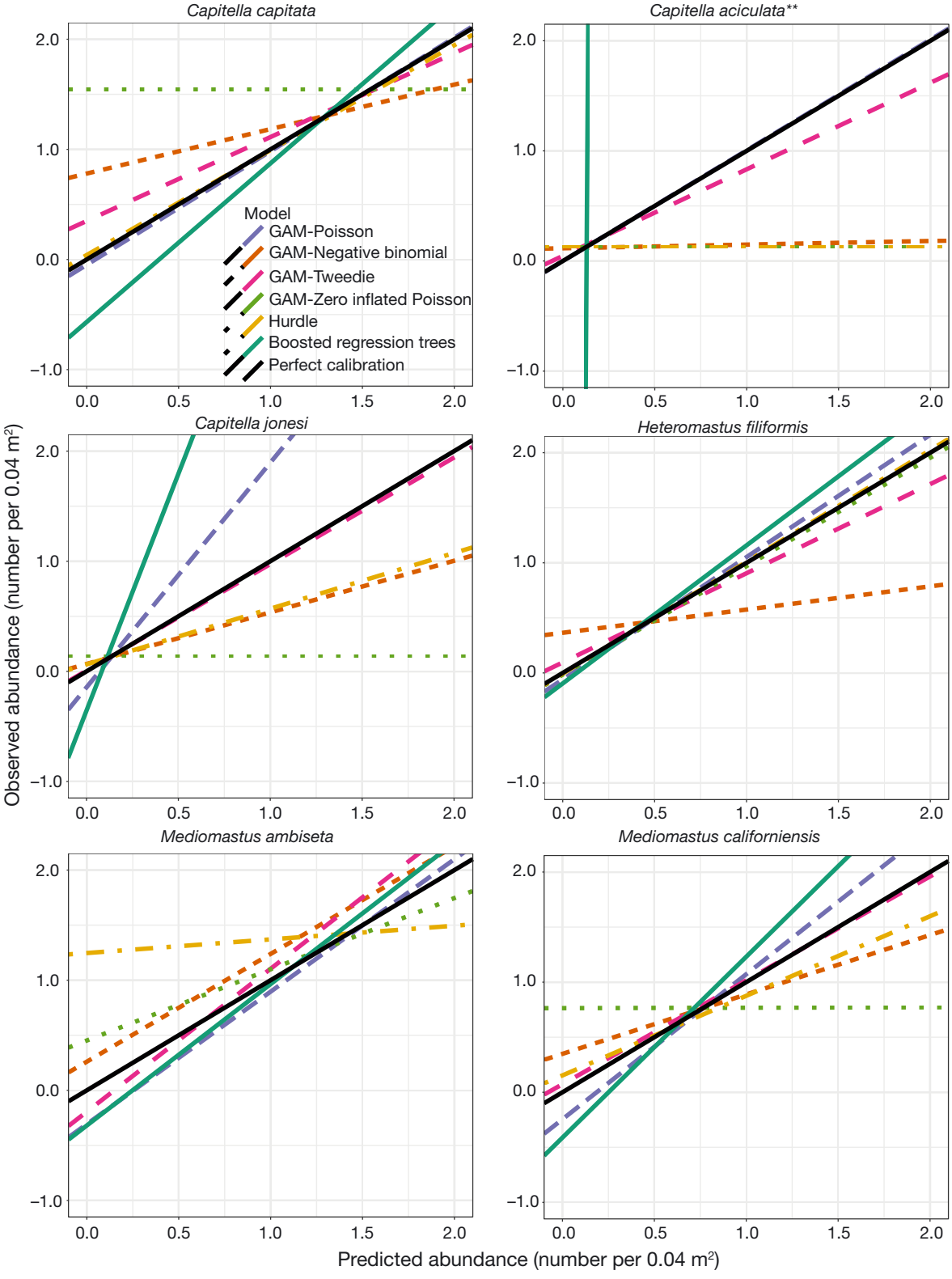

Fig. 4. Regression lines of observed vs. predicted capitellid abundance for all models. The values can be found in Table S2. Axis limits were restricted to a smaller range (2.0 maximum) for easier visualization of the intercepts. The *y*-axis was extended to −1.0 to accommodate negative intercepts. A regression line of perfect calibration was added for ease of interpretation. **: biased, apparent values are used

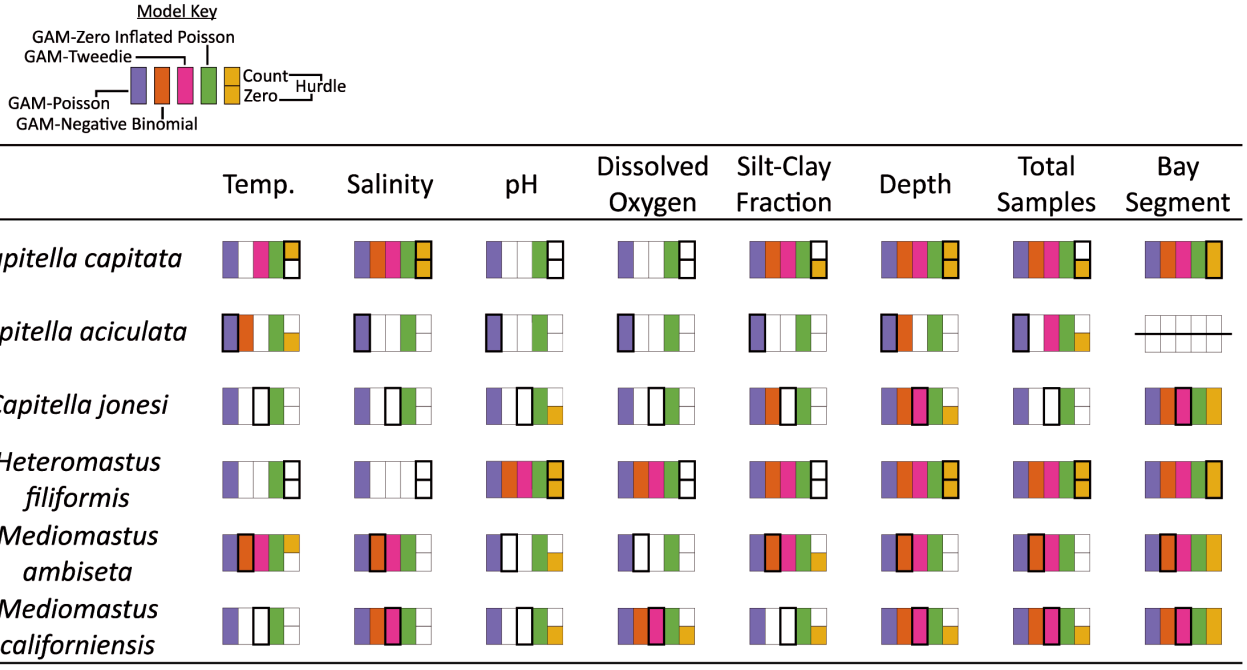

Fig. 5. Term significance for all models except boosted regression trees. If a term was significant (alpha = 0.05) for a given species/model combination, that block was colored. All colors correspond to those used for the models in Fig. 4. Note that the count and zero (presence/absence) parts of the hurdle model are separated for all terms except bay segment. Significance of bay segment was assessed for the hurdle model as a whole using a likelihood-ratio test. Bay segment was not investigated for *Capitella aciculata*. The model chosen as best calibrated for each species is highlighted with a **bold** block. GAM: generalized additive model

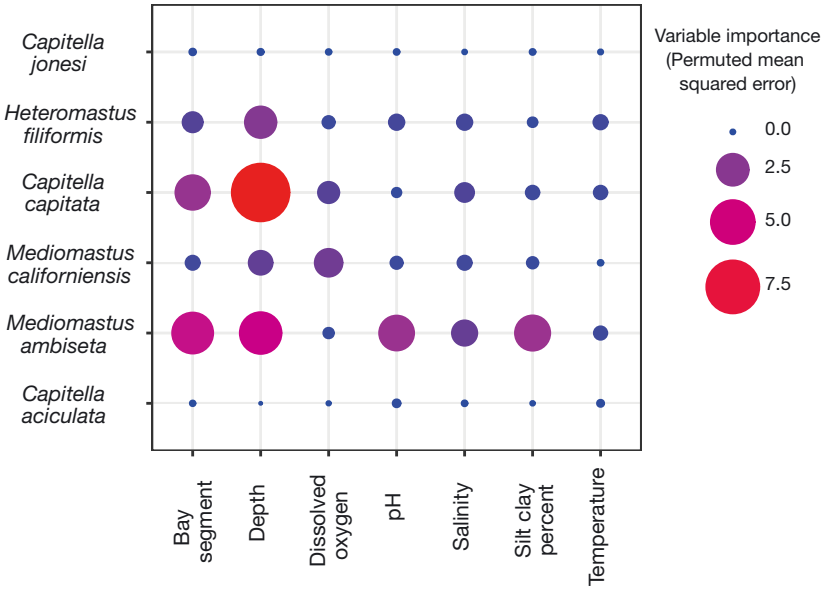

Fig. 6. Conditional random forest variable importance plots with species models organized in rows and environmental factors in columns. Variable importance represents the permutation importance of each factor, or the average effect on mean squared error when the factor is removed. Higher importance equates to higher error when that factor is removed from the model. Importance is scaled by size and color

4. DISCUSSION

Our results demonstrate the presence of spatial autocorrelation structured by bay segment for capitellids and variation in model specification across species. Using Lorenz curves in conjunction with violin plots demonstrated an effective way to assess this and overall species abundance structure across different regions of a large area. Interpretation of Lorenz curves is that the further the curve is from the line of perfect equality, the more unequally distributed the species is. One difference from standard Lorenz curves is that ours do not meet the line of equality at the origin. This is because we used an *x*-axis with categories that are all assumed to have some abundance in the case of perfect equality. The slope of the curve allows for quick assessment of autocorrelation; given a steep slope between 2 regions, it can be inferred that the region with

greater averaged abundance has proportionally more of the total species abundance in the entire bay. Some apparent examples are OTB for *Capitella aciculata*, BCB for *Heteromastus filiformis*, and TCB for *Mediomastus ambiseta* (Fig. 3). Additionally, for *C. jonesi*, the slope of the curve to each bay segment reflects the magnitude of effect that bay segment has in the PDP (results not shown).

Bubble plots of species abundance and LISA plots confirmed the autocorrelation indicated by Lorenz curves (Fig. S2). It is important to keep in mind that the small details do not matter much when interpreting the LISA plots, as this dataset was collected with large-scale patterns in mind. Therefore, what is important is whether or not bay segments appear significantly autocorrelated overall; the fact that one neighborhood is significant and a neighboring group is not has little interpretation because the sampling strategy is not appropriate for such comparisons. For example, *H. filiformis* abundance in BCB (Fig. S2) has an apparent clustering that is clear in the bubble plot. LISA plots show several neighborhoods with high–high (high values surrounded by other high values) and low–high (outliers of high value surrounded by low values) LISA scores (Fig. S2). This indicates several records of high abundance clustering together in the region.

The zero-inflatedness of every species is not surprising. Benthic infaunal invertebrates, especially estuarine polychaetes, are known for having patchy spatial distributions, often associated with grain size or organic matter/food (Warren 1977, James & Gibson 1980, Ansari et al. 1986, Kalejta & Hockey 1991, Widbom & Frithsen 1995, Sánchez-Moyano & García-Asencio 2009). Larval settlement patterns are sometimes attributed to the presence of conspecific adults (Osman & Whitlatch 1995, Snelgrove et al. 2001), chemical cues in sediment (Qian 1999), and/or subjection to near-bottom flow dynamics, with active selection during passive flow (Butman 1986, 1989, Butman & Grassle 1992, Snelgrove et al. 1993). The presence of some bacteria and their metabolites has even been attributed to inducing larval settlement (Harder et al. 2002, Lau et al. 2003, Chung et al. 2010). Some of these factors may have contributed to the low abundance of the capitellids in this study compared to some previous work in the Gulf of Mexico (Montagna et al. 2008, Palmer et al. 2011, Van Diggelen & Montagna 2016).

Capitella are among the better studied marine invertebrates (Grassle & Grassle 1976, Blake et al. 2009, Seaver 2016). One of the cryptic species discovered by Grassle & Grassle (1976) has since been formally described as *C. teleta* (Blake et al. 2009) and has been the subject of several studies on larval settlement (Dubilier 1988, Grassle et al. 1992, Hill & Nelson 1992, Thiyagarajan et al. 2005, Biggers et al. 2012, Burns et al. 2014). There are limited studies on the other genera considered (Hannan 1984, Snelgrove 1994).

Although not considered the best-calibrated model for every species, GAM-Tweedie and/or hurdle models have intercepts closest to, or near, the origin for all species. One of the two was considered best-calibrated for *C. capitata*, *C. jonesi*, *H. filiformis*, and *M. californiensis* (Table S2). *C. aciculata* and *M. ambiseta* have GAM-Tweedie slopes within ±0.3 (Table S2). This is an indication that although these 2 models may not be deemed the best fit in all cases, they are reasonably calibrated and could be considered a good starting point.

The overall high performance of GAM-Tweedie and hurdle models may be attributable to their ability to handle the excess zeros in unique ways. Hurdle models have been specifically used for rare species count data (Cunningham & Lindenmayer 2005). This approach splits the dataset into a binary version (presence/absence) and a zero-truncated abundance version, assuming that processes driving presence/absence are separate from those driving abundance (Cragg 1971, Zuur et al. 2009). All other methods keep the dataset whole and assume the processes driving zero-inflation are also driving abundance. A Tweedie distribution allows more flexibility for the shape of the species abundance distribution, as this is determined by a power term in the variance function (e.g. a power term of 1 is the Poisson distribution) (Jørgensen 1987). The R package 'mgcv' has the option to estimate the power term during model fitting, resulting in an automated distribution choice that may fit the data better than the standard Poisson or negative binomial distributions.

BRT models have shown equal (Martínez-Rincón et al. 2012) and better (Leathwick et al. 2006, França & Cabral 2015) performance compared to GAMs. Our results indicate better performance of GAMs. This is likely due to the 'stump model' restriction (not allowing any interactions), as a key benefit of a BRT is that it can handle very complex interactions that are not possible with the other methods (De'ath 2007, Elith et al. 2008). It is likely that a well-built BRT could perform just as well as, if not better than, the GAM-Tweedie and hurdle models. See Elith et al. (2008) for a guide to assembling BRT models and further references on the topic.

General location within a bay and sampling effort are expected to affect how many worms are col-

lected, so the overall significance of bay segment and total samples is not surprising. What was unexpected was the significance of depth, as Tampa Bay is only 4 m deep on average (Morrison & Yates 2011). This is not a result of collinearity, as standard measures (augmented pairs plots and variance inflation factor values) did not reveal any collinearity among environmental factors (results not shown). Relationships of species abundance and diversity with depth are complex (Houston & Haedrich 1984, Paterson & Lambshead 1995, Sibaja-Cordero et al. 2012). Studies focused on depth gradients in shallow estuarine systems would be useful to better understand this relationship.

We have shown that, despite filling a similar ecological niche (burrowing deposit feeders), there is not one model that is optimal for every species. Much consideration should be given to the biology of a species, especially the shape of its distribution in the area of interest, and the structure of the data frame (e.g. sampling design and scale). For example, *C. aciculata* was especially zero-inflated, which required more exploration of model parameterization. This highlights the complex biology of capitellids, as the extreme zero-inflation may be due to this species truly being rare. It may also be that *C. aciculata* is not a unique species (Hilliard et al. 2016), and the records should be combined with the *C. capitata* complex until there is further resolution of species boundaries. Consideration of the data structure and the sampling scale and design indicated that spatial autocorrelation needed to be accounted for on a bay-scale and comparisons at smaller scales were not appropriate. Taking an approach similar to the one presented here allows for systematic comparison of several modeling strategies at once. The model(s) considered best can then be refined. In the case of a benthic infaunal marine invertebrate with zero-inflated presence/absence records, hurdle and GAM-Tweedie models may be a good place to start if resources are limited.

Acknowledgements. Funding was provided to J.H. through a Texas A&M University Galveston Campus Boost Award. We thank the reviewers for their comments, which helped to greatly improve this manuscript. We thank everyone at the EPCHC for their sampling efforts. Finally, we thank the MEPS editorial staff for their diligent efforts producing this journal.

LITERATURE CITED

Adler D (2005) vioplot: violin plot. https://github.com/TomKellyGenetics/vioplot

Ansari ZA, Ingole BS, Parulekar AH (1986) Effect of high organic enrichment of benthic polychaete population in an estuary. Mar Pollut Bull 17:361–365

Anselin L (1995) Local Indicators of Spatial Association–LISA. Geogr Anal 27:93–115

Anselin L, Syabri I, Kho Y (2006) GeoDa: an introduction to spatial data analysis. Geogr Anal 38:5–22

Baldridge E, Harris DJ, Xiao X, White EP (2016) An extensive comparison of species-abundance distribution models. PeerJ 4:e2823

Biggers WJ, Pires A, Pechenik JA, Johns E, Patel P, Polson T, Polson J (2012) Inhibitors of nitric oxide synthase induce larval settlement and metamorphosis of the polychaete annelid *Capitella teleta*. Invertebr Reprod Dev 56:1–13

Blake JA (2008) Family Capitellidae Grube, 1862. In: Blake JA, Hilbig B, Scott PH (eds) Taxonomic atlas of the benthic fauna of the Santa Maria Basin and western Santa Barbara Channel, Vol 7. Santa Barbara Museum of Natural History, Santa Barbara, CA, p 47–53

Blake JA, Grassle JP, Eckelbarger KJ (2009) *Capitella teleta*, a new species designation for the opportunistic and experimental *Capitella* sp. I, with a review of the literature for confirmed records. Zoosymposia 2:25–53

Breiman L, Friedman JH, Olshen RA, Stone CJ (1984) Classification and regression trees. Chapman and Hall/CRC, Boca Raton, FL

Breiman L (2001) Random forests. Mach Learn 45:5–32

Burns RT, Pechenik JA, Biggers WJ, Scavo G, Lehman C (2014) The B vitamins nicotinamide (B3) and riboflavin (B2) stimulate metamorphosis in larvae of the deposit-feeding polychaete *Capitella teleta*: implications for a sensory ligand-gated ion channel. PLOS ONE 9: e109535

Burt JE, Barber GM, Rigby DL (2009) Elementary statistics for geographers. The Guilford Press, New York, NY

Butman CA (1986) Larval settlement of soft-sediment invertebrates: some predictions based on an analysis of near-bottom velocity profiles. Elsevier Oceanogr Ser 42: 487–513

Butman CA (1989) Sediment-trap experiments on the importance of hydrodynamical processes in distributing settling invertebrate larvae in near-bottom waters. J Exp Mar Biol Ecol 134:37–88

Butman CA, Grassle JP (1992) Active habitat selection by *Capitella* sp. I larvae. I. Two-choice experiments in still water and flume flows. J Mar Res 50:669–715

Carmo RFR, Amorim HP, Vasconcelos SD (2013) Scorpion diversity in two types of seasonally dry tropical forest in the semi-arid region of Northeastern Brazil. Biota Neotrop 13:340–344

Carr CM, Hardy SM, Brown TM, Macdonald TA, Hebert PDN (2011) A tri-oceanic perspective: DNA barcoding reveals geographic structure and cryptic diversity in Canadian polychaetes. PLOS ONE 6:e22232

Chung HC, Lee OO, Huang YL, Mok SY, Kolter R, Qian PY (2010) Bacterial community succession and chemical profiles of subtidal biofilms in relation to larval settlement of the polychaete *Hydroides elegans*. ISME J 4: 817–828

Claparède É (1864) Glanures zootomiques parmi les annélides de Port-Vendres (Pyrénées Orientales). Mém Soc Phys Hist Nat Genève 17:463–600 (https://doi.org/10.5962/bhl.title.1972)

Connolly SR, Dornelas M, Bellwood DR, Hughes TP (2009) Testing species abundance models: a new bootstrap approach applied to Indo-Pacific coral reefs. Ecology 90: 3138–3149

Cragg JG (1971) Some statistical models for limited dependent variables with application to the demand for durable goods. Econometrica 39:829–844

Cunningham RB, Lindenmayer DB (2005) Modeling count data of rare species: some statistical issues. Ecology 86: 1135–1142

De'ath G (2007) Boosted trees for ecological modeling and prediction. Ecology 88:243–251

Dean HK (2008) The use of polychaetes (Annelida) as indicator species of marine pollution: a review. Rev Biol Trop 56:11–38

Dubilier N (1988) H_2S — a settlement cue or a toxic substance for *Capitella* sp. I larvae? Biol Bull 174:30–38

Efron B, Tibshirani R (1997) Improvements on cross-validation: the 632+ bootstrap method. J Am Stat Assoc 92:548–560

Elith J, Leathwick JR (2009) Species distribution models: ecological explanation and prediction across space and time. Annu Rev Ecol Evol Syst 40:677–697

Elith J, Leathwick JR, Hastie T (2008) A working guide to boosted regression trees. J Anim Ecol 77:802–813

Fabricius O (1780) Fauna Groenlandica, systematice sistens animalia Groenlandiae Occidentalis hactenus indagata, quoad nomen specificum, triviale, vernaculumque; synonyma auctorum plurium, descriptionem, locum, victum, generationem, mores, usum, capturamque singuli, prout detegendi occasio fuit, maximaque parte secundum proprias observationes. Hafniae [= Copenhagen] & Lipsiae [= Leipzig], Ioannis Gottlob Rothe. xvi + 452 pp., 1 plate with 12 figures. https://www.biodiversitylibrary.org/page/13442285

França S, Cabral HN (2015) Predicting fish species richness in estuaries: which modelling technique to use? Environ Model Softw 66:17–26

Freund Y, Schapire RE (1996) Experiments with a new boosting algorithm. In: Saitta L (ed) Machine learning: Proceedings of the Thirteenth International Conference. Morgan Kaufmann, Bari, p 148–156

Friedman JH (2001) Greedy function approximation: a gradient boosting machine. Ann Stat 29:1189–1232

Friedman JH, Meulman JJ (2003) Multiple additive regression trees with application in epidemiology. Statist Med 22:1365–1381

Glick P, Clough J (2006) An unfavorable tide — global warming, coastal habitats and sport fishing in Florida. National Wildlife Federation. https://www.nwf.org/Educational-Resources/Reports/2006/06-01-2006-Unfavorable-Tide

Grassle JP, Grassle JF (1976) Sibling species in the marine pollution indicator *Capitella* (Polychaeta). Science 192: 567–569

Grassle JP, Butman CA, Mills SW (1992) Active habitat selection by *Capitella* sp. I larvae. II. Multiple-choice experiments in still water and flume flows. J Mar Res 50: 717–743

Greenwell BM (2017) pdp: an R package for constructing partial dependence plots. R J 9:421–436

Guisan A, Edwards TC, Hastie T (2002) Generalized linear and generalized additive models in studies of species distributions: setting the scene. Ecol Model 157: 89–100

Hannan CA (1984) Planktonic larvae may act like passive particles in turbulent near-bottom flows. Limnol Oceanogr 29:1108–1116

Harder T, Lau SCK, Dahms HU, Qian PY (2002) Isolation of bacterial metabolites as natural inducers for larval settlement in the marine polychaete *Hydroides elegans* (Haswell). J Chem Ecol 28:2029–2043

Hartman O (1944) Polychaetous annelids from California, including the descriptions of two new genera and nine new species. Allan Hancock Pac Exped 10:239–307

Hartman O (1947) Polychaetous annelids. Part VII. Capitellidae. Allan Hancock Pac Exped 10:391–481

Hartman O (1959) Capitellidae and Nereidae (marine annelids) from the Gulf side of Florida, with a review of freshwater Nereidae. Bull Mar Sci Gulf Caribb 9: 153–168

Hastie T, Tibshirani R (1986) Generalized additive models. Stat Sci 1:297–310

Hastie TJ, Tibshirani RJ (1990) Generalized additive models. Chapman and Hall/CRC, Boca Raton, FL

Hegel TM, Cushman SA, Evans J, Huettmann F (2010) Current state of the art for statistical modelling of species distributions. In: Cushman SA, Huettmann F (eds) Spatial complexity, informatics, and wildlife conservation. Springer, Tokyo, p 273–311

Hill SD, Nelson L (1992) Lindane (1, 2, 3, 4, 5, 6-hexachlorocyclohexane) affects metamorphosis and settlement of larvae of *Capitella* species I (Annelida, Polychaeta). Biol Bull 183:376–377

Hilliard J, Hajduk M, Schulze A (2016) Species delineation in the *Capitella* species complex (Annelida: Capitellidae): geographic and genetic variation in the northern Gulf of Mexico. Invertebr Biol 135:415–422

Hothorn T, Bühlmann P, Dudoit S, Molinaro A, Van Der Laan M (2006) Survival ensembles. Biostatistics 7:355–373

Houston KA, Haedrich RL (1984) Abundance and biomass of macrobenthos in the vicinity of Carson Submarine Canyon, northwest Atlantic Ocean. Mar Biol 82:301–305

Jackman S (2017) pscl: classes and methods for R developed in the Political Science Computational Laboratory. United States Studies Centre, University of Sydney, Sydney. R package version 1.5.2. https://github.com/atahk/pscl/

James CJ, Gibson R (1980) The distribution of the polychaete *Capitella capitata* (Fabricius) in dock sediments. Estuar Coast Mar Sci 10:671–683

Jørgensen B (1987) Exponential dispersion models. J R Stat Soc B 49:127–162

Judge J, Barry JP (2016) Macroinvertebrate community assembly on deep-sea wood falls in Monterey Bay is strongly influenced by wood type. Ecology 97:3031–3043

Kalejta B, Hockey PAR (1991) Distribution, abundance, and productivity of benthic invertebrates at the Berg River Estuary, South Africa. Estuar Coast Shelf Sci 33:175–191

Karlen DJ, Dix TL, Goetting BK, Markham SE, Campbell KW, Jernigan JM (2015) Twenty year trends in the benthic community and sediment quality of Tampa Bay: 1993–2012. Tampa Bay Benthic Monitoring Program Interpretive Report. Prepared for: Tampa Bay Estuary Program. https://www.epchc.org/home/showdocument?id=268

Lau SC, Harder T, Qian PY (2003) Induction of larval settlement in the serpulid polychaete *Hydroides elegans* (Haswell): role of bacterial extracellular polymers. Biofouling 19:197–204

Leathwick JR, Elith J, Francis MP, Hastie T, Taylor P (2006) Variation in demersal fish species richness in the oceans surrounding New Zealand: an analysis using boosted regression trees. Mar Ecol Prog Ser 321:267-281

Livi S, Tomassetti P, Vani D, Marino G (2017) Genetic evidences of multiple phyletic lineages of *Capitella capitata*

(Fabricius 1780) complex in the Mediterranean Region. J Mediterr Ecol 15:5–11

Lobo J, Teixeira MAL, Borges LMS, Ferreira MSG and others (2016) Starting a DNA barcode reference library for shallow water polychaetes from the southern European Atlantic coast. Mol Ecol Resour 16:298–313

Lorenz MO (1905) Methods of measuring the concentration of wealth. Publ Am Stat Assoc 9:209–219

Man-Ki J, Wi JH, Suh HL (2017) A reassessment of *Capitella* species (Polychaeta: Capitellidae) from Korean coastal waters, with morphological and molecular evidence. Mar Biodivers 48:1969–1978

Martínez-Rincón R, Ortega-García S, Vaca-Rodríguez JG (2012) Comparative performance of generalized additive models and boosted regression trees for statistical modeling of incidental catch of wahoo (*Acanthocybium solandri*) in the Mexican tuna purse-seine fishery. Ecol Model 233:20–25

McCullagh P, Nelder JA (1983) Generalized linear models, 2nd edn. Monographs on statistics and applied probability, Vol 37. Chapman and Hall/CRC, Boca Raton, FL

Méndez N, Linke-Gamenick I, Forbes VE (2000) Variability in reproductive mode and larval development within the *Capitella capitata* species complex. Invertebr Reprod Dev 38:131–142

Molnar C (2019) Interpretable machine learning. A guide for making black box models explainable. https://christophm.github.io/interpretable-ml-book/

Montagna PA, Palmer TA, Kalke RD, Gossmann A (2008) Suitability of using a limited number of sampling stations to represent benthic habitats in Lavaca-Colorado Estuary, Texas. Environ Bioind 3:156–171

Morrison G, Yates KK (2011) Environmental setting. In: Yates KK, Greening H, Morrison G (eds) Integrating science and resource management in Tampa Bay, Florida. U.S. Geological Survey Circular 1348. https://doi.org/10.3133/cir1348

Nelder JA, Wedderburn RWM (1972) Generalized linear models. J R Stat Soc Ser A 135:370–384

Oppel S, Meirinho A, Ramírez I, Gardner B, O'Connell AF, Miller PI, Louzao M (2012) Comparison of five modelling techniques to predict the spatial distribution and abundance of seabirds. Biol Conserv 156:94–104

Osman RW, Whitlatch RB (1995) The influence of resident adults on larval settlement: experiments with four species of ascidians. J Exp Mar Biol Ecol 190:199–220

Palmer TA, Montagna PA, Pollack JB, Kalke RD, DeYoe HR (2011) The role of freshwater inflow in lagoons, rivers, and bays. Hydrobiologia 667:49–67

Paterson GLJ, Lambshead PJD (1995) Bathymetric patterns of polychaete diversity in the Rockall Trough, northeast Atlantic. Deep Sea Res I 42:1199–1214

Peterson AT, Soberón J (2012) Species distribution modeling and ecological niche modeling: getting the concepts right. Nat Conserv 10:102–107

Potts JM, Elith J (2006) Comparing species abundance models. Ecol Model 199:153–163

Qian PY (1999) Larval settlement of polychaetes. Hydrobiologia 402:239–253

R Core Team (2019) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna

Robinson NM, Nelson WA, Costello MJ, Sutherland JE, Lundquist CJ (2017) A systematic review of marine-based species distribution models (SDMs) with recommendations for best practice. Front Mar Sci 4:421

Sánchez-Moyano JE, García-Asencio I (2009) Distribution and trophic structure of annelid assemblages in a *Caulerpa prolifera* bed from southern Spain. Mar Biol Res 5:122–132

Seaver EC (2016) Annelid models I: *Capitella teleta*. Curr Opin Genet Dev 39:35–41

Shelton AO, Thorson JT, Ward EJ, Feist BE (2014) Spatial semiparametric models improve estimates of species abundance and distribution. Can J Fish Aquat Sci 71: 1655–1666

Sibaja-Cordero JA, Cortés J, Dean HK (2012) Depth diversity profile of polychaete worms in Bahía Chatham, Isla del Coco National Park, Costa Rican Peninsula. Rev Biol Trop 60:293–301

Sillero N (2011) What does ecological modelling model? A proposed classification of ecological niche models based on their underlying methods. Ecol Model 222:1343–1346

Silva CF, Shimabukuro M, Alfaro-Lucas JM, Fujiwara Y, Sumida PYG, Amaral ACZ (2016) A new *Capitella* polychaete worm (Annelida: Capitellidae) living inside whale bones in the abyssal South Atlantic. Deep Sea Res I 108:23–31

Silva CF, Seixas VC, Barroso R, Di Domenico M, Amaral ACZ, Paiva PC (2017) Demystifying the *Capitella capitata* complex (Annelida, Capitellidae) diversity by morphological and molecular data along the Brazilian coast. PLOS ONE 12:e0177760

Snelgrove PVR (1994) Hydrodynamic enhancement of invertebrate larval settlement in microdepositional environments: colonization tray experiments in a muddy habitat. J Exp Mar Biol Ecol 176:149–166

Snelgrove PVR, Butman CA, Grassle JP (1993) Hydrodynamic enhancement of larval settlement in the bivalve *Mulinia lateralis* (Say) and the polychaete *Capitella* sp. I in microdepositional environments. J Exp Mar Biol Ecol 168:71–109

Snelgrove PVR, Grassle JP, Zimmer CA (2001) Adult macrofauna effects on *Capitella* sp. I larval settlement: a laboratory flume study. J Mar Res 59:657–674

Spalding MD, Fox HE, Allen GR, Davidson N and others (2007) Marine ecoregions of the world: a bioregionalization of coastal and shelf areas. Bioscience 57:573–583

Squires A, Janicki A, Heimbuch D, Wade D, Wilson H, Robison D (1993) A monitoring program to assess environmental changes in Tampa Bay, Florida. https://tbeptech.org/TBEP_TECH_PUBS/1993/TBEP_02_93_MonitorProgr-ForEnvironChange.pdf

Steyerberg EW, Harrell FE, Borsboom GJJM, Eijkemans MJC, Vergouwe Y, Habbema JDF (2001) Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. J Clin Epidemiol 54:774–781

Strobl C, Boulesteix A, Zeileis A, Hothorn T (2007) Bias in random forest variable importance measures: illustrations, sources and a solution. BMC Bioinformatics 8:25

Strobl C, Boulesteix A, Kneib T, Augustin T, Zeileis A (2008) Conditional variable importance for random forests. BMC Bioinformatics 9:307

Strobl C, Hothorn T, Zeileis A (2009) Party on! A new, conditional variable-importance measure for random forests available in the party package. R J 1:14–17

Thiyagarajan V, Soo L, Qian PY (2005) The role of sediment organic matter composition in larval habitat selection by the polychaete *Capitella* sp. I. J Exp Mar Biol Ecol 323: 70–83

Tomioka S, Kondoh T, Sato-Okoshi W, Ito K, Kakui K, Kajihara H (2016) Cosmopolitan or cryptic species? A case study of *Capitella teleta* (Annelida: Capitellidae). Zool Sci 33:545–554

Tomioka S, Kakui K, Kajihara H (2018) Molecular phylogeny of the family Capitellidae (Annelida). Zool Sci 35: 436–445

Van Diggelen AD, Montagna PA (2016) Is salinity variability a benthic disturbance in estuaries? Estuaries Coasts 39: 967–980

Warren LM (1977) The ecology of *Capitella capitata* in British waters. J Mar Biol Assoc UK 57:151–159

Wickham H (2016) ggplot2: elegant graphics for data analysis. Springer-Verlag, New York, NY. https://ggplot2.tidyverse.org

Widbom B, Frithsen JB (1995) Structuring factors in a marine soft bottom community during eutrophication — an experiment with radio-labelled phytodetritus. Oecologia 101:156–168

Wood SN (2011) Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. J R Stat Soc Ser B Stat Methodol 73:3–36

Wood SN (2017) Generalized additive models: an introduction with R, 2nd edn. Chapman and Hall/CRC, Boca Raton, FL

Wood SN, Pya N, Säfken B (2016) Smoothing parameter and model selection for general smooth models. J Am Stat Assoc 111:1548–1563

Yates KK, Greening H (2011) An introduction to Tampa Bay. In: Yates KK, Greening H, Morrison G (eds) Integrating science and resource management in Tampa Bay, Florida. U.S. Geological Survey Circular 1348. https://doi.org/10.3133/cir1348

Zeileis A, Hothorn T (2002) Diagnostic checking in regression relationships. R News 2:7–10

Zeileis A, Kleiber C, Jackman S (2008) Regression models for count data in R. J Stat Softw 27:1–25

Zuur AF, Ieno EN, Walker NJ, Saveliev AA, Smith GM (2009) Mixed effects models and extensions in ecology with R. Springer Science + Business Media, New York, NY

Editorial responsibility: Rochelle D. Seitz, Gloucester Point, Virginia, USA
Reviewer: 3 anonymous referees

Submitted: August 22, 2019;
Accepted: September 1, 2020
Proofs received from author(s): October 19, 2020