Using extreme gradient boosting (XGBoost) to evaluate the importance of a suite of environmental variables and to predict recruitment of young-of-the-year spotted seatrout in Florida

Elizabeth Herdter Smith^{1,2}

Fish and Wildlife Research Institute, Florida Fish and Wildlife Conservation Commission, 100 8th Avenue SE, St. Petersburg, Florida 33701¹

University of South Florida, College of Marine Science, 140 7th Avenue South, St. Petersburg, Florida 33701²

Contact Author: Elizabeth Herdter Smith

Elizabeth.Herdter@myfwc.com

eherdter@mail.usf.edu

Abstract

Environmental factors strongly influence the success of juvenile fish recruitment and productivity, but species-specific environment-recruitment relationships have eluded researchers for decades. Most likely, this is because the environment-recruitment relationship is nonlinear, there are multi-level interactions between factors, and environmental variability may differentially affect recruitment among populations due to spatial heterogeneity. Identifying the most influential environmental variables may result in more accurate predictions of future recruitment and productivity of managed species. Here, gradient tree boosting was implemented using XGBoost to identify the most important predictors of recruitment for six estuary populations of spotted seatrout (*Cynoscion nebulosus*), an economically valuable marine resource in Florida. XGBoost, a machine learning method for regression and classification, was employed because it inherently models variable interactions and seamlessly deals with multi-collinearity, both of which are common features of ecological datasets. Additionally, XGBoost operates at a speed faster than many other gradient boosting algorithms due to a regularization factor and parallel computing functionality. In this application of XGBoost, the results indicate that the abundance of pre-recruit, juvenile spotted seatrout in spatially distinct estuaries is influenced by nearly the same set of environmental predictors. But perhaps of greater importance is that the results of this study show that this algorithm is highly effective at predicting species abundance and identifying important environmental factors (i.e. predictors of recruitment). It is strongly encouraged that future research explore the applicability of the XGBoost algorithm to other topics in marine and fisheries science and compare its performance to that of other statistical methods.

Introduction

Environmental factors such as ENSO cycles, salinity, and temperature can strongly influence the success of fish recruitment worldwide (Hjort 1914; Cushing 1982); however the true environment-recruitment relationship is difficult to identify. Several reasons for this difficulty are described by Rose (2000) and Keyl and Wolff (2008) and summarized here. First, the relationship between environmental factors and recruitment is often nonlinear, and many commonly used statistical methods are unable to model such nonlinearity. Additionally, environmental factors may interact to limit or enhance the recruitment response or mitigate or enhance the effects of other environmental factors. Moreover, environmental variability may differentially affect recruitment among populations or subpopulations due to the spatial heterogeneity of the influence of environmental factors. Further, researchers often evaluate the influence of environmental variables at spatial scales that are incongruous with the spatial scale of recruitment. Finally, an environment-recruitment relationship may change due to a shift in the environmental regime (Rose 2000; Keyl and Wolff 2008; Vert-pre et al. 2013).

Nevertheless, identifying environmental variables that are most influential to recruitment and including these sources of variability in the estimation process may yield more accurate estimates of current and future recruitment (Myers 1998). Moreover, accurate estimates of historical and future recruitment are vital to the sustained productivity of exploited populations because parameters that describe the stock-recruitment relationship are used to calculate biological reference points and fishing mortality targets (Hilborn and Walters 1992). Moreover, identifying highly influential environmental variables can inform adaptive management strategies that accommodate for changes in the environment. Thus, this task is of paramount importance (Collie et al. 2016).

In Florida, annual recruitment of spotted seatrout (*Cynoscion nebulosus*), an estuarine fish of great economic importance to the recreational fishery in the Gulf of Mexico, is characterized by deviations unexplainable by changes in adult biomass. In addition, poor fits of Beverton-Holt and Ricker stock-recruitment curves to adult and young-of-the-year (YOY) abundance is suggestive of an unaccounted for source of process error, likely an environmental factor, in the quantitative assessment of this stock (Murphy et al. 2011). Growth rates and recruitment strength differ among genetically similar but spatially separate estuary populations of spotted seatrout, which is indicative of the likely influence of local environmental variables on recruitment strength of this species (Herdter et al. in review, Murphy and McMichael 2003; Kupschus 2004). Further, previous studies have found significant correlations between the abundance of YOY spotted seatrout in several estuaries in Florida and environmental variables, such as freshwater inflow, sea surface temperature, salinity and seagrass cover (Peebles and Tolley 1988; Holt and Holt 2003; Matheson et al. 2003; Kupschus 2004; Purtlebaugh and Allen 2010; Dutterer et al. 2013; Flaherty-Walia et al. 2015).

Such studies are extremely valuable in the quest to identify an environment-recruitment relationship for spotted seatrout in Florida. However, most researchers have employed traditional statistical methods such as generalized linear modeling or correlation analyses, which assume that the environment-recruitment relationship is stationary over time. In contrast, the true relationship may be nonlinear or non-monotonic in nature; thus, the statistical methods used in the aforementioned studies may have resulted in large type I or type II errors (Keyl and Wolff 2008). Additionally, these studies used only short-term data sets and did not exhaustively examine possible temporal lags; thus, the true relationships or temporal lags may have been undetected. To reduce the chance of identifying a spurious environment-recruitment relationship,

the underlying mechanism of such relationships must be defined and supported by cross-validation, all possible environmental predictors must be screened, and the final relationship must be tested on independent data (Myers 1998; Francis 2006).

The objective of this study was to build predictive models of recruitment for six estuary populations of spotted seatrout in Florida using extreme gradient boosting (XGBoost), a new gradient boosted regression tree algorithm developed by Chen and Guestrin (2016), to illuminate spatial differences in predictor importance. Gradient boosted regression trees were employed because they inherently model variable interactions, incorporate k -fold cross-validation into the training algorithm and evaluate the predictive model using a withheld testing dataset (De'ath and Fabricius 2000; Friedman 2001; Elith et al. 2008). This research expands on the current understanding of an environment-recruitment relationship for spotted seatrout in Florida by simultaneously evaluating all available environmental variables at an appropriate spatial scale and considering their influence at several time lags to account for their effects on both spawning success and YOY survival and recruitment. This research provides the foundation for efforts to evaluate environmental variables within a quantitative framework for spotted seatrout in Florida. Such insight is imperative for effective management of this valuable resource.

Methods

Spotted seatrout abundance data. — Observations of YOY spotted seatrout abundance were used as a proxy for spotted seatrout recruitment and were obtained by the Fisheries Independent Monitoring (FIM) program at the Florida Fish and Wildlife Conservation Commission's (FWC) Fish and Wildlife Research Institute (FWRI) in St. Petersburg (FWRI 2016). The FIM program surveys six major estuaries in Florida using a multi-gear, stratified-

random sampling (SRS) approach. The FIM program routinely monitors the Jacksonville area of northeast Florida (JX) which includes the St. Marys River, Cumberland Sound, the Nassau River and Sound, and the St. Johns River. Apalachicola Bay (AP), Cedar Key (CK), the northern and southern portions of the Indian River Lagoon (Mosquito Lagoon through the Sebastian Inlet and Vero Beach to the Jupiter Inlet; IR), Tampa Bay (TB), and Charlotte Harbor (CH; Figure 1) are also routinely monitored.

The FIM program started sampling for YOY and juvenile fishes in Tampa Bay and Charlotte Harbor in 1989 using 21.3 - meter bag seines. Monthly sampling efforts expanded to the northern region of the Indian River Lagoon, Cedar Key, Apalachicola Bay, and northeast Florida in 1990, 1996, 1998, and 2001, respectively. In 1996, the sampling program started monitoring larger fish, including juvenile and adult spotted seatrout, with 183-meter haul seines (FWRI, 2016). Observations of YOY abundance from all six estuaries except for data from the southern portion of the Indian River Lagoon were used for this study. The data from the southern Indian River Lagoon were excluded because the FIM program does not sample the southern Indian River Lagoon with a 21.3-meter seine; therefore, there are no estimates of YOY abundance in that area. At every haul, hereafter referred to as a sampling event, the number of YOY spotted seatrout in the seine was recorded (a zero was recorded if none were present), and water quality was assessed with a YSI® water quality meter. Young-of-the-year fish were defined as those less than or equal to 100 millimeters standard length (SL; Murphy et al. 2006).

Environmental data. — Environmental factors hypothesized to affect adult spawning success and survival and subsequent settlement of YOY spotted seatrout in each area were included in the analysis, when available (Table 1). Factors likely to affect adult spawning success include salinity (at-spawn salinity; Brown-Peterson and Thomas 1988; Saucier and Baltz 1993),

139 water temperature (at-spawn water temperature; Lowerre-Barbieri et al. 1999; Brown-Peterson
140 2003; Kupschus 2004), and water quality (dissolved inorganic nitrogen as a proxy, [DIN 1prior])
141 in the vicinity of spawning and during months of supposed spawning (Wootton 1998; Lambert
142 and Dutil 2000; Greening and Janicki 2006). However, these variables were included in the
143 analysis for only Tampa Bay, Charlotte Harbor, and the Indian River Lagoon because the
144 spawning locations in these areas have been previously identified (Gilmore 2003; Lowerre-
145 Barbieri et al. 2009; Walters et al. 2009; Table 1). Environmental factors hypothesized to affect
146 survival and settlement of YOY fish include salinity (Holt and Holt 2003; Wuenschel et al.
147 2004), water temperature (Alsuth and Gilmore 1994; Wuenschel et al. 2004), dissolved oxygen
148 (Dissolved O₂; Siefert and Spoor 1974; Breitburg et al. 1994; Miller and Kendall 2009), water
149 clarity (attenuation coefficient as a proxy, [Atten coef]; Fiksen et al. 2002), water depth (Bortone
150 2003), monthly flow rate of the closest river to each sampling event (river flow-CLOSEST),
151 average monthly flow rate of all rivers that drain into the sampling area (river flow ALL;
152 Dutterer et al. 2013), precipitation, the Palmer Z drought index which is an indicator for short-
153 term drought (Z anomaly; King et al. 2003; Dolbeth et al. 2010), and maximum and minimum air
154 temperature anomalies (maximum/minimum T anomaly). In addition, the concentration of
155 dissolved inorganic nitrogen (DIN) was included in the analysis for Tampa Bay, Charlotte
156 Harbor, and the Indian River Lagoon, and chlorophyll *a* (Chlor *a*) concentration was included in
157 the analysis for Tampa Bay. Both DIN and Chlor *a* concentrations serve as proxies for water
158 quality and productivity of trophic levels upon which YOY Spotted seatrout rely but were not
159 available for the remaining estuaries (Livingston 2001; Sigua and Tweeddale 2003; Greening and
160 Janicki 2006; Turner et al. 2006). The influence of environmental variability on YOY spotted

seatrout recruitment and abundance was analyzed on an estuary-specific level rather than statewide to examine spatial differences in an environment-recruitment relationship.

Salinity, water temperature, dissolved oxygen, turbidity, and water depth observations were recorded concomitantly with every sampling event by the FIM program. Nitrogen and chlorophyll *a* concentrations within Tampa Bay were recorded by gauges managed by the Hillsborough County Environmental Protection Commission. Nitrogen concentrations in Charlotte Harbor were recorded by gauges managed by the Lee County Environmental Laboratory, Charlotte Harbor Estuaries Volunteer Water Quality Monitoring Network, and the Peace River Manasota Regional Water Supply Authority. The nitrogen and chlorophyll *a* data were accessed through the Tampa Bay Water Atlas program (<http://www.tampabay.wateratlas.usf.edu/datadownload/SelectStations.aspx>). Nitrogen concentrations in the Indian River Lagoon were recorded by gauges managed by the St. Johns River Water Management District and were accessed using the data portal on their website (<https://www.sjrwmd.com/data/water-qual>). The flow rates recorded at the most downstream gauges with the longest data records for each major river within each estuary were obtained from the United States Geological Survey National Water Information System (USGS NWIS; <https://maps.waterdata.usgs.gov/mapper/>; Table 2). When available, tidally filtered flow rates were preferred over general flow rates if the river was tidally influenced. The Z anomaly, and maximum and minimum T anomalies were obtained for each national climatological division (CD) within which each estuary is located from the National Oceanic and Atmospheric Administration National Centers for Environmental Information (NOAA NCEI; <https://www.ncdc.noaa.gov/cag/divisional/ti>; Table 2). Precipitation data recorded at the weather

station with the longest continuous data record and closest to each estuary were also obtained from NOAA NCEI (<https://gis.ncdc.noaa.gov/maps/ncei/s/>; Table 2).

The attenuation coefficient, DIN (including DIN 1prior), monthly river flow, and monthly precipitation were derived before they were included as environmental variables in the estuary-specific datasets. The attenuation coefficient was not available in the FIM dataset so it was calculated by dividing the Secchi depth recorded concomitantly with every sampling event by 1.7 to obtain the attenuation coefficient, k (Murty 1969). However, the attenuation coefficient was not calculated if the Secchi disk reached the bottom. Nitrogen concentration at gauges in Tampa Bay, Charlotte Harbor, and the Indian River Lagoon were reported in terms of concentrations of ammonia (NH_3), ammonium (NH_4) and variants of nitrate/nitrite (NO_x) so observations of DIN, including DIN 1prior, were produced by summing the concentrations of the three components (NH_3 , NH_4 , NO_x). DIN was not calculated if any of the three components were missing. Monthly river flow rate and precipitation were obtained by averaging daily flow rate and daily precipitation, respectively. Before calculating monthly river flow rate in IR, the flow rates for the north and south prongs of the St. Sebastian River in the Indian River Lagoon were summed to obtain the monthly flow rate of the St. Sebastian River in to the Indian River Lagoon. Additionally, monthly flow rates of the Cedar and Ortega Rivers (oriented similar to the north and south prongs of the St. Sebastian River) which flow in to the north portion of the St. Johns River in northeast Florida (JX) were summed to obtain the total monthly flow rate into the St. Johns River. Monthly flow rates of each river in every estuary were averaged to obtain monthly flow rate of all rivers. Finally, each environmental data type was searched for outliers, and any records outside of the 95th percentile of the distribution were removed.

Observations of DIN, Chlor *a*, river flow, precipitation, Z anomaly, and minimum and maximum T were temporally matched to each sampling event by the month in which each sampling event occurred. DIN was matched to the month prior to the sampling event to account for the time lag between a nutrient loading event and succession of or reduction in phyto- and zooplankton populations (Roelke et al. 1997; Livingston 2001). Observations of at-spawn salinity, at-spawn water temperature, were also temporally matched to each sampling event using the estimated month when the juvenile(s) observed in the sampling event was (were) spawned. Observations of DIN 1prior were matched to the month prior to spawning. Spawning month was estimated by calculating the average age (in months) of juveniles present in each sampling event. The average age was calculated by averaging the total length of juveniles present in each sampling event and applying an age-length relationship for juvenile spotted seatrout (McMichael and Peters 1989). The average age of the fish in each sampling event was then used to estimate an expected spawning month, assuming a pelagic larval duration of two weeks, by subtracting the average age from the sampling month. Therefore, at-spawn salinity, at-spawn water temperature, and DIN 1prior were calculable only when at least one juvenile spotted seatrout was observed in a sampling event.

River flow, precipitation, minimum and maximum T anomalies and the Z anomaly exhibited seasonal lags and were included in each estuary-specific dataset because each variable may influence estuarine circulation and water quality necessary for successful transport, settlement and recruitment of juveniles (Mann and Lazier 1992; Bolle et al. 2009; Dutterer et al. 2013). Average values for each environmental factor during the previous autumn and winter were calculated; the previous autumn encompassed October, November and December of the year prior to each sampling event, and the winter season encompassed January and February of

the sampling year. For example, seasonally lagged flow of the closest river was temporally matched to each sampling event by determining the average river flow during October, November, and December of the previous year (thus, previous autumn river flow) and the average river flow during January and February of the current year (thus, winter river flow). The average flow rate of all rivers into each area was also considered on a seasonally lagging timescale. Additionally, the average concentrations of DIN for the two and three months prior (DIN 2prior and DIN 3prior, respectively) to spawning were also considered because of the possible influence of long-term, lower trophic level productivity on spawning stock health, spawning success, and fecundity (Wootton 1998).

Total DIN, Chlor *a*, river flow, at-spawn salinity, at-spawn water temperature, DIN 1prior, DIN 2prior, and DIN 3prior were also spatially matched to each sampling event. Observations of DIN and Chlor *a* from gauges within a 2.5-mile buffer around each sampling event were selected, and the closest observations were matched to the sampling event if the observations were from the same year and month, or month prior for DIN, of the sampling event. Not all sampling events were assigned DIN or Chlor *a* values because many were located more than 2.5 miles away from the closest gauge. The flow rate of the closest river was assigned to each sampling event using the approximate latitudinal and longitudinal coordinates of all river mouths and spatially distinct sampling events in each estuary. Unlike DIN and Chlor *a*, which were matched to the closest sampling event, DIN 1prior, at-spawn salinity and at-spawn water temperature were calculated by averaging the observations of these variables across gauge locations located within probable spawning grounds and during the estimated spawning period of the fish within each sampling event. Observed DIN 2prior and DIN 3prior were assigned to each

sampling event only if observations of DIN were available in the vicinity of the spawning grounds and during the two and three months leading up to supposed spawning.

Data filtering. — Structural zeros (true zeros; Zuur et al. 2009) were removed from each estuary-specific dataset to reduce zero inflation using two occurrence rate thresholds. A structural zero was defined as a zero count that occurred at a sampling site because of either timing or habitat reasons. First, a zero count may have occurred because it was too early in the year and new juveniles had yet to settle to the nursery habitat and recruit to the sampling gear. Alternatively, sampling may have been too late in the year so that new juveniles had already undergone ontogenetic habitat shifts and went undetected by sampling. Second, a zero count may have occurred because the structural habitat (for example, bottom vegetation, bottom type, shoreline type) or sampling zone was not and would never be suitable for juvenile spotted seatrout even if it were optimally favorable in terms of other water quality characteristics such as temperature or salinity. A 10% occurrence rate threshold was used for sampling month, and an 85% occurrence rate threshold was used for the structural habitat category. For example, if at least one juvenile spotted seatrout was observed only 10% of the time or less within a certain month across all sampling years within an estuary the sampling events performed during those months were removed. Then, if at least one juvenile spotted seatrout was observed only 15% of the time or less within a specific category of bottom vegetation type, bottom type, shoreline type or sampling zone across all sampling years within an estuary, the sampling events performed in those habitat types were also removed. Finally, sampling events where more than 50 juvenile spotted seatrout were captured were lumped together into a 50+ bin.

Evaluating environmental predictor importance with XGBoost. — The importance of each environmental factor (environmental predictor variable) to the recruitment of YOY spotted

seatrout was assessed using gradient boosting regression trees (GBRT) or “gradient tree boosting”, which is a type of classification and regression tree algorithm (CART; Breimen et al. 1993). Classification and regression trees use dominant patterns in predictor variables to partition the response data into homogenous nodes, or leaves, using a series of decisions, i.e., decision rule or decision tree. The resulting decision rule is then used to predict a new series of response data (Breimen et al. 1993; Elith et al. 2008)

Gradient tree boosting is based on principals from classical statistical estimation and machine learning theory (De’ath and Fabricius 2000; Friedman 2001). However, unlike traditional statistical methods that start with a data model, the GBRT algorithm assumes the process by which the response data were produced is complex, unknown, and may have resulted from many inherent variable interactions (Elith et al. 2008). The GBRT algorithm is well-suited for ecological studies exploring predictor importance, particularly those exploring an environment-recruitment relationship, because it can model inherent, multi-level interactions between environmental variables, is generally insensitive to multi-collinearity, can accommodate for outliers and missing observations of predictor variables, and model nonlinearity likely to occur between predictor and response variables following an environmental regime shift (Beamish et al. 1999; De’ath and Fabricius 2000; Polovina 2005). Furthermore, the GBRT algorithm produces a decision rule using k -fold cross-validation that can be applied and tested on a withheld portion of data, which is strongly encouraged when testing environment-recruitment relationships (Myers 1998). Gradient tree boosting has been successfully used by fisheries scientists in the past to predict the abundances of spotted seatrout and southern flounder in Texas as well as demersal fish species richness around New Zealand (Leathwick et al. 2006; Froeschke and Froeschke 2011, 2016).

Here, gradient tree boosting was implemented in the R software environment using XGBoost (extreme gradient boosting; R package *xgboost* [Chen et al. 2018]), which is a machine learning method for regression or classification tree boosting that builds many shallow decision trees. Each individual tree describes only a portion of the system but produces a highly accurate rule for prediction when joined in an ensemble (Chen and Guestrin 2016). The XGBoost algorithm is a unique form of gradient tree boosting because decision trees are built to minimize an objective function composed of not only a loss function but also a regularization factor that prevents tree over-fitting (Chen and Guestrin 2016). Additionally, the XGBoost algorithm works in parallel among all available computing cores in combination with the regularization factor to produce a prediction rule faster than most other tree boosting methods (Chen and Guestrin 2016). XGBoost has been used in the ecological literature to forecast bird migration and identify fishing vessel type based on vessel monitoring system trajectories in the East China Sea (Huang et al. 2018; Van Doren and Horton 2018).

Each estuary-specific dataset was randomly shuffled and split into to a training and testing dataset using a 70%/30% split, where 70% of each shuffled dataset was used to train the XGBoost prediction rule, and the remaining 30% was withheld and used to test the decision rule. The 70%/30% ratio is commonly used to split datasets and is supported by validation studies (Guyon 1997; Kuhn et al. 2018). A count Poisson learning objective and a Poisson negative log-likelihood loss objective function were used to create a prediction rule for each estuary-specific dataset. Each XGBoost prediction rule was trained with 10-fold cross-validation to identify the number of trees (*ntree*) that minimized the objective function. Then, using the optimal number of trees and functions within the R *caret* package (Kuhn et al. 2018), the prediction rule was fine-tuned by identifying the optimal combination of hyperparameters that further minimized the

objective function for each area. The hyperparameters included the learning rate (*eta*), maximum depth of each tree (*max_depth*), the number of observations in each leaf node of the tree (*min_child_weight*), the minimum loss reduction required to further partition a leaf node on a single tree (*gamma*), the proportion of observed data used by XGBoost to grow each tree (*subsample*), and the proportion of predictor variables used at each level of tree splitting (*colsample_bytree*; DMLC 2016). With the optimal values of the hyperparameters and number of trees, the decision rule was retrained and applied to the withheld testing data to predict a new series of count observations and evaluate the accuracy of the decision rule. The gain value (improvement in accuracy each variable brings to the decision rule, i.e., variable importance) of each environmental predictor variable was also obtained using the XGBoost algorithm.

Next, the combination of environmental predictor variables that produced the lowest mean absolute error (MAE) between the observed and predicted count observations was determined with recursive variable addition. The datasets were stripped of all environmental predictor variables except for salinity and temperature. The data were shuffled, and the training procedure described above was used to determine a decision rule. Additional environmental predictor variables were added to the dataset one at a time, in the order of importance defined by the XGBoost algorithm in the previous step. A decision rule was reproduced after every variable addition and applied to the withheld data to predict a new series of count data. Mean absolute error and mean error (ME) were also calculated after the addition of each variable to explore the accuracy and bias, respectively, of each decision rule (Walther and Moore 2005). Last, the optimal combination of environmental predictor variables for each estuary was used to build a final decision rule; MAE and ME were recalculated, and the predictive performance of each environmental variable was ranked by the XGBoost algorithm.

Generalized linear models. — Generalized linear models (GLMs) were fitted to the same datasets and compared to the XGBoost results. Each dataset was once again shuffled and separated into a testing and training dataset, and the count data were modeled with a Poisson distribution. Step wise model selection was used to identify the environmental predictor variables that explained the most deviance, and only variables that reduced the deviance by 0.5% or more were retained in the final model for each estuary. The final model was applied to the withheld testing data to predict a new series of count data. Lastly, MAE and ME between the predicted and observed data were calculated for each estuary-specific model.

Results

A total of 902 sampling events in Apalachicola, 2034 sampling events in Cedar Key, 7239 sampling events in Tampa Bay, 6038 sampling events in Charlotte Harbor, 1996 sampling events in northeast Florida, and 6449 sampling events in the Indian River Lagoon along with the associated observations of environmental predictor variables were used to build estuary-specific predictive models of YOY spotted seatrout recruitment (Table 3). Young-of-the-year spotted seatrout were most abundant in Apalachicola, Tampa Bay, Charlotte Harbor, and the Indian River Lagoon where, two YOY spotted seatrout were captured per sampling event on average (Table 3; Figure 2). In contrast, YOY fish were less abundant in Cedar Key and northeast Florida where only one individual was captured per sampling event on average (Table 3; Figure 2).

Observations of salinity, water temperature, dissolved O₂, depth, precipitation, Z anomaly, and maximum and minimum temperature were available for nearly every sampling event in each estuary (Table 3). River flow rates were missing for a small portion of the sampling events in northeast Florida and the Indian River Lagoon due to either missing river gauge

observations or data that did not span the entire sampling timeframe. In contrast, the attenuation coefficient was calculable only 38% of the time, on average, because of frequent sampling in shallow areas where the Secchi disk remained visible when it reached the bottom (Table 3). Additionally, observations of spatially matched variables including concentrations of DIN and chlorophyll *a*, at-spawn salinity, at-spawn water temperature, and DIN 1prior were infrequent throughout the datasets. Observed DIN was available for only 53% of the sampling events in Tampa Bay and 13% of the sampling events in Charlotte Harbor and Indian River Lagoon on average. Observed chlorophyll *a* was available for only 36% of the sampling events in Tampa Bay. Additionally, at-spawn salinity and at – spawn water temperature were matched to only 26% of the sampling events in Tampa Bay, Charlotte Harbor, and the Indian River Lagoon, on average. Additionally, DIN 1prior, DIN 2prior, and DIN 3prior were matched to only 18% of the sampling events in these estuaries, on average.

Salinity, water temperature, dissolved O₂ and precipitation were generally consistent among areas except for in Cedar Key where the mean salinity and temperature were lowest, mean depth and the attenuation coefficient varied considerably among all areas (Figure 2). In general, the attenuation coefficient was larger in deeper estuaries as observed in northeast Florida (Figure 2). Additionally, concentration of DIN, river flow rate, at-spawn salinity, at-spawn water temperature, and seasonally lagged precipitation varied widely among areas. The mean concentration of DIN was highest in the Indian River Lagoon and the mean flow rates (including seasonally lagged variants) were highest among rivers in northeast Florida. Both at-spawn salinity and at-spawn water temperature were highest in Tampa Bay and lowest in Apalachicola, while seasonally lagged precipitation was highest in Apalachicola (winter precipitation) and the

Indian River Lagoon (previous autumn precipitation) and lowest in Tampa Bay and Charlotte Harbor (previous autumn precipitation and winter precipitation; Figure 2).

The XGBoost algorithm was used to evaluate the predictive performance of 22 environmental variables in Apalachicola, 23 environmental variables in Cedar Key, 29 environmental variables in Charlotte Harbor, 30 environmental variables Tampa Bay, 23 environmental variables in northeast Florida, and 29 environmental predictors in Indian River Lagoon (Table 3). The XGBoost algorithm identified 16 environmental variables that best predicted YOY abundance in Apalachicola, 20 variables for Cedar Key, 28 for Tampa Bay, 26 for Charlotte Harbor, 12 for northeast Florida, and 25 for the Indian River Lagoon (see Table 4 for a list of variables providing gain values greater than 0.01). Salinity, temperature, depth, and dissolved O₂ exhibited the greatest gains in predictive performance for all areas. River flow during the previous autumn was the most important predictor variable for Cedar Key, and monthly flow of the closest river was among the top ten most important predictor variables for all areas. Precipitation and the Z anomaly were also among the top ten most important predictors for all areas except for northeast Florida and the Indian River Lagoon where the Z anomaly provided a gain of less than 0.05 (Table 4; Figure 3).

In Apalachicola, the top ten most important environmental predictors were salinity (gain = 0.17), water temperature (gain = 0.15), dissolved O₂ (gain = 0.14), depth (gain = 0.07), attenuation coefficient (gain = 0.06), river flow (gain = 0.05), minimum temperature anomaly (gain = 0.05), Z anomaly (gain = 0.04), average river flow during the previous autumn (gain = 0.03), and precipitation (gain = 0.03). In addition, precipitation during the previous autumn and winter seasons, maximum temperature anomaly, winter river flow, and sampling year and month provided a combined gain of 0.12 (Table 4; Figure 3). In Cedar Key, the top ten most important

environmental predictors were average river flow during the previous autumn (gain = 0.14), dissolved O₂ (gain = 0.14), salinity (gain = 0.11), depth (gain = 0.11), temperature (gain = 0.09), year (gain = 0.08), attenuation coefficient (gain = 0.07), river flow (gain = 0.05), precipitation (gain = 0.04), and Z anomaly (gain = 0.03). In addition, maximum and minimum temperature anomalies, river flow of all rivers, sampling month, average river flow during the previous autumn and winter, average maximum temperature anomaly during the previous autumn and winter, and average Z anomaly during the winter provided a combined gain of 0.13 (Table 4; Figure 3). In Tampa Bay, the top ten most important environmental predictors were salinity (gain = 0.14), dissolved O₂ (gain = 0.12), water temperature (gain = 0.11), concentration of DIN (gain = 0.08), depth (gain = 0.07), concentration of chlorophyll *a* (gain = 0.06), river flow (gain = 0.05), precipitation (gain = 0.03), river flow of all rivers (gain = 0.03), and the Z anomaly (gain = 0.03). In addition, minimum and maximum temperature anomalies, average river flow during the previous autumn and winter, the attenuation coefficient, sampling year and month, average minimum temperature anomaly during winter, average Z anomaly and maximum temperature anomaly during the previous autumn provided a combined gain of 0.19 (Table 4; Figure 3). In Charlotte Harbor, the top ten most important environmental predictors were salinity (gain = 0.19), dissolved O₂ (gain = 0.13), temperature (gain = 0.13), depth (gain = 0.07), precipitation (gain = 0.05), river flow (gain = 0.05), river flow of all rivers (gain = 0.05), sampling month (gain = 0.04), Z anomaly (gain = 0.04), and average river flow during the previous autumn (gain = 0.04). In addition, minimum and maximum temperature anomalies, average river flow during winter, concentration of DIN, the attenuation coefficient, average maximum temperature anomaly during winter, average Z anomaly during the previous autumn and sampling year provided a combined gain of 0.15 (Table 4; Figure 3).

In northeast Florida, the top ten most important environmental predictors were salinity (gain = 0.16), temperature (gain = 0.14), attenuation coefficient (gain = 0.12), dissolved O₂ (gain = 0.11), depth (gain = 0.09), river flow (gain = 0.08), precipitation (gain = 0.07), river flow of all rivers (gain = 0.07), and average river flow during winter (all and closest, gain = 0.05 and gain = 0.07, respectively). In addition, average river flow of the closest and all rivers during the previous autumn provided a combined gain of 0.05 (Table 4; Figure 3). In the Indian River Lagoon, the top ten most important environmental predictors were temperature (gain = 0.15), dissolved O₂ (gain = 0.15), salinity (gain = 0.13), depth (gain = 0.07), river flow (gain = 0.07), precipitation (gain = 0.06), the attenuation coefficient (gain = 0.04), average river flow during the previous autumn (gain = 0.04), river flow of all rivers (gain = 0.03), and concentration of DIN (gain = 0.03). In addition, the Z anomaly, average river flow, precipitation, and maximum temperature anomaly, Z anomaly during winter, maximum temperature anomaly, minimum temperature anomaly, sampling year and month provided a combined gain of 0.15 (Table 4; Figure 3).

The mean absolute error between the observed and predicted count data in all estuaries indicated good predictive performance by the XGBoost decision rule. The XGBoost algorithm produced a decision rule predicted the count data distribution more accurately than the GLMs (MAE = 2.15 vs 3.19 for AP; 0.90 vs 1.08 for CK; 2.75 vs 3.23 for TB; 1.56 vs. 1.99 for CH; 0.53 vs 0.71 for JX; 1.96 vs 2.37 for IR; Table 5). However, the distributions were slightly more biased (higher ME) than those produced by the GLMs (0.90 vs -0.94 for AP; 0.53 vs 0.09 for CK; 1.06 vs 0.04 for TB; 0.60 vs -0.18 for CH; 0.48 vs -0.02 for JX; 0.8 vs -0.11 for IR; Table 5).

Discussion

This study synthesized the results of a large body of research on the environmental factors affecting spotted seatrout and ranked the importance of environmental predictor variables using a new gradient tree boosting algorithm. The relative importance of each environmental predictor variable was consistent among areas, and any differences are reflective of estuary-specific differences in geomorphology, circulation patterns and habitat characteristics. Salinity, temperature, dissolved O₂, and depth were all among the five most important predictor variables for all estuaries. In addition, river flow was among the seven most important environmental predictor variables in all estuaries, except for in Cedar Key where it was the most important variable. Precipitation and drought were among the ten most important predictor variables in all estuaries.

Spotted seatrout are euryhaline species and have been found in waters ranging widely in salinity and temperature (Bortone 2003). However, rapidly changing water conditions can impart lethal osmoregulatory stress on larvae and juveniles as well as limit metabolic activity and oxygen transfer ability and, therefore, strongly influence the distribution and abundance of juveniles and adults (Banks et al. 1991; Alsuth and Gilmore 1994; Holt and Holt 2003; Whitfield and Harrison 2003; Wuenschel et al. 2004). Salinity and temperature conditions also greatly influence spawning success by affecting fertilization and hatching rates, egg buoyancy, and the metabolic rates of reproductive adults (Holliday 1969; Fry 1971; Gray et al. 1991; Holt and Holt 2003; Kupschus 2004). In addition, the concentration of dissolved oxygen affects growth, survival, and respiration of larval and juvenile fishes and can influence predator-prey interactions by altering the vertical distribution of species (Breitburg et al. 1994, 1999; Secor, D.H.; Gunderson 1998). Our results echo those of Froeschke and Froeschke (2011), who also used gradient boosting to explore a suite of environmental predictor variables for spotted seatrout

populations in Texas and found that salinity, temperature, turbidity, and dissolved oxygen were the most important variables among nine estuary populations.

The influx of freshwater from rivers and other point sources can significantly alter the salinity, temperature, dissolved oxygen, turbidity, light availability, and nutrient conditions within estuary systems thereby affecting the abundance and distribution of juvenile fishes (Kimmerer 2002; Halliday et al. 2008; Gillanders et al. 2012; Whaley et al. 2016; de Mutsert et al. 2017). Variability in freshwater flow can also impact spawning success and the subsequent shoreward larval transport and settlement to essential nursery habitat by altering the salinity gradient, stability of the halocline, and estuary circulation (Drinkwater and Frank 1994; Pineda et al. 2007; Gillson 2011). In addition, freshwater flow can impact the quality and distribution of nursery habitat as salinity intrusions or even overall habitat loss can occur during reduced flow or drought conditions (Drinkwater and Frank 1994; Gillson 2011; Whaley et al. 2016). Our results support the findings of previous research on spotted seatrout and freshwater flow in Florida. Purtlebaugh and Allen (2010) found a positive relationship between freshwater flow from the Suwannee River and abundance of YOY spotted seatrout. In addition, Matheson et al. (2003) and Flaherty et al. (2015) found a positive relationship between YOY spotted seatrout abundance and freshwater flow in Tampa Bay.

Flow rate of the closest river offered greater gain than the average flow rate of all rivers for each estuary with one exception. Average nutrient and water quality conditions in estuaries are influenced by water circulation which is controlled by tides, winds, and the salinity gradient between the less saline freshwater and the resident salt water (Mann and Lazier 2013). As such, estuarine water quality may remain heterogeneous even when the average flow rate of all rivers increases. For example, the Indian River Lagoon is long and narrow and has relatively sluggish

wind driven circulation with limited tidal flushing. Therefore, while portions of the Indian River Lagoon may suddenly receive increased nutrient-laden freshwater, other areas and the surrounding flora and fauna may remain unaffected (Sigua and Tweedale 2003). Additionally, major salinity gradients limit the circulation among Tampa Bay's four major segments, thus, the nutrient-laden freshwater flowing into the northern portion of the bay may not necessarily result in a net increase in nutrients in the lower portions of the bay (Greening and Janicki 2006). Therefore, the flow rate of the river closest to the capture location or home range of a pre-recruit or juvenile spotted seatrout is more influential to the surrounding water quality and nutrient conditions than the average flow rate of all rivers to an estuary.

In contrast, the average flow of both rivers in Cedar Key (Suwannee River and Waccasassa River) during the previous autumn was the most important environmental predictor variable. We hypothesize that this result occurred because the average flow rate of the Suwannee River (which flows at the second highest rate of all rivers in Florida) from 1996 through 2015 ($8175 \text{ f}^3/\text{s}$) was nearly 33 times that of the Waccasassa River ($245 \text{ f}^3/\text{s}$) therefore, the juvenile spotted seatrout that were captured closer to the mouth of the Waccasassa River in the southern portion of the Cedar Key sampling area were influenced by freshwater from both the Suwannee River and the Waccasassa River. The results from Purtlebaugh and Allen (2010) corroborate our findings. They found that spotted seatrout in the southern portion of the Cedar Key sampling area were strongly influenced by freshwater from the Suwannee River. However, the seasonal lag contrasts with that identified by Purtlebaugh and Allen (2010), who found that river flow from March through May explained most of the variation in spotted seatrout abundance. This current study evaluated average flow rates from only October through December and January through February. Our results may be more comparable had we evaluated average flow rates from March

through May, as evaluated in Purtlebaugh and Allen (2010). Average flow rates of the Suwannee and Waccasassa Rivers peak between February and April, and a smaller peak occurs between August and October (Purtlebaugh and Allen 2010); thus, it is unclear why the average winter flow rate, which includes the month of February in this study, was less important. One hypothesis is that the influx of nutrient-laden freshwater in autumn sets the stage for the productivity of the Cedar Key estuary in the coming year by affecting upper level productivity and spawning stock health. However, future studies are necessary to explore a link between autumn conditions and juvenile spotted seatrout abundance in Cedar Key.

In general, river flow is strongly influenced by precipitation and drought. Although decreased precipitation during drought conditions may result in decreased river flow rates, the two variables may not always correlate. In Florida, minimum flow and water levels of major water resources are maintained by water management districts to mitigate the effects of water withdrawals, and conversely, to prevent flooding (Olexa et al. 2017). While patterns of increased precipitation or short-term drought conditions may alter salinity gradients or the distribution and availability of essential nursery habitat, precipitation and drought cycles are not directly responsible for the influx of nutrients to the estuaries. Therefore, both precipitation and drought, while among the ten most important variables, were less important than river flow in all estuaries.

The importance of several environmental predictors varied among estuaries. In Tampa Bay, total dissolved inorganic nitrogen as well as chlorophyll *a* were more important than river flow, as excess nutrients from agricultural runoff or increased river flow stimulate plant growth and can induce oxygen-depleting, algal blooms that may result in hypoxic or anoxic conditions. In addition, increased nutrients can induce a shift from macrophyte to phytoplankton or algal-

based communities (Sigua and Tweedale 2003; Greening and Janicki 2006). For example, increased nutrient loading in Tampa Bay during the late 1970's and early 1980's was responsible for the precipitous decline in seagrass, which is important habitat for juvenile spotted seatrout (Flaherty-Walia et al. 2015). While highly important to juvenile spotted seatrout abundance in Tampa Bay, DIN was of low importance in Charlotte Harbor and the Indian River Lagoon. This discrepancy is possibly a result of disproportionate number of observations among estuaries. While more than half (54%) of the sampling events in Tampa Bay had associated DIN observations, only 14% of the sampling events in the Indian River Lagoon and 10% of the sampling events in Charlotte Harbor had such observations. While the XGBoost algorithm can infer missing values to a certain degree (Chen and Guestrin 2016), the paucity of observations in Charlotte Harbor and the Indian River Lagoon may have exceeded the ability of the XGBoost algorithm to identify a decision rule based on limited observations. The predictive performance of the XGBoost decision rule tuned to varying levels of missing data requires further investigation.

Light attenuation, which is partially a function of phytoplankton biomass, turbidity and eutrophication affects the distribution and abundance of seagrass, which is important habitat structure for pre-recruits and juvenile fish (Flaherty-Walia et al. 2015). In addition, light availability affects the distribution and abundance of benthic macro-algae which is a basal resource for many estuarine species including juvenile spotted seatrout (Burghart et al. 2013). An attenuation coefficient was not calculated if the Secchi disk was still visible at the bottom of the water column; thus, sampling events in deeper, less clear water were more likely to have an associated attenuation coefficient than those in shallower, clearer water. For example, the attenuation coefficient was the third most important variable in northeast Florida where it was

missing for only 26% of the sampling events. In contrast, in shallower or clearer estuaries such as Tampa Bay and Charlotte Harbor, only an average of 15.5% of the sampling events had associated attenuation coefficients, thus, the XGBoost algorithm had few observations to inform the decision rule training process. Direct measures of the factors contributing to light attenuation in estuaries including turbidity, phytoplankton biomass or water color, should be evaluated in future predictive studies to provide a more complete series of environmental predictors variables and more data with which to tune a decision rule.

Among all estuaries, maximum and minimum air temperature anomalies and their seasonally lagged versions were of little importance. The temperatures of water and air change at different rates, so seasonally anomalous air temperature conditions are not immediately reflected by the water temperature. Air temperature anomalies have little predictive power when evaluated on a monthly and or seasonal timeframe but may have greater predictive power if they are considered on a wider timeframe. This hypothesis requires further investigation. In addition, temperature, salinity, and DIN at estimated time of spawning were not included in the final XGBoost decision rule for any estuary dataset on which they were evaluated. While temperature and salinity affect spawning success in spotted seatrout (Kupschus 2004), the relationship between these variables and juvenile abundance was not captured by the XGBoost decision rule, which is likely a result of the limited observations and generalizations in supposed spawning areas, timing and average environmental conditions in spawning grounds.

While the tolerance of the XGBoost algorithm to missing values requires further investigation, the XGBoost algorithm produced a decision rule that predicted the count distribution more accurately than a traditional GLM. The use of gradient boosting in the fisheries and marine ecology literature has increased over the past ten years. A Google Scholar search to

find articles with the words “fish”, “marine” and the exact phrase “gradient boosting” published between 2008 and 2012 yielded 142 results, while the same search for articles published between 2013 and 2018 yielded 316 results. As the XGBoost algorithm is new to the machine learning literature, there are few published studies in the fields of fisheries or marine ecology that detail XGBoost usage. A Google Scholar search to find articles with the words “fish”, “marine”, and “XGBoost” published between 2014 and 2018 yielded 6 results, but the results of this study show that this algorithm has strong predictive power when evaluating species abundance and identifying important environmental factors. It is strongly encouraged that future research explore the applicability of the XGBoost algorithm to other topics in marine and fisheries science and compare its performance to that of other statistical methods.

This study showed that the abundance of pre-recruit, juvenile spotted seatrout in spatially distinct estuaries is influenced by nearly the same set of environmental predictors and each predictor was similarly important. Water characteristics including salinity, temperature, concentration of dissolved oxygen, and water quality are most important. River flow is also highly important. While the goal of this study was to evaluate and rank the importance of every possible environmental variable, toxic algal blooms and seagrass acreage can also affect the growth and mortality of juveniles (Flaherty and Landsberg 2011; Flaherty-Walia et al. 2015). These variables were not included in this study, and future research should evaluate their importance in a decision rule framework. The results of this study can inform future assessment and management efforts for spotted seatrout. The environment can strongly affect population productivity, so the inclusion of highly important environmental variables in future stock assessments should be evaluated. Such inclusion may explain a portion of the process error

common to many modern stock assessments and afford more accurate estimates of important population parameters and management reference points for this species.

Acknowledgments

The survey data for spotted seatrout were provided courtesy of the Fisheries Independent Monitoring program of the Florida Fish and Wildlife Conservation Commission's Fish and Wildlife Research Institute (FWRI) in St. Petersburg, FL. The author would like to acknowledge Dr. Robert Muller and Shanae Allen of FWRI, who provided suggestions on methodology that greatly helped this manuscript. Special thanks go to Ms. Sue Connors and Dr. Margaret Guyette of the St. Johns River Water Management District who assisted with data procurement for the Indian River Lagoon. This manuscript was completed while E.H.S. was employed by FWRI.

References

- Alsuth, S., and G. Gilmore. 1994. Salinity and temperature tolerance limits for larval Spotted Seatrout *Cynoscion nebulosus*. ICES Council Meeting Papers, ICES-CM-1994/L:17, Biol. Oceanogr. Cttee.
- Banks, M. A., G. J. Holt, and J. M. Wakeman. 1991. Age-linked changes in salinity tolerance of larval spotted seatrout (*Cynoscion nebulosus*, Cuvier). Journal of Fish Biology 39(4):505–514.
- Beamish, R. J., D. J. Noakes, G. a McFarlane, L. Klyashtorin, V. V Ivanov, and V. Kurashov. 1999. The regime concept and natural trends in the production of Pacific salmon. Canadian Journal of Fisheries and Aquatic Sciences 56:516–526.
- Bolle, L. J., M. Dickey-Collas, J. K. L. Van Beek, P. L. A. Erftemeijer, J. I. Witte, H. W. Van Der Veer, and A. D. Rijnsdorp. 2009. Variability in transport of fish eggs and larvae. III. Effects of hydrodynamics and larval behaviour on recruitment in plaice. Marine Ecology Progress Series 390:195–211.
- Bortone, S. 2003. Biology of the Spotted Seatrout. CRC Press, Boca Raton.
- Breimen, L., J. Freidman, C. J. Stone, and O. R.A. 1993. Classification and Regression Trees. Machine Learning 3(3):81–106.
- Breitburg, D. L., K. A. Rose, and J. H. Cowan. 1999. Linking water quality to larval survival: Predation mortality of fish larvae in an oxygen-stratified water column. Marine Ecology Progress Series 178:39–54.
- Breitburg, D., N. Steinberg, S. DuBeau, C. Cooksey, and E. Houde. 1994. Effects of low dissolved oxygen on predation of estuarine fish larvae. Marine Ecology Progress Series 104:235–246.

- 655 Brown-Peterson, N. J. 2003. The reproductive biology of Spotted seatrout. Pages 99–133 in S.
656 Bortone, editor. Biology of the Spotted Seatrout. CRC Press, Boca Raton.
- 657 Brown-Peterson, N. J., and P. Thomas. 1988. Differing reproductive life histories between
658 temperate and subtropical groups of *Cynoscion nebulosus*. Continental Marine Science
659 30:71–78.
- 660 Burghart, S. E., D. L. Jones, and E. B. Peebles. 2013. Variation in estuarine consumer
661 communities along an assembled eutrophication gradient: implications for trophic
662 instability. Estuaries and Coasts 36:951–965.
- 663 Chen, T., and C. Guestrin. 2016. XGBoost, a scalable tree boosting system. Pages 785–794
664 Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery
665 and Data Mining - KDD '16.
- 666 Chen, T., T. He, M. Benesty, V. Khotilovich, and Y. Tang. 2018. xgboost: Extreme Gradient
667 Boosting. R package version 0.6.4.1.
- 668 Collie, J. S., L. W. Botsford, A. Hastings, I. C. Kaplan, J. L. Largier, P. A. Livingston, É.
669 Plagányi, K. A. Rose, B. K. Wells, and F. E. Werner. 2016. Ecosystem models for fisheries
670 management: Finding the sweet spot. Fish and Fisheries 17(1):101–125.
- 671 Cushing, D. H. 1982. Climate and fisheries. Academic Press, New York.
- 672 De'ath, G., and K. E. Fabricius. 2000. Classification and regression trees: A powerful yet simple
673 technique for ecological data analysis. Ecology 81(11):3178–3192.
- 674 DMLC. 2016. Scalable and Flexible Gradient Boosting.
675 <http://xgboost.readthedocs.io/en/latest/parameter.html>.
- 676 Dolbeth, M., F. Martinho, V. Freitas, S. Costa-Dias, J. Campos, and M. Â. Pardal. 2010. Multi-
677 year comparisons of fish recruitment, growth and production in two drought-affected
678 Iberian estuaries. Marine and Freshwater Research 61(12):1399–1415.
- 679 Drinkwater, K. F., and K. T. Frank. 1994. Effects of river regulation and diversion on marine fish
680 and invertebrates. Aquatic Conservation: Marine and Freshwater Ecosystems 4(2):135–151.
- 681 Dutterer, A. C., C. Mesing, R. Cailteux, M. S. Allen, W. E. Pine, and P. A. Strickland. 2013. Fish
682 recruitment is influenced by river flows and floodplain inundation at Apalachicola river,
683 Florida. River Research and Applications 29(9):1110–1118.
- 684 Elith, J., J. R. Leathwick, and T. Hastie. 2008. A working guide to boosted regression trees.
685 Journal of Animal Ecology 77(4):802–813.
- 686 Fiksen, Ø., D. L. Aksnes, M. H. Flyum, and J. Giske. 2002. The influence of turbidity on growth
687 and survival of fish larvae: A numerical analysis. Hydrobiologia 484:49–59.
- 688 Flaherty-Walia, K. E., R. E. Matheson, and R. Paperno. 2015. Juvenile Spotted seatrout
689 (*Cynoscion nebulosus*) habitat use in an eastern Gulf of Mexico estuary: the effects of
690 seagrass bed architecture, seagrass species composition, and varying degrees of freshwater
691 influence. Estuaries and Coasts 38(1):353–366.
- 692 Flaherty, K. E., and J. H. Landsberg. 2011. Effects of a persistent red tide (*Karenia brevis*)
693 bloom on community structure and species-specific relative abundance of nekton in a Gulf
694 of Mexico estuary. Estuaries and Coasts 34(2):417–439.
- 695 Francis, R. I. C. C. 2006. Measuring the strength of environment- recruitment relationships: the
696 importance of including predictor screening within cross-validations. ICES Journal of
697 Marine Science 63:594–599.
- 698 Friedman, J. H. 2001. Greedy function approximation: A gradient boosting machine. Annals of
699 Statistics 29(5):1189–1232.
- 700 Froeschke, J. T., and B. F. Froeschke. 2011. Spatio-temporal predictive model based on

environmental factors for juvenile spotted seatrout in Texas estuaries using boosted regression trees. *Fisheries Research* 111(3):131–138.

Froeschke, J. T., and B. F. Froeschke. 2016. Two-stage boosted regression tree model to characterize southern flounder distribution in Texas estuaries at varying population sizes. *Marine and Coastal Fisheries* 8(1):222–231.

Fry, F. E. J. 1971. The effect of environmental factors on the physiology of fish. *Fish Physiology* 6(C):1–98.

FWRI. 2016. Fisheries-independent Monitoring program 2015 Annual Data Summary Report. IHR-2011-001. St. Petersburg.

Gillanders, B. M., B. a. Black, M. G. Meekan, and M. a. Morrison. 2012. Climatic effects on the growth of a temperate reef fish from the Southern Hemisphere: a biochronological approach. *Marine Biology* 159(6):1327–1333.

Gillson, J. 2011. Freshwater flow and fisheries production in estuarine and coastal systems: Where a drop of rain is not lost. *Reviews in Fisheries Science* 19(3):168–186.

Gilmore, G. 2003. Sound production and communication in the Spotted Seatrout. Pages 177–193 in S. Bortone, editor. *Biology of the Spotted Seatrout*. CRC Press, Boca Raton.

Gray, J. D., T. L. King, and R. L. Colura. 1991. Effect of temperature and hypersalinity on hatching success of spotted seatrout eggs. *The Progressive Fish-Culturist* 53(2):81–84.

Greening, H., and A. Janicki. 2006. Toward reversal of eutrophic conditions in a subtropical estuary: Water quality and seagrass response to nitrogen loading reductions in Tampa Bay, Florida, USA. *Environmental Management* 38(2):163–178.

Guyon, I. 1997. A scaling law for the validation-set training-set size ratio. *AT&T Bell Laboratories*:1–11.

Halliday, I. A., J. B. Robins, D. G. Mayer, J. Staunton-Smith, and M. J. Sellin. 2008. Effects of freshwater flow on the year-class strength of a non-diadromous estuarine finfish, king threadfin (*Polydactylus macrochir*), in a dry-tropical estuary. *Marine and Freshwater Research* 59(2):157–164.

Hilborn, R., and C. J. Walters. 1992. *Quantitative Fisheries Stock Assessment*. Chapman and Hall, New York.

Hjort, J. 1914. Fluctuations in the great fisheries of northern Europe viewed in the light of biological research. *Rapports et Procès-Verbaux des Réunions du Conseil Permanent International pour l'Exploration de la Mer* 20(1):228.

Holliday, F. G. T. 1969. The effects of salinity on the eggs and larvae of teleosts. *Fish Physiology: Excretion, Ionic Regulation, and Metabolism* 1:293–311.

Holt, G. J., and S. A. Holt. 2003. Effects of variable salinity on reproduction and early life stages of Spotted Seatrout. Pages 135–145 in S. A. Bortone, editor. *Biology of the Spotted Seatrout*. CRC Press, Boca Raton.

Huang, H., F. Hong, J. Liu, C. Liu, Y. Feng, and Z. Guo. 2018. FVID: Fishing Vessel Type Identification Based on VMS Trajectories. *Journal of Ocean University of China* 17(5):1–10.

Keyl, F., and M. Wolff. 2008. Environmental variability and fisheries: what can models do? *Reviews in Fish Biology and Fisheries* 18(3):273–299.

Kimmerer, W. J. 2002. Effects of freshwater flow on abundance of estuarine organisms: Physical effects or trophic linkages? *Marine Ecology Progress Series* 243:39–55.

King, A. J., P. Humphries, and P. S. Lake. 2003. Fish recruitment on floodplains: the roles of patterns of flooding and life history characteristics. *Canadian Journal of Fisheries and*

- 747 Aquatic Sciences 60(7):773–786.
- 748 Kuhn, M., C. From, S. Weston, C. Keefer, A. Engelhardt, T. Cooper, Z. Mayer, B. Kenkel, R. C.
- 749 Team, M. Benesty, R. Lescarbeau, A. Ziem, L. Scrucca, Y. Tang, and C. Candan. 2018.
- 750 caret: Classification and Regression Training. R package version 6.0-79.
- 751 Kupschus, S. 2004. A temperature-dependent reproductive model for Spotted seatrout
- 752 (*Cynoscion nebulosus*) explaining spatio-temporal variations in reproduction and young-of-
- 753 the-year recruitment in Florida estuaries. ICES Journal of Marine Science 61(1):3–11.
- 754 Lambert, Y., and J.-D. Dutil. 2000. Energetic consequences of reproduction in Atlantic cod (
- 755 *Gadus morhua*) in relation to spawning level of somatic energy reserves. Canadian Journal
- 756 of Fisheries and Aquatic Sciences 57(4):815–825.
- 757 Leathwick, J. R., J. Elith, M. P. Francis, T. Hastie, and P. Taylor. 2006. Variation in demersal
- 758 fish species richness in the oceans surrounding New Zealand: An analysis using boosted
- 759 regression trees. Marine Ecology Progress Series 321:267–281.
- 760 Livingston, R. 2001. Eutrophication Processes in Coastal Systems. CRC Press, Boca Raton.
- 761 Lowerre-Barbieri, S. K., L. R. Barbieri, and J. J. Alberts. 1999. Reproductive parameters needed
- 762 to evaluate recruitment overfishing of Spotted Seatrout in the southeastern U.S., Final report
- 763 to the Saltonstall-Kennedy (S-K) Grant Program (grant No. NA77FD0074).
- 764 Lowerre-Barbieri, S. K., N. Henderson, J. Llopiz, S. Walters, J. Bickford, and R. Muller. 2009.
- 765 Defining a spawning population (spotted seatrout *Cynoscion nebulosus*) over temporal,
- 766 spatial, and demographic scales. Marine Ecology Progress Series 394:231–245.
- 767 Mann, K. H., and J. R. Lazier. 1992. Dynamics of marine ecosystems: Biological-physical
- 768 interactions in the ocean. Reviews in Fish Biology and Fisheries 2(1):86–87.
- 769 Mann, K. H., and J. R. N. Lazier. 2013. Dynamics of Marine Ecosystems: Biological-Physical
- 770 Interactions in the Oceans: Third Edition. Page Dynamics of Marine Ecosystems:
- 771 Biological-Physical Interactions in the Oceans: Third Edition.
- 772 Matheson, R. E., R. H. McMichael, D. Leffler, and T. MacDonald. 2003. Populations of juvenile
- 773 and small-adult fishes in Tampa Bay: a decadal perspective. Pages 1–18 in S. F. Treat,
- 774 editor. The Fourth Tampa Bay Area Scientific Information Symposium 4. St. Petersburg.
- 775 McMichael, R. H., and K. M. Peters. 1989. Early life history of Spotted seatrout, *Cynoscion*
- 776 *nebulosus* (Pisces: Sciaenidae), in Tampa Bay, Florida. Estuaries 12(2):98–110.
- 777 Miller, B. S., and A. W. Kendall. 2009. Early Life History of Marine Fishes. University of
- 778 California Press.
- 779 Murphy, M. D., D. Chagaris, and D. Addis. 2011. An assessment of the status of Spotted seatrout
- 780 in Florida waters through 2009, In-House Report IHR 2011-002. St. Petersburg.
- 781 Murphy, M. D., C. Guenther, and B. Mahmoudi. 2006. An assessment of the status of Spotted
- 782 seatrout in Florida waters through 2005, In-House Report IHR2006-017. St. Petersburg.
- 783 Murphy, M. D., and R. H. McMichael. 2003. Age determination and growth of Spotted Seatrout,
- 784 *Cynoscion nebulosus*. Pages 41–56 in S. A. Bortone, editor. Biology of the Spotted
- 785 Seatrout. CRC Press, Boca Raton.
- 786 Murty, A. 1969. A theoretical approach to the attenuation coefficient of light in sea water.
- 787 Indian Journal of Fisheries 16(1–2):151–155.
- 788 de Mutsert, K., K. Lewis, S. Milroy, J. Buszowski, and J. Steenbeek. 2017. Using ecosystem
- 789 modeling to evaluate trade-offs in coastal management: Effects of large-scale river
- 790 diversions on fish and fisheries. Ecological Modelling 360:14–26.
- 791 Myers, R. A. 1998. When do environment-recruitment correlations work? Reviews in Fish
- 792 Biology and Fisheries 8:285–305.

- Olexa, M., T. Borisova, and J. Davis. 2017. Handbook of Florida Water Regulation: Florida Water Management Districts.
- Peebles, E. B., and S. G. Tolley. 1988. Distribution, growth and mortality of larval Spotted seatrout, *Cynoscion nebulosus* : A comparison between two adjacent estuarine areas of southwest Florida. *Bulletin of Marine Science* 42(3):397–410.
- Pineda, J., J. Hare, and S. Sponaugle. 2007. Larval transport and dispersal in the coastal ocean and consequences for population connectivity. *Oceanography* 20(3):22–39.
- Polovina, J. J. 2005. Climate ariation, regime shifts, and implications for sustainable fisheries. *Science* 76(2):233–244.
- Purtlebaugh, C. H., and M. S. Allen. 2010. Relative abundance, growth, and mortality of five age-0 estuarine fishes in relation to discharge of the Suwannee River, Florida. *Transactions of the American Fisheries Society* 139(4):1233–1246.
- Roelke, D. L., L. a. Cifuentes, and P. M. Eldridge. 1997. Nutrient and phytoplankton dynamics in a sewage-impacted gulf coast estuary: a field test of the PEG-model and equilibrium resource competition theory. *Estuaries* 20(4):725.
- Rose, K. 2000. Why are quantitative relationships between environmental quality and fish populations so elusive ? *Ecological Applications* 10(2):367–385.
- Saucier, M. H., and D. M. Baltz. 1993. Spawning site selection by spotted seatrout, *Cynoscion nebulosus* , and black drum, *Pogonias cromis* , in Louisiana. *Environmental Biology of Fishes* 36(3):257–272.
- Secor, D.H.; Gunderson, T. E. 1998. Effects of hypoxia and temperature on survival, growth and respiration of juvenile Atlantic sturgeon, *Acipenser oxyrinchus* . *Fishery Bulletin*:603–613.
- Siefert, R. ., and W. A. Spoor. 1974. Effects of reduced oxygen on embryos and larvae of the white sucker, coho salmon, brook trout and walleye. Pages 487–495 in J. H. S. Blaxter, editor. *The early life history of fishes*. Springer-Verlag, New York.
- Sigua, G. C., and W. A. Tweedale. 2003. Watershed scale assessment of nitrogen and phosphorus loadings in the Indian River Lagoon basin, Florida. *Journal of Environmental Management* 67(4):363–372.
- Turner, R. E., N. N. Rabalais, B. Fry, N. Atilla, C. S. Milan, J. M. Lee, C. Normandeau, T. A. Oswald, E. M. Swenson, and D. A. Tomasko. 2006. Paleo-indicators and water quality change in the Charlotte Harbor estuary (Florida). *Limnology and Oceanography* 51(1):518–533.
- Van Doren, B. M., and K. G. Horton. 2018. A continental system for forecasting bird migration. *bioRxiv*.
- Vert-pre, K. A., R. O. Amoroso, O. P. Jensen, and R. Hilborn. 2013. Frequency and intensity of productivity regime shifts in marine fish stocks. *Proceedings of the National Academy of Sciences* 110(5):1779–1784.
- Walters, S., S. Lowerre-Barbieri, J. Bickford, and D. Mann. 2009. Using a passive acoustic survey to identify Spotted Seatrout spawning sites and associated habitat in Tampa Bay, Florida. *Transactions of the American Fisheries Society* 138:88–98.
- Walther, B. A., and J. L. Moore. 2005. The concepts of bias, precision and accuracy, and their use in testing the performance of species richness estimators, with a literature review of estimator performance. *Ecography* 28(6):815–829.
- Whaley, S. D., M. C. Christman, and J. J. Burd. 2016. Spatial distribution–abundance relationships in juvenile (age 0) Red Drum (*Sciaenops ocellatus*) and Spotted Seatrout (*Cynoscion nebulosus*). I: Influence of freshwater inflow. *Estuaries and Coasts* 39(3):742–

751.
 Whitfield, A. K., and T. D. Harrison. 2003. River flow and fish abundance in a South African
 estuary. *Journal of Fish Biology* 62(6):1467–1472.
 Wootton, R. J. 1998. *Ecology of teleost fishes*, second edition. Kluwer Academic Publishers,
 London.
 Wuenschel, M. J., R. G. Werner, and D. E. Hoss. 2004. Effect of body size, temperature, and
 salinity on the routine metabolism of larval and juvenile Spotted Seatrout. *Journal of Fish*
Biology 64(4):1088–1102.
 Zuur, A. F., E. N. Ieno, N. J. Walker, A. a. Saveliev, and G. M. Smith. 2009. *Mixed Effects*
Models and Extensions in Ecology with R. Springer Science+Business Media, New York,
 USA,:574.

Tables

TABLE 1. Environmental predictor variables assessed by the XGBoost algorithm and GLM.

¹Available for Tampa Bay, Charlotte Harbor, and Indian River Lagoon only.

²Available for Tampa Bay only.

³Apalachicola area has only one major river therefore river flow CLOSEST is equivalent to river flow ALL

Variable Name	Description	Life Stage Affected
Salinity (ppt)	Salinity at sampling	Recruit survival
Temperature (°C)	Water temperature at sampling	Recruit survival
Dissolved O ₂ (mg/L)	Dissolved oxygen at sampling	Recruit survival
Depth (m)	Water depth at sampling location	Recruit survival
Attenuation coefficient	Calculated attenuation coefficient if Secchi disk does not hit bottom at sampling	Recruit survival
DIN (dissolved inorganic nitrogen) ¹ (µg/L)	Closest dissolved inorganic nitrogen (µg/L) measurement within the defined spatial buffer surrounding the sampling location; one month prior to sampling month	Recruit survival
Chlor <i>a</i> (Chlorophyll) ² (µg/L)	Closest chlorophyll <i>a</i> measurement within the defined spatial buffer surrounding the sampling location; sampling month	Recruit survival
River flow CLOSEST (f ³ /s) ³	Average monthly flow rate of closest river (f ³ /second); sampling month	Recruit survival
River flow ALL (f ³ /s) ³	Average monthly flow rate of all rivers (f ³ /second); sampling month	Recruit survival
Precipitation (inches)	Total monthly precipitation (rainfall); sampling month	Recruit survival
Z anomaly	Palmer Z drought anomaly index ; sampling month	Recruit survival
Maximum T anomaly	Monthly maximum air temperature anomaly; sampling month	Recruit survival
Minimum T anomaly	Monthly minimum air temperature anomaly ; sampling month	Recruit survival
At-spawn salinity ¹ (ppt)	Mean salinity measurement in likely spawning area and at estimated spawning time	Spawning success
At-spawn water temperature ¹ (°C)	Mean water temperature in likely spawning area and at estimated spawning time	Spawning success

DIN 1prior ¹ (µg/L)	Mean dissolved inorganic nitrogen across likely spawning areas and one month prior to estimated spawning time	Spawning success
Previous autumn river flow (f ³ /s)	Mean monthly flow rate of closest river during previous autumn	Recruit survival, Spawning success
Winter river flow (f ³ /s)	Mean monthly flow rate of closest river during winter	Recruit survival, Spawning success
Previous autumn river flow rate ALL (f ³ /s)	Mean monthly flow rate of all rivers within area during previous autumn	Recruit survival, Spawning success
Winter river flow rate ALL (f ³ /s)	Mean monthly flow rate of all rivers within area during winter	Recruit survival, Spawning success
Previous autumn precipitation (inches)	Mean monthly precipitation during previous autumn	Recruit survival, Spawning success
Winter precipitation (inches)	Mean monthly precipitation during winter	Recruit survival, Spawning success
Previous autumn Z anomaly	Mean Palmer Z drought anomaly index during previous autumn	Recruit survival, Spawning success
Winter Z anomaly	Mean Palmer Z drought anomaly index during winter	Recruit survival, Spawning success
Previous autumn maximum T anomaly	Mean monthly maximum air temperature anomaly during previous autumn	Recruit survival, Spawning success
Winter maximum T anomaly	Mean monthly maximum air temperature anomaly during winter	Recruit survival, Spawning success
Previous autumn minimum T anomaly	Mean monthly minimum air temperature anomaly during previous autumn	Recruit survival, Spawning success
Winter minimum T anomaly	Mean monthly minimum air temperature anomaly during winter	Recruit survival, Spawning success
Mean DIN 2prior ¹ (µg/L)	Mean dissolved inorganic nitrogen across likely spawning areas within the two months prior to estimated spawning time	Spawning success
Mean DIN 3prior ¹ (µg/L)	Mean dissolved inorganic nitrogen across likely spawning areas within the three months prior to estimated spawning time	Spawning success

TABLE 2. River gauge numbers, weather station locations, and climate division assignments for each estuary: Apalachicola (AP), Cedar Key (CK), Tampa Bay (TB), Charlotte Harbor (CH), and northeast Florida (JX), and the Indian River Lagoon (IR).

Area	Rivers	Weather Station	Climate Division
AP	Apalachicola River (gauge #02359170)	Apalachicola Airport Station: USW00012832	1
CK	Suwannee River (station # 02323500) Waccasassa River (station #02313700)	Usher Tower, FL Station: USC00089120	2
TB	Alafia River (gauge #2301500) Hillsborough River (gauge #2304500) Little Manatee River (gauge #2300500)	Tampa International Airport, Station: USW00012842	4
CH	Caloosahatchee River (gauge #2292900) Myakka River (gauge #2298830) Peace River (gauge #2296750)	Fort Myers Page Field Airport, Station:USW00012835	5
JX	Cedar River (gauge #02246459) Ortega River (gauge #02246318) St. Johns River at Buffalo Bluff (gauge #02244040) St. Johns River (gauge #02246500) St. Marys River (gauge #02231254)	Jacksonville Naval Air Station, Station: USW00093837	2
IR	Crane Creek (gauge # 02249500) Eau Gallie River (gauge #02249007) North Prong St. Sebastian River (gauge #02251500) South Prong St. Sebastian River (gauge #02251000) Turkey Creek (gauge # 02250030)	Melbourne Weather Forecast Office, Station: USC00085612	3

TABLE 3. Total number of sampling events within sampling years, number of events where spotted seatrout were captured, mean capture rate, and environmental predictor variables evaluated using the XGBoost algorithm in each estuary: Apalachicola (AP), Cedar Key (CK), Tampa Bay (TB), and Charlotte Harbor (CH), northeast Florida (JX) and the Indian River Lagoon (IR). A check mark indicates whether the variable was used and a number to the right of the check mark is the number of missing observations within the estuary-specific datasets.

	AP	CK	TB	CH	JX	IR
Years	1998-2015	1996-2015	1989-2015	1989-2015	2001-2015	1990-2015
Sampling events	920	2034	7239	6038	1996	6449
# events where YOY <i>C. nebulosus</i> were captured	293	470	2176	1942	361	1835
Mean capture rate ($\pm$ SD)	2.05 $\pm$ 5.98	0.74 $\pm$ 2.85	1.89 $\pm$ 5.67	1.56 $\pm$ 4.26	0.68 $\pm$ 2.87	1.68 $\pm$ 5.17
Salinity	✓ 2	✓	✓ 17	✓	✓ 1	✓ 6
Temperature	✓	✓	✓ 4	✓	✓	✓ 6
Dissolved O ₂	✓ 19	✓ 66	✓ 140	✓ 51	✓ 12	✓ 70
Depth	✓	✓	✓ 14	✓	✓ 1	✓ 1
Attenuation Coef.	✓ 593	✓ 758	✓ 6125	✓ 5184	✓ 531	✓ 4849
DIN	-	-	✓ 3358	✓ 5450	-	✓ 5435
Chlorophyll <i>a</i>	-	-	✓ 4586	-	-	-
River flow rate	✓	✓ 8	✓	✓	✓ 159	✓ 254
River flow rate (ALL rivers)	-	✓	✓	✓	✓	✓
Precipitation	✓ 3	✓ 19	✓	✓	✓ 84	✓
Z anom	✓	✓	✓	✓	✓	✓
Maximum T anom.	✓	✓	✓	✓	✓	✓
Minimum T anom.	✓	✓	✓	✓	✓	✓
At spawn salinity	-	-	✓ 5127	✓ 4738	-	✓ 4645
At spawn water temperature	-	-	✓ 5127	✓ 4738	-	✓ 4675
DIN 1prior	-	-	✓ 5092	✓ 4815	-	✓ 6223
Previous autumn river flow rate	✓	✓	✓	✓	✓ 143	✓ 478
Winter river flow rate	✓	✓	✓	✓	✓ 143	✓ 299
Previous autumn river flow (ALL rivers)	✓	✓	✓	✓	✓	✓
Winter river flow (ALL rivers)	✓	✓	✓	✓	✓	✓
Previous autumn precipitation	✓	✓	✓ 82	✓ 1013	✓ 138	✓

Winter precipitation	✓	✓	✓	✓ 932	✓ 141	✓
Previous autumn Z anom.	✓	✓	✓ 82	✓ 81	✓	✓
Winter Z anom.	✓	✓	✓	✓	✓	✓
Previous autumn Max T anom.	✓	✓	✓ 82	✓ 81	✓	✓
Winter Max T anom.	✓	✓	✓	✓	✓	✓
Previous autumn Min T anom.	✓920	✓	✓ 82	✓ 81	✓	✓
Winter Min T anom.	✓920	✓	✓	✓	✓	✓
DIN 2prior	-	-	✓ 5092	✓ 4741	-	✓ 6100
DIN 3prior	-	-	✓ 5092	✓ 4675	-	✓ 6000

TABLE 4. Environmental predictor variables that resulted in the lowest mean average error between the observed and predicted counts for each estuary. Predictor variables are listed in order from those resulting in most to least gain as calculated by XGBoost for estuaries on the west coast of Florida; Apalachicola (AP), Cedar Key (CK), Tampa Bay (TB), and Charlotte Harbor (CH)). Note: predictor variables with less than 0.01 gain are not reported here.

AP		CK		TB		CH	
Important Variables	Gain	Important Variables	Gain	Important Variables	Gain	Important Variables	Gain
Salinity	0.17	Prev. aut. river flow ALL	0.14	Salinity	0.14	Salinity	0.19
Temperature	0.15	Dissolved O ₂	0.14	Dissolved O ₂	0.11	Dissolved O ₂	0.13
Dissolved O ₂	0.14	Salinity	0.11	Temperature	0.11	Temperature	0.12
Depth	0.07	Depth	0.11	DIN	0.08	Depth	0.07
Atten. coef.	0.06	Temperature	0.09	Depth	0.07	Precipitation	0.05
River flow	0.05	Year	0.08	Chlor <i>a</i>	0.06	River flow	0.05
Min. T anom.	0.05	Atten. coef.	0.07	River flow	0.05	River flow ALL	0.05
Z anom.	0.04	River flow	0.05	Precipitation	0.03	Month	0.04
Prev. aut. river flow	0.03	Precipitation	0.04	River flow ALL	0.03	Z anom.	0.04
Precipitation	0.03	Z anom.	0.03	Z anom.	0.03	Prev. aut. river flow	0.04
Winter Precip.	0.03	Max. T anom	0.03	Winter river flow	0.03	Min. T anom.	0.03
Max. T anom.	0.03	Min. T anom	0.02	Max. T anom.	0.03	Winter river flow	0.03
Winter river flow	0.02	River flow ALL	0.02	Winter river flow	0.03	DIN	0.02
Prev. aut. Precip.	0.02	Month	0.02	Atten. coef.	0.02	Max T. anom.	0.02
Year	0.01	Winter river flow	0.01	Month	0.02	Atten. coef.	0.02
Month	0.01	Winter max T anom.	0.01	Prev. aut. river flow	0.02	Winter max T. anom.	0.01
		Prev. aut. max T anom	0.01	Min. T anom.	0.01	Prev. aut. Z anom.	0.01
		Prev. aut. Z anom	0.01	Prev aut. Z anom.	0.01	Year	0.01
				Year	0.01		

				Winter river flow	0.01		
--	--	--	--	-------------------	------	--	--

TABLE 4 (cont.). Environmental predictor variables that resulted in the lowest mean average error between the observed and predicted counts for each estuary. Predictor variables are listed in order from those resulting in most to least gain as calculated by XGBoost for estuaries on the east coast of Florida: northeast Florida (JX) and Indian River Lagoon (IR). Note: predictor variables with less than 0.01 gain are not reported here.

JX		IR	
Important Variables	Gain	Important Variables	Gain
Salinity	0.16	Temperature	0.15
Temperature	0.14	Dissolved O ₂	0.15
Atten. coef.	0.12	Salinity	0.13
Dissolved O ₂	0.11	Depth	0.07
Depth	0.09	River flow	0.07
River flow	0.08	Precipitation	0.06
Precipitation	0.07	Atten. coef.	0.04
River flow ALL	0.07	Prev. aut. river flow	0.04
Winter river flow ALL	0.05	River flow ALL	0.03
Winter river flow	0.04	DIN	0.03
Prev. aut. river flow	0.04	Z anom.	0.03
Prev. aut river flow ALL	0.01	Winter river flow	0.03
		Month	0.02
		Min T. anom.	0.02
		Max. T. anom.	0.02
		Winter Z anomaly	0.01
		Winter precipitation	0.01
		Year	0.01

TABLE 5. Sample sizes of the training and testing datasets, mean absolute error (MAE), and mean error (ME) between observed and predicted counts produced by the XGBoost algorithm (XGB) and generalized linear model (GLM). Estuaries are Apalachicola (AP), Cedar Key (CK), Tampa Bay (TB), Charlotte Harbor (CH), Indian River Lagoon (IR), and northeast Florida (JX).

Estuary	Model	Sample size (N) train; test	MAE	ME
AP	XGB	644; 276	2.45	0.90
	GLM	554; 237	3.19	-0.94
CK	XGB	1424; 610	0.90	0.53
	GLM	580; 248	1.08	0.09
TB	XGB	5067; 2172	2.38	1.06
	GLM	1714; 735	3.23	0.04
CH	XGB	4227; 1811	1.84	0.60
	GLM	3023; 1296	1.99	-0.18
JX	XGB	1397; 599	0.79	0.48
	GLM	1090; 467	0.71	-0.02
IR	XGB	4514; 1935	2.07	0.8
	GLM	4127; 1769	2.37	-0.11

Figures

FIGURE 1. Map of statewide sampling for juvenile spotted seatrout in six sampling estuaries in Florida: Apalachicola (AP), Cedar Key (CK), Tampa Bay (TB), Charlotte Harbor (CH), northeast Florida (JX), and the Indian River Lagoon (IR). The red triangles indicate unique sampling stations across all years within each estuary.

FIGURE 2. Mean number ($\pm 1 SD$) of YOY spotted seatrout captured per sampling event, and a summary (mean $\pm 1 SD$) of the environmental predictor variables used in the analysis for every estuary: Apalachicola (AP), Cedar Key (CK), Tampa Bay (TB), Charlotte Harbor (CH), northeast Florida (JX), and the Indian River Lagoon (IR).

FIGURE 3. Importance of environmental predictor variables by estuary: Apalachicola (AP), Cedar Key (CK), Tampa Bay (TB), Charlotte Harbor (CH), northeast Florida (JX), and the Indian River Lagoon (IR).