Improving species distribution model predictive accuracy using species abundance: Application with boosted regression trees

Hao Yu⁎, Arthur R. Cooper, Dana M. Infante
Department of Fisheries and Wildlife, Michigan State University, East Lansing, MI, USA

ARTICLE INFO

Keywords:
Species distribution models
Boosted regression trees
Species abundance
Fluvial fishes

ABSTRACT

Auxiliary information in the form of species abundance is frequently available as part of data collected for ecological investigations, yet when modeling distributions of species over large regions, species presence (and sometimes absence) are typically used. Incorporating abundances into species distribution models may greatly improve model predictive accuracy in practice. Boosted regression trees (BRT) models have been widely used in species distribution modeling, however no ecological study has been conducted to date that has assessed the predictive accuracy of BRT models that incorporates species abundance weights. We compared traditional, unweighted BRTs with species abundance-weighted BRTs for 55 fluvial fish species native to the Northeastern U.S. Overall model deviance explained and six diagnostic measures of predictive performance were compared between traditional BRTs and weighted BRTs. These comparisons indicated that unweighted BRTs performed better for fluvial fish species considered common, including those with greater numbers of presences and higher prevalence. Conversely, weighted BRTs were better suited for modeling distributions of species that had fewer presences, lower prevalence, and higher rarity, indicating the potential of species abundance-weighted distribution modeling to improve results for species of high conservation importance. Last, we offer insights into the applicability of using weighted approaches with other commonly used species distribution modeling methods.

1. Introduction

Species distribution models (SDMs), also called ecological (environmental) niche models, play an important role in quantifying species-habitat relationships and predicting species distributions in ecological research, conservation, and environmental management (Guisan and Zimmermann 2000; Robinson et al. 2017). SDMs are used to predict the probability that a target species is present at a given location or to quantify habitat suitability as a function of multiple predictor variables representing key environmental conditions linked to species habitat usage and persistence. SDMs have been developed for many types of organisms residing in terrestrial, freshwater, and marine environments (Elith and Leathwick, 2009). In early stages of SDM use and development (prior to 2000), regression-based models (e.g., general generalized linear models) were frequently utilized (Guisan and Zimmermann, 2000). However, based on improved methodology and ecological understanding, more complex statistical approaches have been implemented for SDMs, increasing accuracy of model predictions. These advances have provided a mechanism for understanding complex, non-linear relationships and interactions among environmental predictors, providing gains in ecological understanding of species-environment relationships. In particular, the application of machine learning techniques in SDMs has increased dramatically over the past two decades, with boosted regression trees (BRT) models being one of the most widely used approaches (Elith et al., 2008). BRTs are adept at handling nonlinearity, selecting predictor variables, accounting for interactions among predictors, and quantifying predictor relative importance, all of which can be difficult to address in regression-based models. In numerous studies, BRTs have outperformed regression-based models, such as generalized linear models (GLMs) and generalized additive models (GAMs) in analyzing complex species-habitat relationships (Elith et al., 2008; but, see Shabani et al., 2016; Norberg et al., 2019). Despite the success of BRTs in developing SDMs for a wide variety of organisms and environments, potential still exists for improvements that could further bolster model accuracy for this modeling method.

Data availability is critical for SDM development, with characteristics of the data (type, method of collection, spatial extent, etc.) driving choices in SDM models and approaches. A number of studies have focused on data deficiency (Warton and Shepherd 2010; Elith et al. 2011; Fithian and Hastie 2013; Yackulic et al. 2013; Radosavljevic and Anderson 2014; Renner et al. 2015). For instance, when only presence
data are available, presence-only approaches can be used to model species distributions (Merow et al. 2013; Phillips et al. 2017). While species information may be collected as presence-only or presence-ab- sence data, auxiliary information in the form of species counts or abundance is frequently available in ecology. However, this information is often reduced to presence-only or presence-absence prior to SDM development, thus eliminating information that could further elucidate complex species-habitat relationships in SDMs. In certain cases, variabil- ity in species abundance could be linked to variation in habitat suitability, with species being more abundant within highly suitable habitats. For instance, Weber et al. (2017) found a positive relationship between abundance of species from many taxonomic groups and envi- ronmental suitability. Additional studies used occurrence data or SDMs to predict abundance distributions (Van Couwenbergh et al., Yañez-Arenas et al., 2014) or have used abundance data to improve the predictive abilities of SDMs (Howard et al., 2014). However, other studies found this type of relationship (between abundances and suit- ability) to be weak or non-existent (Dallas and Hastings, 2018). Col- lectively, these studies indicate that much still needs to be learned with respect to abundance-habitat relationships in SDMs, providing a research opportunity to explore the conditions under which abundance-informed SDMs could improve model predictions.

Modifications to existing SDM modeling methods have often been implemented in ecology for correcting imbalanced survey data. Commonly used SDMs can be divided into regression-based models and machine learning-based models. Logistic regressions are widely used in species distribution modeling when the response variable is dichotomous. However, if the response variable in a logistic regression has many more absences than presences, the accuracy and precision of parameters, as well as predictive performance, will be affected (Salas-Eljatib et al., 2018). For instance, King and Zeng (2001) introduced a corrective approach to deal with this imbalanced data issue known as “rare events logistic regression,” while Warton and Shepherd (2010) used Poisson point process logistic regression models to solve the “pseudo-absence problem.” Further, Stolar and Nielsen (2015) im- proved model performance dealing with spatially biased sampling by adding a weighting term in the logistic regression. Machine learning- based SDMs (e.g., BRTs, Maxent, random forest, artificial neural networks, etc.) are more complex than regression-based models and often treated as a “black box” with respect to species distribution modeling. Therefore, modified versions are relatively rare compared to regression- based models in ecology. However, just like regression-based SDMs, machine learning-based SDMs can be improved once their “black box” properties are uncovered. Through the explanation of machine-learning approaches such as BRT (Elith et al., 2008), these models have become more tangible to scientists and subsequently more frequently applied in ecological studies. Modifications to improve model fit and predictive abilities of these widely used machine learning methods, such as BRTs, have great potential in improving ecological research.

The goal of this study is to compare the predictive abilities of tra- ditional, unweighted presence-absence species distribution models with those that are weighted by species abundance using a common, robust species distribution modeling method: Boosted Regression Trees. We develop weighted vs. unweighted species distribution models for 55 fluvial fish species native to the Northeastern U.S. using a standard 10- fold cross-validation modeling approach. Weighted vs. unweighted models are compared using model deviance explained to measure overall model fit, six diagnostic measures of predictive performance, evaluation of patterns in species presence, prevalence, and rarity, and predictor variable importance and rankings. We provide recom- mendations on the use of abundance-weighted vs. unweighted models and offer insights into the applicability of using abundance-weighted approaches with other commonly used species distribution modeling methods.

2. Material and methods

Four primary steps were conducted to implement this study: bio- logical and environmental predictor data preparation, development of abundance weightings for each species, species distribution modeling using weighted and unweighted approaches, and comparison of results among weighted and unweighted models for each species (Figure 2).

2.1. Biological and environmental data

We developed species distribution models for 55 fluvial fish species native to 22 Northeastern U.S. states (Figure 1; Table A1) to compare the model performance and predictive capabilities of the weighted and unweighted BRT approaches. Community-wide fish data collected using single-pass electrofishing methods spanning 1990–2013 were obtained from academic institutions and local, state, and federal agencies (see Daniel et al., 2015) and used in model development. Species presence-absence data locations were designated as falling within either native or non-native portions of their overall range based on species-level 8-digit USGS Hydrologic Unit Code (HUC) maps of native and introduced status provided by the USGS Nonindigenous Aquatic Species Program. This step ensured that only presence-absence locations considered na- tive were used in modeled development and excluded non-native pres-ences for species outside of their native range which could represent novel conditions for model predictors and thus affect native species distribution model development. Fish survey site locations were spa- tially linked to stream reaches of the National Hydrography Dataset Plus V1 (NHDPusV1, USGS 2005), allowing for the use of an existing suite of 23 natural and anthropogenic landscape variables as predictors in modeling (Table A2). These include commonly utilized predictors

![Fig. 1. Flow chart of the four primary steps used to develop and compare BRT vs. WBRT models for 55 fish species.](image-url)
known to influence the distribution and abundance of fluvial fishes (Cooper et al., 2019), with natural factors representing catchment area, climate, elevation, and groundwater contribution to stream flow, and anthropogenic factors representing urban and agricultural land uses, dams, roads, nutrient inputs, and mines characterized over multiple spatial extents including catchments and riparian buffers (Table A2; Daniel et al., 2015; Cooper et al., 2017).

2.2. Description of the models

2.2.1. Unweighted Boosted Regression Trees (BRT)

Boosted Regression Trees (BRT) combine regression trees and a powerful boosting technique that iteratively fits tree models using binary splits of predictor variables (Elith et al., 2008). Boosting is an ensemble procedure for improving model prediction by reducing model deviance through linking successive tree models focused on weak learners, i.e., the residuals from predictors performing poorly in previous steps. In BRT models, three parameters must be considered: learning rate, tree complexity, and bag fraction. Learning rate is used to control the contribution of each individual tree to the overall model. Tree complexity adjusts the number of nodes in a tree, governing the interaction complexity in the model (e.g., if tree complexity is 2, up to two-way interactions can be fit). Finally, bag fraction is the proportion of training data that is used in each iteration, which controls the stochasticity of boosting. Imbalanced presence/absence data (very rare or very common species) will often require differing learning rates (Elith et al., 2008). During preliminary model runs we evaluated differing combinations of learning rate values, number of trees and bag fraction values. This process identified that larger learning rates might result in models for rare species that did not converge, while smaller learning rates for common species might result in model overfitting. As a result, we used an initial learning rate of 0.05 for species with > 100 presences and a learning rate 0.01 for species with ≤ 100 presences. We iteratively reduced the learning rate by half to ensure a minimum of 1000 trees at 10,000 to avoid overfitting. For all models we used the default bag fraction of 0.75 and a tree complexity of 5 (i.e., five nodes in each tree), with models being developed with the 'dismo' R package.

2.2.2. Weighted Boosted Regression Trees (WBRT)

A weighted Boosted Regression Tree (WBRT) model is a modified BRT that applies a weight \( w_{ij} \) to each species at each sampling site based on individual species abundance and overall species richness, differing from a standard BRT where all species presence-absences are effectively weighted equally. Weighting sites based on numbers of species supported could account for potential differences in habitat suitability across sites, as sites with higher individual species abundance may reflect greater overall habitat suitability (Weber et al., 2017) and result in improved model fit.

In a logistic WBRT, the loss function (residual deviance) of species \( j \) is:

\[
L_j = -2 \frac{1}{\sum w_{ij} \sum w_j} \sum w_{ij} \left[ (1 - y_{ij}) \log(1 - \hat{y}_{ij}) + y_{ij} \log(\hat{y}_{ij}) \right]
\]

where \( n_j \) is the number of sites in the model training data set of species \( j \), \( y_{ij} \) is the observed value (0 or 1) of species \( j \) at site \( i \), \( \hat{y}_{ij} \) is the predicted value of species \( j \) at site \( i \), and \( w_{ij} \) is the weight of species \( j \) at site \( i \). For the WBRT, we set the weight \( w_{ij} \) equal to 1 for sites with species absences, while \( w_{ij} \) is a scaled product (ranging from 1 to 101) of relative abundance and species richness \( j \) at site \( i \) with presence of target species \( j \):

\[
w_{ij} = \frac{\text{unscaled weight of species } j \text{ at site } i}{\text{product of the relative abundance of species } j \text{ and richness at site } i, \text{ calculated using the following formulas:}}
\]

\[
\text{RA}_{ij} = \frac{A_{ij}}{\sum_{j=1}^{m} A_{ij}}
\]

\[
\text{unscaled weight of species } j \text{ at site } i = \text{RA}_{ij} \times \text{Richness}
\]

where \( \text{RA}_{ij} \) is the relative abundance of species \( j \) at site \( i \), \( A_{ij} \) is the raw abundance of species \( j \) at site \( i \), and \( m \) is the species richness at site \( i \). Here we use relative abundance as it is readily available form of abundance for fish data that can be calculated simply from community species count data as opposed to effort-based abundance measures (commonly referred to as catch per unit effort or CPUE). This is due to the fact that effort measures (e.g., length, time, area, etc.) are sometimes lacking for fish community data provided by various sources collected under differing sampling objectives.

In developing this weighting factor, we account for the dual influences of both relative abundance and species richness. For instance, using the formula above, a species with a relative abundance of 0.2 at a
site with an overall species richness of 10 would have a weight of 2, however the same species would have a lower weight of 0.4 given a site species richness of 2 and the same relative abundance (0.2). In effect, this weighting controls for both site-level species abundance and species dominance, resulting in higher weighting for sites with higher species abundance relative to higher overall species richness and providing a potential indicator of higher overall site-specific habitat suitability.

BRT and WBRT models were developed using a 10-fold cross validation with BRT and WBRT being fitted using the same training set with the optimal number of trees in each model being estimated using 10-fold cross validation. During this validation process, the dataset was divided into 10 non-overlapping groups, with each unique group being withheld as a test dataset while remaining groups were used as a training dataset for model fitting.

2.3. Model comparison

We compared results between the WBRT and BRT approaches among the 55 fishes modeled using: 1) overall model deviance explained as a measure of model fit, 2) six diagnostic metrics evaluating model predictive performance, 3) comparison of patterns in species presences, prevalence, and rarity between models, and 4) relative importance of predictors and overall predictor rankings.

2.3.1. Model deviance explained

We used deviance explained based on models developed from the overall initial dataset to compare model fitting, where cross validation residual deviance is the mean of the residual deviance from each fold of the cross validation:

\[ \text{Deviance explained} = \frac{\text{total deviance} - \text{cross validation residual deviance}}{\text{total deviance}} \]

A t-test was used to compare mean deviance explained between the WBRT and BRT approaches.

2.3.2. Metrics comparing predictive performance

Model comparison is often a crucial aspect of evaluating potential improvements to an existing method. The measures and methods to compare the accuracy or performance of SDMs are diverse and controversial (Liu et al., 2011; Leroy et al., 2018). An intuitive measure is overall accuracy, defined as the proportion of sites predicted to support a species where a species is actually found. However, it has been repeatedly criticized as not being suitable for imbalanced data, which includes data with many more absences than presences or vice versa (Fielding and Bell, 1997; Manel et al., 2001). Two alternative measures are sensitivity (proportion of presences correctly predicted) and specificity (proportion of absences correctly predicted; Swets, 1988). Both are calculated from a confusion matrix and are independent of prevalence (the proportion of presences in the dataset) (Allouche et al., 2006).

One of the most commonly used measures to compare SDM performance is the area under the receiver operating characteristic (ROC) curve, known as AUC, which is developed from a 2-dimensional plot with sensitivity as the vertical axis and 1-specificity as the horizontal axis. AUC is a threshold-independent method, avoiding the subjective selection of threshold values where a single presence/absence cutoff is chosen to develop a confusion matrix for model evaluation. AUC ranges between 0 and 1, with an AUC of 0.5 indicating that the prediction capability of the model is no better than random and values greater than 0.7 are considered adequate in modeling species distributions (Swets, 1988). However, AUC has been criticized in several studies for giving misleading results for imbalanced data (Lobo et al., 2008; Peterson et al., 2008; Jimenez-Valverde, 2012). Frequently, ecological sample data are imbalanced, especially over large regions, and therefore AUC may not be appropriate to evaluate and/or compare SDMs alone. An alternative evaluation metric is the area under the precision-recall curve (also called sensitivity) curve (AUPRC), which is also a threshold-independent metric. This metric can evaluate SDMs with imbalanced data as it is not dependent on model specificity (Davis and Goadrich, 2006; Sofaer et al., 2018). Similarly, AUPRC measures the area under a 2-dimensional curve in which the vertical axis is precision and the horizontal axis is sensitivity (also called recall) (Sofaer et al., 2019). AUPRC can range between 0 and 1. However, its minimum value increases with prevalence and there is no established cut-off point for identifying adequate models with AUPRC, though higher AUPRC indicates a better model prediction.

Cohen’s kappa is another commonly used metric to evaluate SDM performance, however it depends on prevalence and therefore may result in statistical inaccuracies in estimating SDM accuracy (Allouche, 2006; Delgadillo and Tibau, 2019). Cohen’s kappa is calculated using three parameters: prevalence, sensitivity, and specificity:

\[ \text{Kappa} = \frac{P_0 - P_2}{1 - P_2} \]

where \( P_0 = \text{sensitivity} + (1 - \text{prevalence}) \times \text{specificity} \), \( P_2 = -2(\text{sensitivity} + \text{specificity} - 1) \times (1 - \text{prevalence}) + P_0 \)

Kappa ranges between −1 and 1 (Cohen, 1960) with higher kappa values indicating better model predictions. A more appropriate alternative is the true skill statistic (TSS), which is equal to the sum of sensitivity and specificity minus one (Fielding and Bell, 1997). TSS retains all the advantages of kappa but is also largely immune to prevalence of the sample data (Allouche, 2006). In this study, predicted presences and absences were separated by a threshold value at which the TSS is maximized (Manel et al., 2001; Hernandez et al., 2006).

Use of a wide variety of diagnostic measures of model accuracy, such as those described, can provide a means to effectively compare models developed using alternative approaches including potential modification of an existing modeling method. Results for BRT and WBRT models were compared using sensitivity, specificity, AUC, AUPRC, Cohen’s kappa, and TSS to evaluate the models’ predictive capability (Table 1). Values for these six diagnostic metrics were compared for each species for the WBRT and BRT modeling approaches to determine the number of metrics that performed better using WBRT vs. BRT. For subsequent analyses, we identified species that performed better using WBRT if ≥ 4 metrics had higher values compared to BRT.

<table>
<thead>
<tr>
<th>Metric</th>
<th>Description</th>
<th>Range</th>
<th>Threshold</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sensitivity</td>
<td>Proportion of presences correctly predicted (i.e., recall)</td>
<td>0 – 1</td>
<td>Larger value indicates better model</td>
<td>Swets (1988)</td>
</tr>
<tr>
<td>Specificity</td>
<td>Proportion of absences correctly predicted</td>
<td>0 – 1</td>
<td>Larger value indicates better model</td>
<td>Swets (1988)</td>
</tr>
<tr>
<td>AUC</td>
<td>Area under the curve of the receiver operating characteristic (ROC)</td>
<td>0 – 1</td>
<td>&gt; 0.7</td>
<td>Lusted (1971)</td>
</tr>
<tr>
<td>AUPRC</td>
<td>Area under the precision-recall curve</td>
<td>0 – 1</td>
<td>Larger value indicates better model</td>
<td>Raghavan et al. (1989)</td>
</tr>
<tr>
<td>Cohen’s kappa</td>
<td>Measurement of interrater reliability</td>
<td>−1 – 1</td>
<td>&gt; 0.2</td>
<td>Cohen (1960)</td>
</tr>
<tr>
<td>TSS</td>
<td>True skill statistic: sensitivity + specificity - 1</td>
<td>−1 – 1</td>
<td>Larger value indicates better model</td>
<td>Fielding and Bell (1997)</td>
</tr>
</tbody>
</table>
Conversely, species identified as performing better using BRT had \( \geq 4 \) metrics with higher values compared to WBRT. Results for these two groups were evaluated using radar plots to identify which metrics corresponded to differing performance among the two modeling approaches.

2.3.3. Differences in species presence, prevalence, and rarity

To test whether differences in model performance were related to species presence, prevalence, and multiple aspects of rarity, empirical distribution functions were applied. We used this approach to compare these factors for species performing better using WBRT with distributions from species better suited to BRT modeling (described above). For this analysis, we performed a two sample Kolmogorov-Smirnov (KS) test and plotted results using empirical cumulative density functions. The KS test is a non-parametric method for comparing two samples to determine whether they follow the same distribution (Rohlf and Sokal, 1981). The KS test statistic is:

\[
D = \sup_{x} |F_{n}(x) - F_{o}(x)|
\]

where \( F_{n} \) and \( F_{o} \) are the empirical cumulative distribution function of the first and second sample, respectively. When sample sizes are large (\( n_{1} > 50 \) and \( n_{2} > 50 \)), the critical value is:

\[
D_{n} = K_{\alpha} \sqrt{\frac{n_{1} n_{2}}{n_{1} + n_{2}}} \quad \text{where} \quad K_{\alpha} = \frac{\sqrt{\pi}}{2\alpha}, \quad \alpha = \text{level of significance.}
\]

When the KS test statistic \( D \) is greater than the critical value \( D_{n} \), these two samples’ distributions are significantly different.

To quantify species rarity, we utilized an Integrated Rarity (IR) index (Leitao et al., 2016) representing a continuous gradient of species rarity by combining measures of species range size, habitat usage, and mean species weights (described in ‘WBRT’ section above). Species native range areas (km\(^{2}\)) obtained from range maps (described above) were utilized as a measure of overall range size, with species having smaller ranges being geographically rarer. Range in habitat usage was derived with the Index of Habitat Specificity (IHS; Pritt and Frimpong, 2010) which sums the number of unique freshwater habitat types (e.g., substrate, flow velocity, etc.) attributed to individual freshwater fish species (Frimpong and Angermeier, 2009). The IHS has a theoretical maximum value of 25 (i.e., 25 total habitat types assigned) with species that have lower IHS scores indicating a lower range of habitat use and potentially greater rarity (Pritt and Frimpong, 2010). Mean species weights utilized as the weighting factor in WBRT were used as a measure of species abundance, with species that have lower mean weights representing those that have lower overall relative abundance normalized by species richness at occupied sites. The Integrated Rarity (IR\(_{i}\)) of species \( i \) was calculated by combining species native range size (SNR\(_{i}\)), Index of Habitat Specificity (IHS\(_{i}\)), and mean species weights (MSWI\(_{i}\)) in order to integrate habitat diversity appropriate to individual freshwater fish species:

\[
IR_{i} = \frac{(SNR_{i} \times w_{SNR_{i}} + IHS_{i} \times w_{IHS_{i}} + MSW_{i} \times w_{MSW_{i}})}{w_{SNR_{i}} + w_{IHS_{i}} + w_{MSW_{i}}}
\]

where \( w_{SNR_{i}}, w_{IHS_{i}}, \) and \( w_{MSW_{i}} \) are the weights of SNR\(_{i}\), IHS\(_{i}\), and MSWI\(_{i}\), respectively. In this study, all three weights in this formula were equal to 1/3, reflecting equal weighting of these three respective factors. SNR\(_{i}\), IHS\(_{i}\), and MSWI\(_{i}\) were standardized to a 0 to 1 scale prior to calculation of \( IR_{i} \) values using the following formula:

\[
\text{Standardized Index}_{i} = \frac{\text{Index}_{i} - \min(\text{Index})}{\max(\text{Index}) - \min(\text{Index})}
\]

where \( \min() \) and \( \max() \) represent the minimum and maximum values, respectively. Resulting IR values range from a theoretical minimum of 0 to a theoretical maximum of 1, with lower values for rare species and higher values for more common species. In general, species considered rare according to the IR index would have smaller ranges, utilize fewer habitat types, and have lower abundance relative to other species.

2.3.4. Predictor relative importance and overall predictor rankings

The relative importance of predictor variables was calculated for each species using WBRT and BRT model results in order to compare the relative contributions of predictor variables among approaches. Predictor variable importance is calculated as:

\[
RI_{i} = \frac{1}{M} \sum_{m=1}^{M} I^{m}(T_{m})
\]

where \( RI_{i} \) stands for the relative importance for the \( i \)th predictor variable, \( M \) is the number of trees, and \( I^{m}(T_{m}) \) is the squared relevance of each predictor weighted by the number of times it was chosen as the splitting variable in tree \( m \) (Hastie et al., 2009). In addition, we assigned ranks to predictor variables based on their relative importance, calculating an overall mean rank across species for each predictor:

\[
\text{mean rank}_{i} = \frac{\sum_{j=1}^{N} \text{rank}_{ij}}{N}
\]

where \( \text{rank}_{ij} \) is the rank of predictor \( i \) for species \( j \), and \( N \) is the total number of species (55 in this study). All analyses were conducted in R (version 3.6.1, R Core Team, Vienna, Austria).

3. Results

3.1. Comparing BRT and WBRT model results

3.1.1. Model deviance explained

Percentage of total deviance explained, used to measure model fitting, was higher for WBRT than BRT for 49 of 55 species (Table A3). Mean deviance explained for WBRT was 0.4769 (SE 0.0192, Range 0.1924–0.8132) compared to 0.3743 (SE = 0.0137, Range 0.1429–0.6131) for BRT, and was significantly higher based on a paired \( t \)-test (\( p < 0.01 \)) (Fig. 3A). Differences in deviance explained between WBRT and BRT models varied over the Integrated Rarity index. In general, these differences were positive, indicating that WBRTs had improved model fit compared to BRTs, however when Integrated Rarity was larger than 0.4 (indicating more common species) differences among the WBRT and BRT approaches decreased (Fig. 3B).

3.1.2. Metrics comparing predictive performance

For 50 of 55 species, WBRT had a higher metric value (indicating better predictive performance) than BRT for at least one metric. When multiple metrics are considered, 13 species have higher values for WBRT for four or more metrics (Fig. 4A; Table A3), while 24 species performed better with BRT based on having higher values for four or more metrics. Generally, kappa and TSS led to the same model preferences and in particular, results for sensitivity and TSS matched (both performed better for either WBRT or BRT) for 52 of 55 species. In contrast, sensitivity and specificity had differing model preferences being mismatched in 50 of 55 species. AUPRC and AUC had moderate congruence with only 15 out 55 species mismatched. Among the 13 species for which WBRT performed better, results were mainly driven by higher metric values for sensitivity (13), Kappa (13), TSS (13), and AUPRC (12) (Fig. 4B). Only two species showed a higher specificity value, while eight species had higher AUC values. Similar to WBRT, results for the 24 species for which BRT performed better included higher metric values for sensitivity (21), Kappa (24), TSS (24), and AUC (23) (Fig. 4C) while eight species and 16 species had higher specificity and AUPRC values, respectively.

3.2. Differences in species presences, prevalence, and rarity

Among the two groups of species performing better in either WBRT (\( n = 13 \)) or BRT (\( n = 24 \)) (Table 2), the number of presences for species with better predictions in WBRT was significantly lower than those with better predictions in BRT (Welch \( t \)-test, \( p < 0.01 \)), with a mean of 785
and richness in these datasets in SDMs as opposed to reducing com-
sors, particularly in fisheries re-
nity-based abundance sampling e
he species data utilized in SDM development originate from commu-
Stevens and Conway, 2020
Stevens and Conway, 2020
Stevens and Conway, 2020

4. Discussion

The importance of predictor variables varied primarily by species, with differences among WBRT and BRT models among species being minimal as indicated by similar mean relative importance and ranks of predictors (Table 3). The top predictor was catchment area, accounting for 23.8% and 18.1% of relative importance for WBRT and BRT, respectively, with mean annual air temperature (9.7 vs. 9.3%) and mean annual precipitation (7.7% vs. 6.9%) being the next two most important variables. Overall, the order of predictors according to mean relative importance and mean rank was identical in WBRT and BRT (Table 3).

4.1. Applying species abundance and richness-based weights to SDMs: implications for improved modeling of less prevalent and rare species

Our results suggest that neither BRTs nor weighted BRTs (WBRTs) as a whole were a better choice for all stream fish species modeled in this study for the Northeastern U.S. However, in general, WBRTs outperformed unweighted BRTs for stream fish species with fewer pre-
sences, lower prevalence, and higher rarity. These are characteristics that can be shared by species of conservation importance; thus, this weighting approach has the potential to improve models for these types of species. Although species with low prevalence can result in im-
balanced data, causing biased predictions in SDMs as has been in-
dicated in a number of studies (Manel et al., 2001; McPherson and
Jetz, 2007; Santika, 2011), this issue has seldom been discussed in the context of BRT models. This is perhaps due to the more powerful pre-
dictive ability of BRTs, as issues related to low species prevalence and imbalanced data are not as apparent as regression-based SDMs. Iden-
tifying candidate species for use of weighting BRTs can be challenging given that the discrimination and definition of rare species have not been widely accepted by researchers (Pritt and Frimpong, 2010). The rarity index utilized in this study that combines species' abundance weights, habitat usage, and native range size provides a reasonable measure of rarity that can be applied to other taxa to identify species best suited for weighted BRTs. When prevalence data are not available, the number of presences could be applied to substitute for prevalence as they are likely to be highly correlated in many cases. For instance, the number of presences and prevalence were highly correlated (Pearson's correlation coefficient = 0.85, p < 0.01) in the current study. Based on the results of this study, WBRTs should be considered for fluvial fish species with high rarity and low number of presences.

Both species relative abundance and richness were used to calculate the weight of each observation in this study. Unweighted binary (pre-
sence-absence) species models treat all presences the same when testing habitat suitability, whether a single individual was observed or many
(potentially 1000’s of) individuals were observed. A more logical assumption is that species will be more abundant in more suitable habitats (Aguirre-Gutiérrez et al., 2013; VanDerWal et al., 2009). There are, however, limits to the information that these weightings can provide in the context of SDMs, as WBRTs may not be ideal for common species as they are usually widely distributed and have generalist adaptions to habitats (Pritt and Frimpong, 2010). Given that common species tend to be widespread and generally of high abundance at locations where they are found, it is unsurprising that weights for these species add little value to BRTs. While a limited number of studies have concluded that habitat suitability and species abundance are unrelated (e.g., Dallas and Hasting, 2018; Filz et al., 2013; Nielsen et al., 2005), these studies have focused on different taxa groups (e.g., trees, mammals, insects, and vascular plants), have not incorporated the effects of species richness, and/or had much coarser data spatial resolutions than the current study. Although we incorporated species richness in WBRT models, the effect of species interactions on SDMs needs to be further studied. While the current study supports a positive linkage between species abundance and habitat suitability for less prevalent and rare species, this relationship requires further research to aid in use of species weightings for other taxa groups beyond fluvial fishes.

In addition to prevalence and rarity, there are other factors that may affect model performance. For example, fish sampling provides a snapshot of the relationship between fish communities and their habitats, with fish presence and/or abundance information being influenced by sampling effort, season, date, time, location, and identification (and/or counting) errors to differing degrees. Those biases and uncertainties can vary by species. For instance, rarer species can be more difficult to

Fig. 4. Comparison of six predictive performance metrics (A), sensitivity, specificity, kappa, TSS, AUPRC, and AUC, between WBRT and BRT with blue bars indicating better performance by WBRT while white bars indicate better performance by BRT. Radar plots of the six performances metrics for 13 species performing better in WBRT (B) and 24 species performing better for BRT (C), indicating the number of species that performed better based on that metric. See Table A1 for species names corresponding to reference numbers in (A).
detect or may have more variability in abundance across sampling locations (Wenger and Freeman, 2008; Steenweg et al., 2019), which could result in less model certainty for these species in certain cases. In addition, fish species have differing physiological tolerances to habitat conditions or may undergo seasonal migrations. In addition, fish species have differing physiological tolerances to habitat conditions. As a result, locations with the highest species’ abundance may not always correspond to the most suitable habitats. Nevertheless, weighted SDMs still have a great potential in improving predictive ability of SDMs.

4.2. Use of multiple diagnostic metrics to evaluate alternative models

When applying the BRT or WBRT approaches for other species, model predictive performance should be evaluated on a species by species basis. We implemented six metrics to compare the performance of the BRTs and WBRTs and found differing results in model preference among metrics. If only one metric was used to measure the performance of SDMs, biased conclusions in model performance would likely result. For this reason, model performance should not be measured using one single diagnostic metric. Further, survey data for many species are imbalanced (i.e., prevalence is much lower than 0.5), yet some commonly used diagnostic metrics are designed for balanced data. AUC, in particular, weights sensitivity and specificity equally when evaluating model performance. In ecological surveys, presences are usually more valuable in exploring habitat similarity than absences, which can be incorrectly obtained for multiple reasons. For instance, all fish species may not have been captured due to sampling or gear problems. Additionally, habitat may be suitable, however individuals from a particular fish species may have vacated the area at the time when sampling was conducted due to seasonal or short-term variation in habitat use based on environmental conditions (e.g., stream flow, temperature, etc.). Similarly, correctly predicting presences for an SDM is often of greater importance than correctly predicting absences in most cases. Therefore, sensitivity and AUPRC are better metrics than specificity and AUC for imbalanced data in ecology. Allouche et al. (2006) found that TSS is a better metric of predictive accuracy than kappa for SDM evaluation as kappa is dependent on prevalence. However, in the present study, kappa and TSS exhibited accuracy than kappa for SDM evaluation as kappa is dependent on most cases. Therefore, sensitivity and AUPRC are better metrics than specificity and AUC for imbalanced data in ecology. Allouche et al. (2006) found that TSS is a better metric of predictive accuracy than kappa for SDM evaluation as kappa is dependent on prevalence. However, in the present study, kappa and TSS exhibited high correlation (Pearson’s correlation coefficient = 0.98, p < 0.01). One possible reason is that the prevalence for 53 of 55 fluvial fish species is this study is less than 0.5, and it is not possible to test the performance of kappa or TSS across the whole prevalence range (0–1). When evaluating the performance of SDMs developed using the BRT and WBRT approaches, we recommend the use of multiple diagnostic metrics that best represent a given study’s objectives in use of SDM results.

4.3. Applicability of species weightings with other SDM approaches and additional weighting factors

This study offers insights into the applicability of using abundance-weighted approaches with other commonly used species distribution modeling methods, such as logistic regression (LR) and random forest (RF). Species abundance information can be added into logistic regression models by similarly adjusting the weights in the likelihood function. However, compared to BRT models, LR models have...
difficulties in dealing with multicollinearity, interactions, non-linearity, and predictor variable selection. These disadvantages may affect the application of weighted LR models and therefore model comparison and validation is needed prior to implementation. RF models use the Gini index and entropy to grow trees. An abundance index can be added into the Gini index and entropy function by adjusting case weights of sample data. However, the performance of weighted RF and its corresponding influence on SDM predictive performance needs further

Fig. 5. Boxplots and empirical cumulative density function (ECDF) curves comparing presences (A), prevalence (B), and rarity index (C) for the species performing better using the WBRT ($n = 13$) vs. BRT ($n = 24$) modeling approaches (see Table 2 for species lists).
In the present study, we found that WBRT models outperformed BRT models for fish species with lower prevalence and higher rarity in the Northeastern U.S., while BRT models performed better for common species with higher prevalence. Further, use of a single model evaluation metric should be avoided in model comparison in favor of multiple diagnostic metrics, as certain metrics may be less robust for evaluating model performance. Instead, multiple metrics should be used to provide a framework for developing weighted SDMs into useful tools for policy makers and managers.

### 4.4. Conclusion

We propose an approach to developing weighted SDMs using species abundance and richness presented in this study can be applied to other commonly used SDMs and is not limited solely to BRTs. Since SDMs are inherently species-specific and data dependent, model evaluation should be applied for any new species, taxa, or datasets utilizing this approach.

### CRediT authorship contribution statement

Hao Yu: Conceptualization, Methodology, Formal analysis, Writing - original draft. Arthur R. Cooper: Conceptualization, Data curation, Investigation, Writing - review & editing. Dana M. Infante: Supervision, Conceptualization, Writing - review & editing.

### Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

### Acknowledgments

This study was funded by the USGS Aquatic GAP program. We thank Alexa McKerrow and Daniel Wieferich for providing feedback on this project. We also thank Kyle Herreman and Jared Ross for GIS and database assistance. We are grateful to the numerous federal and state agencies, academic institutions, and non-profit organizations that provided the fish data facilitating this study (http://assessment.fishhabitat.org/#578a9a43e4b0c1aacab89763/578a9a47e4b0c1aacab8976a).

### Supplementary materials

Supplementary material associated with this article can be found in the online version, at doi:10.1016/j.ecolmodel.2020.109202.

### References


