



Modeling fish habitat: model tuning, fit metrics, and applications

Jacob W. Brownscombe¹ · Jonathan D. Midwood¹ · Steven J. Cooke²

Received: 18 November 2020 / Accepted: 7 March 2021
© Crown 2021

Abstract

Knowledge of the habitat associations and spatial–temporal distributions of wild animals is essential for successful ecosystem management, and effective analytical approaches are key to develop accurate models of these relationships. We explore the influence of several modeling techniques, tuning parameters, and assignment thresholds on a variety of model fit metrics to characterize habitat associations and make spatial–temporal predictions of species distribution based on a nine-year acoustic telemetry fish tracking dataset from a freshwater system. Unweighted generalized linear mixed models (GLMM) and random forests (RF) had the highest prediction accuracy of fish occupancy (> 84%) and precision (positive predictive value accuracy), but because the data were imbalanced (> 70% absences), predictions had low sensitivity (accuracy of true presences, < 45%), and therefore, low accuracy balance. Model weighting to prioritize presences and lowered presence probability thresholds both produced more balanced models, but RF exhibited low sensitivity to alterations in probability thresholds. Model weighting presents a straightforward approach to balance class accuracy in imbalanced datasets, which are common in species distribution samples. However, there is a wide range of weighting options and an important trade-off between model sensitivity and precision, either of which may be favoured depending on the research question or management application.

Keywords Environmental management · Conservation · Machine learning · Occupancy models · Habitat suitability

Introduction

Knowledge of species distributions, habitat associations, and spatial–temporal patterns of occupancy is essential for effective ecosystem management (Fielding and Bell 1997; Loiselle et al. 2003; de Kerckhove et al. 2008). Measurement of these characteristics is challenging due to environmental monitoring limitations, and complex modeling techniques are often required to develop accurate representations for inference of habitat associations and spatial–temporal predictions (Elith et al. 2006; Guillera-Arroita et al. 2015). Rapid advances in both data collection and modeling techniques are improving our capacity to develop accurate representations of biological systems; however, they also present

numerous experimental design and analytical challenges that can be difficult for researchers to navigate. Moreover, given that the spatial ecology of wild animals is of great relevance to environmental managers (Allen and Singh 2016), data collection and modeling strategies can also have implications for management decisions and interventions (e.g., restoration planning).

Historically, animal habitat has been measured by in situ sampling of animal distribution relative to environmental conditions. For example, fish distributions are often sampled through techniques such as netting, electrofishing, longlining, or visual observations (Bonar and Hubert 2002). Fish habitat use is particularly challenging to characterize due to constraints of detecting the presence of mobile and cryptic fish in underwater environments. However, new technologies are allowing for more continuous measurement of fish habitat use using video, telemetry, and passive acoustics (Cappo et al. 1999; Rountree et al. 2006; Hussey et al. 2015). Telemetry involves tracking tagged animals with mobile or stationary receivers, which has the added value of being able to monitor the movement and space use of individuals long term (i.e., multiple years), helping to address key knowledge requirements related to habitat connectivity and movement

✉ Jacob W. Brownscombe
jakebrownscombe@gmail.com

¹ Great Lakes Laboratory for Fisheries and Aquatic Sciences, Fisheries and Oceans Canada, Burlington, ON L7S 1A1, Canada

² Fish Ecology and Conservation Physiology Laboratory, Department of Biology, Carleton University, 1125 Colonel By Drive, Ottawa, ON K1S 5B6, Canada

pathways (Crossin et al. 2017; Brooks et al. 2018). Continuous measurement of the presence/absence of tagged fish at telemetry receiver locations can also provide more comprehensive measures of habitat use, suitability, and function than instantaneous measures of fish habitat use, which are typically conducted more seldom due to resource constraints (Brownscombe et al. 2019). However, applications of telemetry techniques to characterize animal habitat use and selection are generally better developed with other animal taxa and ecosystems (e.g., Marzluff et al. 2001; Long et al. 2009).

Using data from field sampling of animal distributions relative to both biotic and abiotic conditions, a wide variety of models of animal presence or presence/absence are commonly used to describe habitat associations and make spatiotemporal predictions, such as simple or multiple regression, generalized additive models, Bayesian inference, and machine learning (Elith et al. 2006). In selecting model type and tuning, model fit is the primary consideration, which should ideally be assessed in separate test data held out of the model training process [see Fielding and Bell (1997) for overview of common approaches]. For classification problems such as animal occupancy, there are a wide variety of model fit metrics available (e.g., see Fielding and Bell 1997; Kuhn et al. 2019). Focusing on overall model fit metrics such as model accuracy is highly common (e.g., Elith et al. 2006; MacLeod et al. 2008), but can be perilous in some circumstances because these metrics do

not consider accuracy amongst classes (Fielding and Bell 1997; Termansen et al. 2006; Jiménez-Valverde and Lobo 2007; Lobo et al. 2008). This is particularly concerning for species distribution and habitat models, since false positive and negative prediction rates can have a major impact on environmental decision making (Loiselle et al. 2003; Rondinini et al. 2006). To address this, models can be tuned in a variety of ways including selecting probability thresholds for presence assignment, down sampling the majority class, or model weighting to penalize misclassification of the minority response class (Fielding and Bell 1997; MacKenzie et al. 2002; Liu et al. 2005; Jiménez-Valverde and Lobo 2007; Evans et al. 2011).

Advances in our ability to remotely measure animal distributions and environmental associations, combined with complex analytical techniques, have valuable applications for environmental management through description of habitat associations and prediction of species distributions. Here we apply two commonly employed modeling techniques, generalized linear mixed effects models (GLMM) and random forests (RF), to a nine-year fish tracking data set, with which we explore a range of model tuning approaches (i.e., weighting, probability of occurrence thresholds) in relation to a diversity of model fit metrics (Table 1). Due to the flexible nature of RF for habitat modelling, we further apply the RF models to make spatial-temporal predictions of fish occupancy and habitat suitability indices. In doing

Table 1 Select model fit metrics relevant to classification problems

Metric	Calculation	Definition
Accuracy (A)	$(TP + TN)/(P + N)$	Overall prediction accuracy
Sensitivity (Se) Recall True positive rate	$TP/(TP + FN)$	True positive value prediction accuracy
Specificity (Sp) Selectivity True negative rate	$TN/(TN + FP)$	True negative value prediction accuracy
Precision (Pc) Positive predictive value (PPV)	$TP/(TP + FP)$	Positive predictive value accuracy
Negative predictive value (NPV)	$TN/(TN + FN)$	Negative predictive value accuracy
Area under the curve (AUC)	Integral of TPR vs FPR	Model performance in TP vs FP. Perfect value = 1, random selection in binary response = 0.5
Balanced accuracy	$Se + Sp/2$	Accuracy amongst classes weighted equally
True skill statistic (TSS)	$Se + Sp - 1$	Accuracy amongst classes weighted equally
No information rate (NIR)	$P/P + N/N + P$	Model accuracy if the dominant class was predicted in all cases
Kappa	Complex. See Cohen (1968)	Accuracy normalized for baseline of random chance
F1	$(2 * Pc * Se)/(Pc + Se)$	Harmonic mean of Precision and Sensitivity, integrating positive prediction accuracy and error
Minimized difference threshold (MDT)	$abs(Se - Sp)$	Measure of model fit that focuses on minimizing the differential between class accuracy, with lower values indicating better balance
Maximized sum threshold (MST)	$Se + Sp$	Measure of model fit that integrates balance amongst classes, with higher values indicating better balance

TP true positive, *FP* false positive, *TN* true negative, *FN* false negative, *P* total positives, *N* total negatives

so we identify optimal modeling techniques, model tuning, and focal fit metrics relative to the environmental management context. The findings are of relevance to any situation where spatial or spatial–temporal distributions of animal occupancy or habitat associations are being modelled—a common activity in fisheries management (e.g., Boisclair 2001).

Methods

Data collection

Largemouth bass (*Micropterus salmoides*) space use was measured over a nine-year period using acoustic telemetry in Toronto Harbour, Ontario, Canada (43.63°N, 79.36°W). Largemouth bass ($n = 144$; 400 ± 94 mm total length; mean \pm SD; 156–535 mm range) were captured via boat electrofishing and surgically implanted with acoustic transmitters (V7-4x, V9TP-2x, V13-1x, V13TP-1x, V13A-1x; 60–180 s or 130 to 270 s transmission delay; Vemco/InnovaSea Inc., Halifax, Nova Scotia) from 2010-09-08 to 2016-07-15 [see Midwood et al. (2019) for more details on fish capture and tagging]. All procedures were conducted in accordance with the Carleton University Animal Care Committee (application 110,723). Tagged largemouth bass were tracked with stationary acoustic receivers (Vemco VR2W; $n = 67$), which were grouped into 36 unique receiver nodes with corresponding habitat conditions in Toronto Harbour (Appendix S1; Fig. S1). Deployment periods at each receiver node were variable (Appendix S1; Fig. S2). At each receiver node, habitat variables were measured including the water depth (meters), percent cover of submerged aquatic vegetation (SAV), and wind exposure. Depth and SAV were measured by sonar within the detection range regions surrounding receivers, and wind exposure (fetch) via GIS measurement [see Midwood et al. (2019) for more details on habitat sampling]. Detection ranges were measured at a subset of receivers using stationary reference tags, and varied from 400 to 1500 m (see Veilleux 2014).

Data analysis

All data analyses were conducted in R v1.2.5019 (R Core Team 2019). Largemouth bass detections were passed through a series of filters to reduce the potential for false detections (Simpfendorfer et al. 2015). Firstly, detections that occurred simultaneously within a period shorter than the minimum tag delay were removed. Secondly, single detections that occurred at an acoustic receiver within a 2-h period were removed (Pincock 2012). Thirdly, spatial–temporal patterns of detections were plotted to visually examine for any shed tags or dead fish, which manifest as consistent

detections at a single acoustic receiver without subsequent detections on another (Klinard and Matley 2020; Matley et al. 2020). No tags exhibited this pattern. Filtered detections were used to calculate daily largemouth bass presence/absence at each acoustic receiver station over a nine-year period (2010-09-22 to 2019-10-31) only including days when a given acoustic receiver station was deployed, and the period when a minimum of five largemouth bass were being tracked.

To model largemouth bass habitat use and make spatial–temporal predictions of occupancy, daily largemouth bass presence/absence at each of the 36 receiver nodes was modelled with generalized linear mixed effects models (GLMM; Bolker et al. 2009) and random forests (RF; Breiman 2001). Predictors included water depth, SAV, wind exposure, and season. For GLMMs these predictors were fit as fixed effects, with receiver station as a random intercept and a logit link function. RF were fit with 1000 trees and the default number of variables was tried at each split at the square root of the number of predictors. Prior to modelling, the daily location-level largemouth bass presence/absence dataset ($n = 116,445$) was subset into ten partitions for cross validation. Initially, a single partition was used to fit GLMMs and RFs, with the remaining nine partitions held out as test data. In addition to unweighted GLMM and RF, weighted versions of each model type were fit to penalize misclassification of the minority class (presences) and balance prediction accuracy. This was accomplished by iteratively testing weighting combinations; balanced class accuracy was accomplished by weighting presences: absences at 5:1 for GLMM and 1:0.95 for RF. Overweighted versions of the models were also fit, with GLMM at 10:1 and RF 1:2. The weighting schemes are different with GLMM and RF implementation; these schemes produced similar model fit metrics at the default probability threshold (0.5). Relationships amongst probability of occurrence threshold values (zero to one) and true positive rate (TPR; accurately predicted presences), false positive rate (FPR; inaccurately predicted presences), true negative rate (TNR; accurately predicted absences), and false negative rate (FNR; inaccurately predicted absences) were examined for each model. TPR vs FPR were also plotted to visualize the Receiver Operating Characteristic (ROC) curve. Further, a range of fit metrics (Table 1) were calculated for each model by cross validation on test data (nine hold-out folds) with threshold presence probabilities ranging from 0.05 to 0.5.

We found RF were generally more flexible for development of non-linear fish-habitat relationships, and therefore the influence of model weighting was further explored with these models. Unweighted, weighted, and overweighted RF were fit with tenfold cross validation using one fold at a time to fit the models, and assessing model fit with the remaining nine folds. Variable importance was assessed using

mean decrease in accuracy (MDA), which is the percent decrease in model accuracy in trees where the variable was not included. MDA values were corrected to a proportion by dividing each variable MDA value by the sum of MDA values for all variables. Two-way variable interactions were calculated using Friedman's H-statistic, which also ranges from zero to one (Friedman and Popescu 2008). To examine modeled relationships between predictors and largemouth bass presence, each RF model (default, weighted, overweighted) was fit to the entire dataset and partial dependencies (\hat{y}) were calculated for each predictor-response combination, as well as two-way predictor interactions as a measure of seasonal habitat associations, or habitat suitability indices. Lastly, full RF models were used to make predictions for each season across the study area in Toronto Harbour to compare predictions amongst the weighting schemes. GLMM models were fit with the 'lme4' R package (Bates et al. 2015). RF were fit with the 'randomForests' package (Liaw and Wiener 2002), model fit metrics were calculated with the 'caret' package (Kuhn et al. 2019), variable interaction values were calculated with the 'iml' package (Molnar 2019), and partial dependencies were calculated with the 'pdp' package (Greenwell 2017).

Results

Over 9 years of tracking data resulted in 6,742,588 filtered detections of 144 largemouth bass at 67 acoustic receivers in 36 receiver nodes throughout Toronto Harbour (Appendix S1; Fig. S1). Tracking durations were variable amongst individuals (514 ± 348 days; mean \pm SD; 5–1250 day range; Appendix S1; Table S1, Fig. S3). Over time, largemouth bass were present at 23% of day/station combinations. GLMM and RF models with a range of weighting structures showed varied patterns of largemouth bass presence prediction rates depending on presence assignment threshold probability values (Fig. 1). Decreasing threshold values consistently resulted in higher TPR, FPR and lower TNR and FNR with both model types, but with a far greater impact on GLMMs (Fig. 1). Interestingly, altering probability thresholds with RF models had a limited capacity to achieve a range of TPR.

Unweighted GLMM and RF with a default threshold value of 0.5 had similar performance, including the highest accuracy of all models/thresholds (RF: 0.85; GLMM: 0.84; Fig. 2) but favoured the dominant class (absences), with high precision, specificity, and MDT, moderate Kappa and F1 scores, and low sensitivity, accuracy balance, and MST (see Table 1 for definition and calculation of fit metrics). Despite having high overall accuracy, sensitivity was low (RF: 0.40, GLMM: 0.43) indicating only 40–43% of true presences were predicted accurately. Weighted models prioritize presence accuracy, hence,

sensitivity increased drastically (RF: 0.77; GLMM: 0.79), with greater accuracy balance and a small increase in F1 score, but at a penalty to overall model accuracy (RF: 0.76; GLMM: 0.74), specificity, and Kappa (Fig. 2). These differences were more pronounced in overweighted models, increasing sensitivity (RF: 0.91; GLMM: 0.91) at the penalty of accuracy (RF: 0.64; GLMM: 0.63), precision, sensitivity, and accuracy balance. Decreasing probability thresholds for presence assignment generally had a similar effect to model weighting, and both combined had cumulative effects shifting models to higher sensitivity, but lower accuracy and precision. However, the effect of threshold variation was much greater in GLMM than RF (Fig. 2). Overall, weighted and overweighted GLMM and RF at a presence threshold of 0.5 had similar performance metrics and generally offered a better trade-off with sensitivity, precision, specificity, accuracy, and accuracy balance than decreasing model thresholds (Fig. 2).

Examining model weighting more closely with RF models, differently weighted models resulted in variations in variable importance and interaction scores (Fig. 3b) along with model fit metrics (Fig. 3a). Interactions between season and depth, SAV, and exposure were all important, as were all of the predictors alone, explaining a substantial amount of variation in largemouth bass occupancy (Fig. 3b). However, the order of predictor importance changed with different model weighting.

Largemouth bass were generally present most often at sites in Toronto Harbour in the summer, spring, and fall seasons, with marked decline in the winter (Fig. 4). Largemouth bass associated most often with either very low or moderate to high SAV, moderately shallow water depths, and low exposure environments, although sampling of higher exposure locations was sparse (Fig. 4). Examining the interaction between season and SAV, different patterns were generated amongst model weightings (Fig. 5). The best fitting RF model, RF weighted, found consistently high use of moderate-high SAV habitats amongst seasons, with the highest association with high SAV in the summer, and very low SAV in all seasons but winter (Fig. 5). Use of moderate water depths was more consistent seasonally (Fig. 5). The unweighted RF model generally failed to capture largemouth bass habitat associations, while the overweighted model exaggerated them, predicting greater probability of presences than raw data would suggest (Figs. 4, 5). This is reflected in the spatial-temporal patterns of model prediction amongst seasons in Toronto Harbour (Fig. 6).

Discussion

Modeling the spatial distribution and habitat associations of wild animals is of major relevance to ecosystem management and conservation. For example, understanding essential

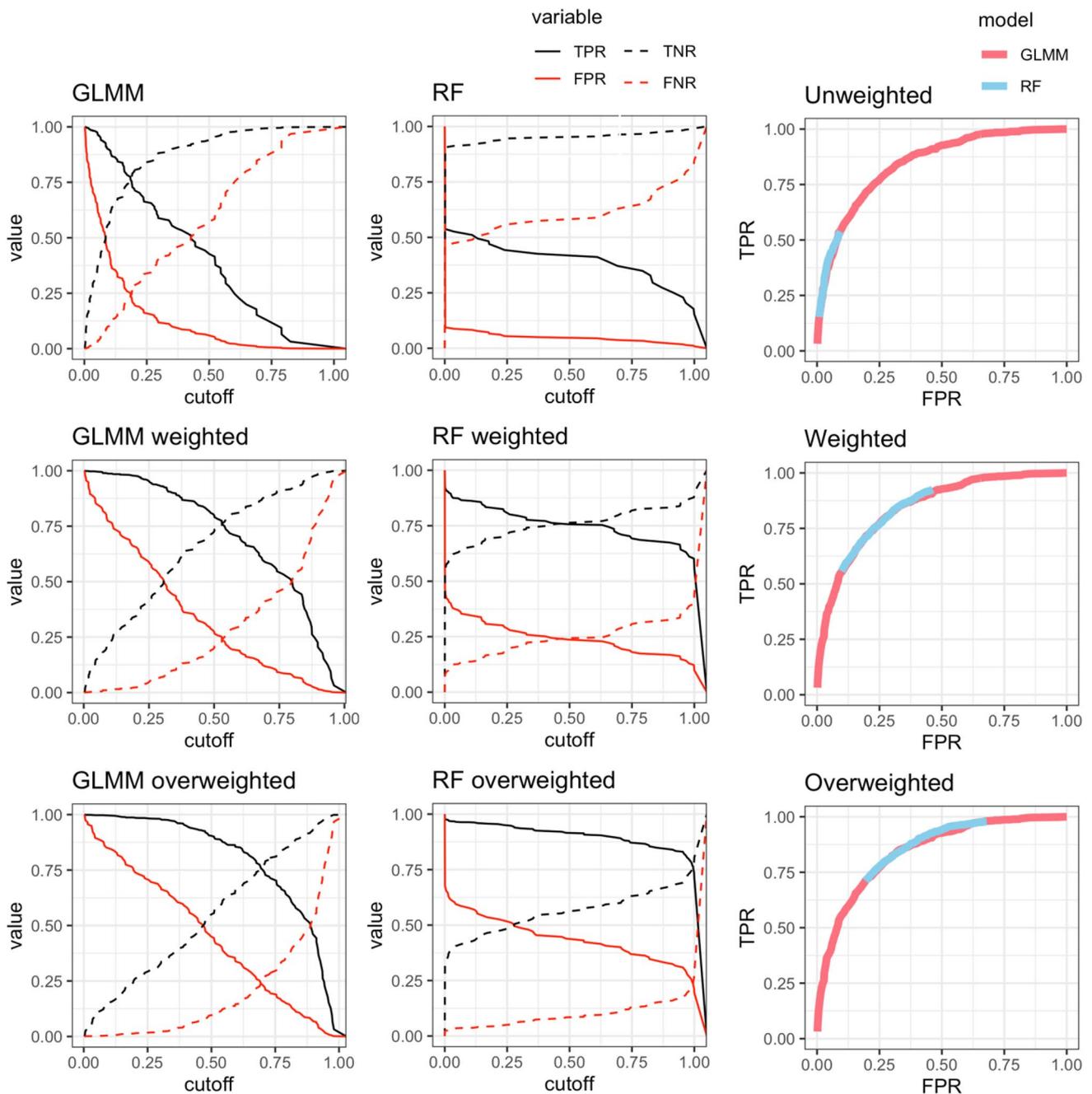


Fig. 1 Generalized mixed effects model (GLMM) and random forests (RF) model fit including true positive rate (TPR) false positive rate (FPR), true negative rate (TNR) and false negative rate (FNR) across a range of probability threshold values for predicting fish presences (left, center panels) and TPR vs FPR (receiver operating char-

acteristic; right panels). Weighted models are fit with penalization to prioritize the minority class where prediction accuracy is balanced amongst classes; overweighted models highly prioritize the minority class at 10:1

habitat needed to complete life cycles enables development of habitat protection measures (Goodchild 2004; Levin and Stunz 2005; Rosenfeld and Hatfield 2006; Heinrichs et al. 2010; Camaclang et al. 2015) and can also be used to identify opportunities for habitat restoration and enhancement (Miller et al. 2016). Advanced techniques like telemetry

enable more extensive measurements of animal space use, habitat associations, and distributions than point estimates from in person sampling, but large and complex datasets require particular analytical techniques to translate these data into usable knowledge. Treating acoustic telemetry data as a presence/absence classification exercise and focusing

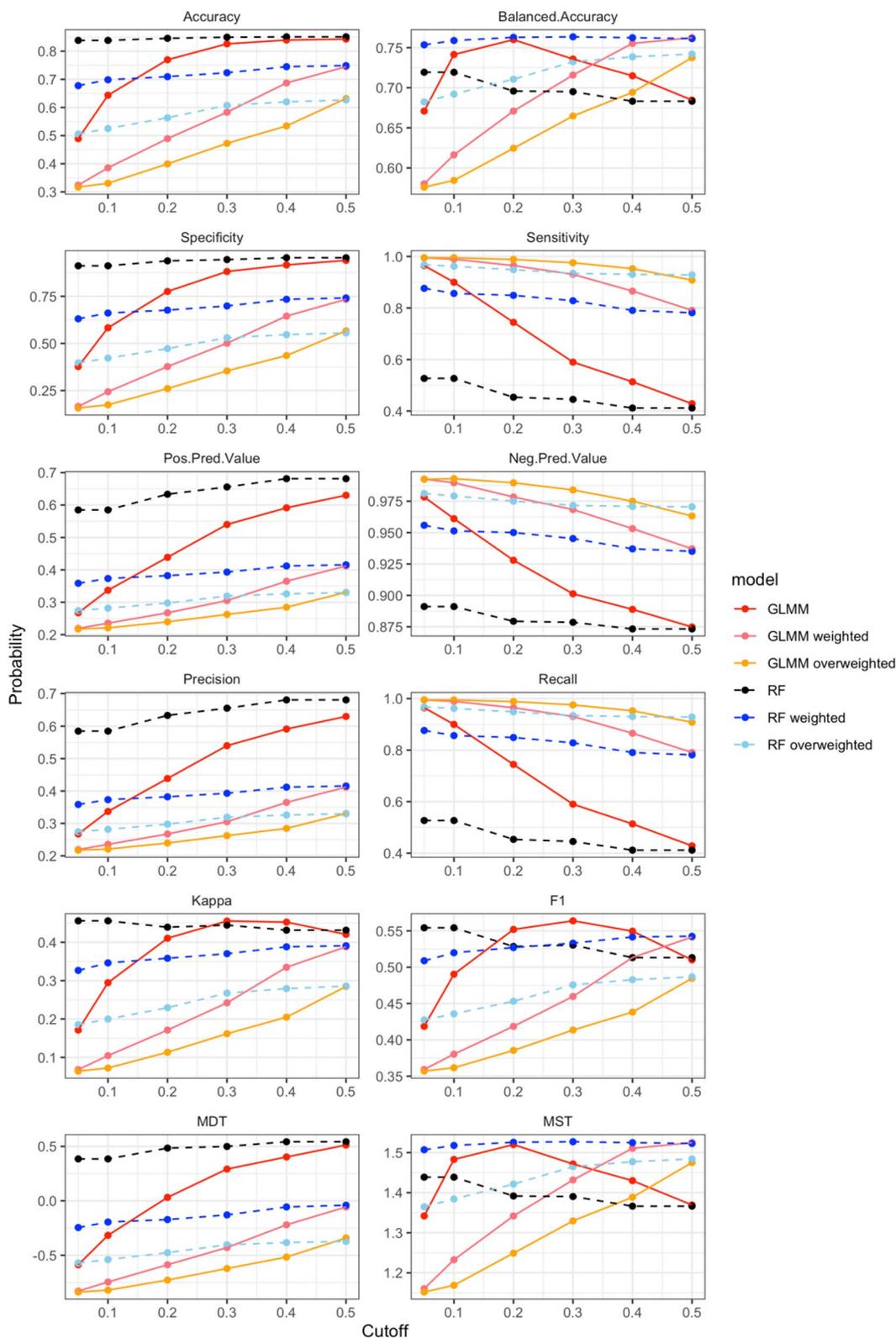


Fig. 2 Model fit metrics for a GLMM and RF models with a range of weighting structures and presence probability thresholds (Cutoff). See Table 1 for definition of metrics

on a variety of model fit metrics in non-training data, we demonstrated the importance of model tuning on derived habitat associations and predictions of spatial distribution using these techniques. Here we further discuss modeling considerations, including which model fit metrics should be the focus depending on the conservation application, as well as the ecological implications and further applications of our approach and findings.

Modeling considerations

In modeling species distributions and habitat associations, model fit is a universal concern regardless of which of the wide variety of model types are being applied (Elith et al. 2006; Guillera-Arroita et al. 2015). The perils of focusing solely on overall model fit are well documented, especially when data are imbalanced (e.g., Fielding and Bell 1997; Tarmansen et al. 2006; Jiménez-Valverde and Lobo 2007; Velez et al. 2007). This was demonstrated with the largemouth bass telemetry dataset—the unweighted (default) RF model had the highest prediction accuracy (85% in non-training data), yet only 40% of the largemouth bass presences were classified accurately. The derived habitat suitability indices (Fig. 4) and spatial distribution predictions (Fig. 5) were therefore inaccurate, and application of this model for any environmental management purpose would likely be ineffective. Altering model probability thresholds for presence assignment, as well as model weighting to prioritizing accurate classification of the minority class, both effectively addressed this issue with GLMM. However, RF were far less sensitive to probability thresholds and a limited range of TPR could be achieved without model weighting.

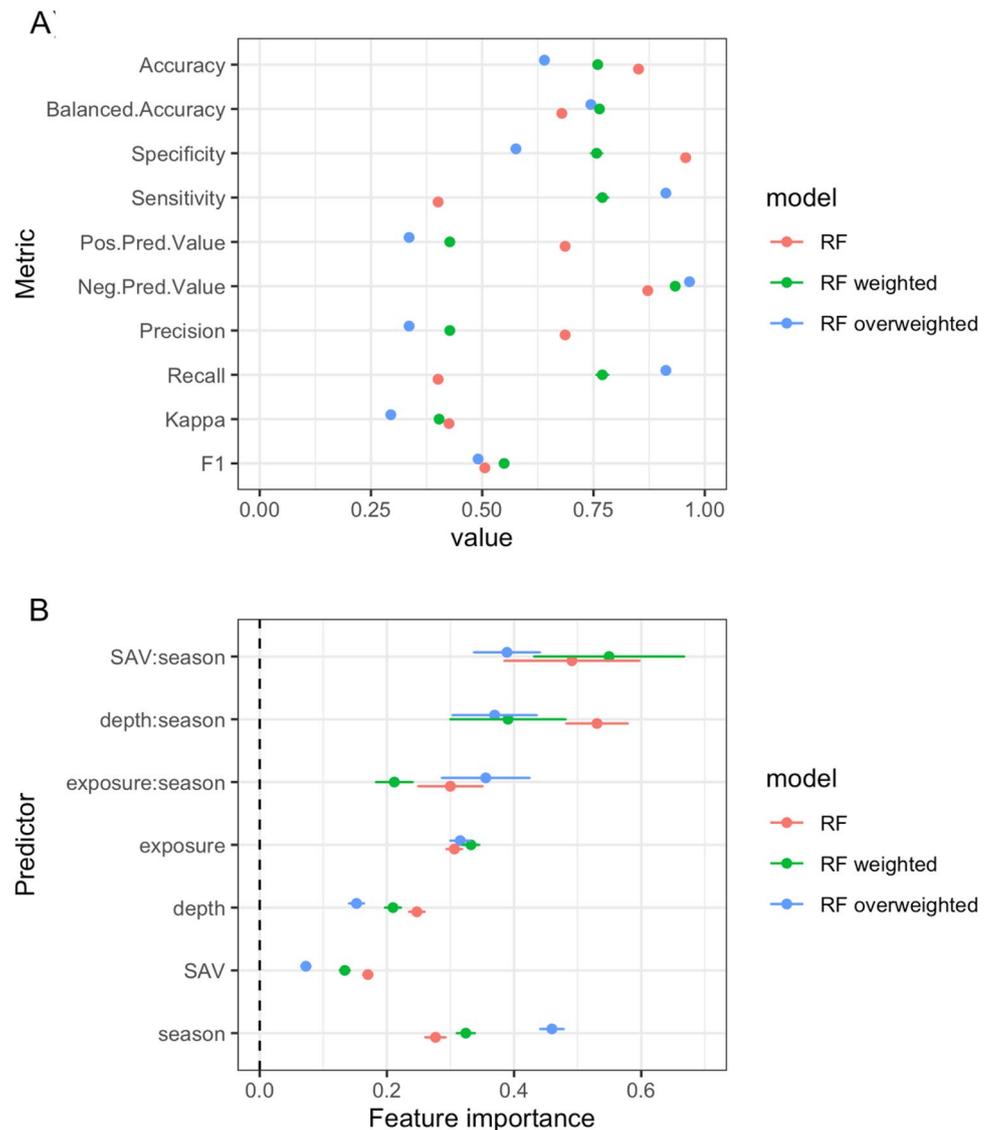
Weighting RF and GLMM models to balance class accuracy produced better models in terms of a complex set of fit metrics, warranting further discussion here. With classification problems there are a wide range of model fit metrics available (e.g., Fielding and Bell 1997; Kuhn et al. 2019), of which we consider a subset of relevant ones (Table 1). Despite the overwhelming number of fit metrics available, at their core classification problems come down to the balance of true positives, false positives, true negatives, and false negatives, from which fit metrics are derived. Some of these also consider the prevalence of each level of the response variable in the dataset (e.g., Kappa). For conservation issues related to animal distributions and habitat, in addition to overall model accuracy, researchers should focus on sensitivity (true presence accuracy), precision (predictive presence accuracy), and specificity (true negative accuracy), with a particular focus on the trade-off between sensitivity and precision. Although other commonly used metrics such as Area under the ROC curve (AUC), Kappa, and F1 can be informative of overall model fit and are insensitive to species prevalence (Lawson et al. 2014), for ecological/

conservation applications it is also prudent to consider the optimal trade-off between sensitivity and precision (i.e., the location along the ROC curve) given their influence on derived habitat associations and spatial distribution predictions (Figs. 3,4,5,6). In this way, consideration of model sensitivity and precision may be warranted in many applications regardless of whether the data are imbalanced.

By fitting GLMMs with a random effect of receiver location, they were capable of making similar quality of predictions as RF at the default probability assignment threshold (0.5). However, this case was a very simple one with a minimal number of predictors; with more complex datasets machine learning models such as RF, boosted regression trees, and neural networks have a number of advantages over common frequentist-based models (Cutler et al. 2007; Elith et al. 2008; Christin et al. 2019). Importantly, SAV and water depth were highly correlated, which we accepted for making predictions with GLMMs, but would likely bias the predictor coefficients (Morrissey and Ruxton 2018). RF and other machine learning algorithms assume no relationship between the predictors and response, but instead learn it algorithmically. They are generally more robust in this respect, enabling the inclusion of a large number of diverse predictor types to model complex, hierarchical relationships (Breiman 2001; Cutler et al. 2007; Elith et al. 2008). Although not well illustrated with our relatively simple application here, more advanced machine learning algorithms have many advantages for animal distribution and habitat models, especially for complex datasets derived from methods like telemetry. However, the relative insensitivity of RF to probability thresholds may result in poor model fit if continuous probability of occurrence is of interest (unless modelled explicitly), opposed to presence/absence (Pearce and Ferrier 2000). The major advantage that frequentist-based and Bayesian inference-based approaches have over machine learning algorithms is the capacity to integrate random effects, correlation structures, and variance structures to account for dependencies and autocorrelation in the data (Bolker et al. 2009; Zuur et al. 2009, 2017). Generally, machine learning algorithms rarely consider this (but see Buston and Elith 2011). Model fit is certainly something modelers should always assess amongst the predictors and over space and time where applicable. In this case we observed no obvious issues in this respect with either model type, perhaps due to the random subsampling regime implemented with cross validation, as well as the scale of the predictors chosen (i.e., season, rather than day).

A final modelling consideration is that animal presences are rarely, if ever, 100% detectable. It is for this reason that occupancy models often integrate uncertainty measures for measured ‘absences’ by generating presences at a frequency relative to the inverse of detection probability (MacKenzie et al. 2002; Lamothe et al. 2019). Acoustic

Fig. 3 **a** Model fit metrics and **b** predictor importance values from random forest models with a range of weighting structures



telemetry detection efficiency does vary predictably with environmental conditions (Brownscombe et al. 2020), and analytical approaches have been developed to integrate this in some contexts (e.g. Winton et al. 2018), yet, to date we are unaware of any studies that have integrated it into telemetry-based habitat or occupancy models. Temporal scale is likely an important consideration; for example, in our application with largemouth bass we focused on daily level presence/absence, with which only two detections were required on a given day to assign a presence. This approach may help to buffer the impacts of variations in detection efficiency, but we are not aware of any studies that have examined how telemetry receiver detection efficiency varies at this scale, nor do we have the data available to do so.

Applications

The optimal balance between model sensitivity and precision discussed above ultimately depends on the conservation application. In some scenarios, having a highly sensitive model would be a precautionary approach. For example, in a situation where environmental managers are advising a shoreline development project, they may wish to ensure a species (perhaps a highly sensitive species at risk) is unlikely to be occupying this location/habitat at a specific time to advise when and where the work can be completed. Overweighting a model to have high sensitivity for making predictions would be more precautionary, despite the fact that precision would be low, i.e., there is a reasonable chance the species is not there. In this case false negatives are more of a conservation concern. Highly sensitive models may also

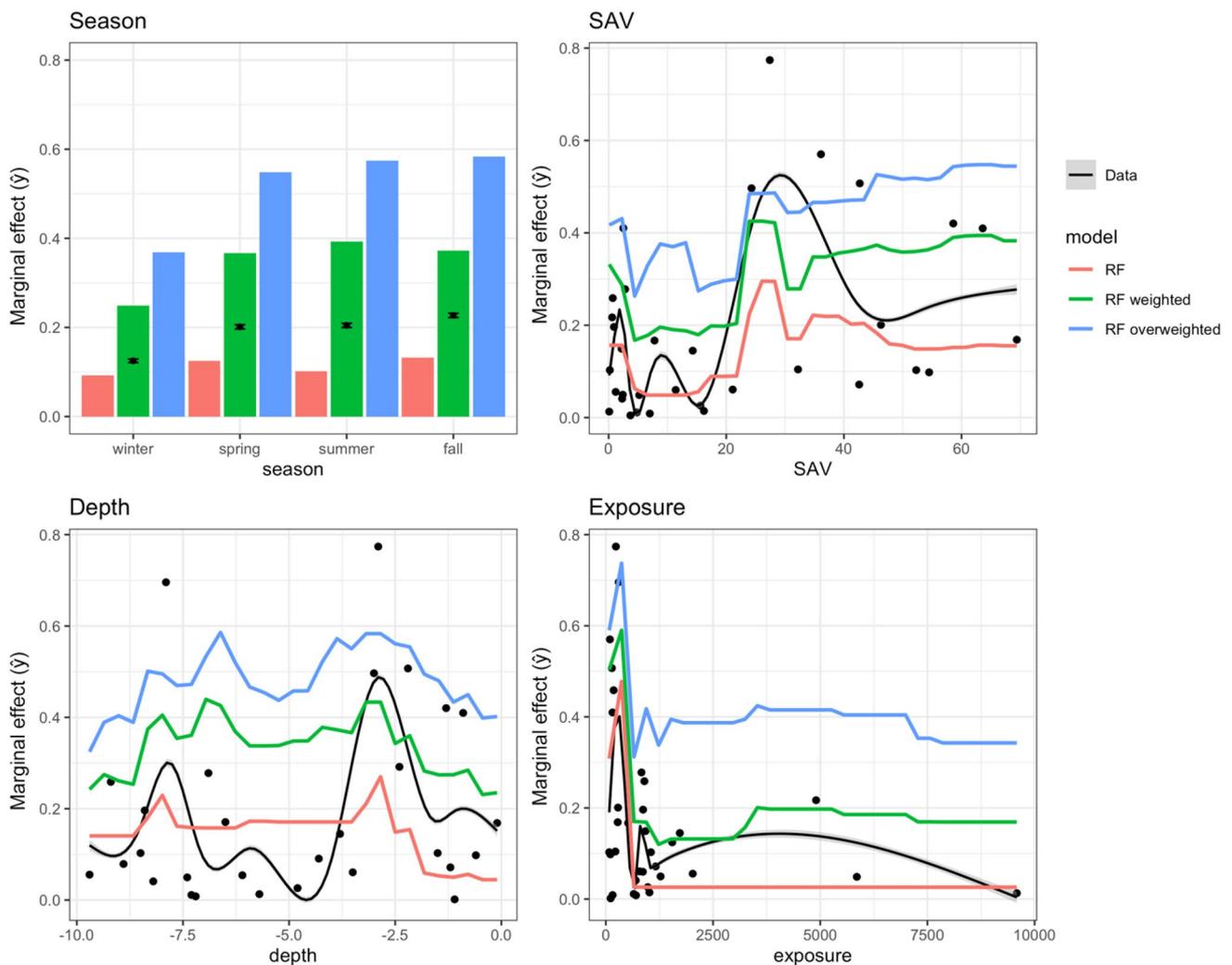


Fig. 4 Partial dependency plots of marginal predicted values (\hat{y} ; colors) from random forests models of largemouth bass site occupancy amongst environmental and seasonal predictors including raw data fit with a loess smoother (black)

help to buffer detection probability errors, although this has not been explored experimentally.

In other contexts, practitioners may prefer to prioritize model precision over sensitivity. For example, overestimation of species distributions could create pitfalls in conservation planning by focusing conservation and management efforts (e.g., habitat restoration) on the wrong areas (Loiselle et al. 2003; Rondinini et al. 2006). Similarly, in working with species at risk, models that generate overestimates of species distribution may result in failures to take necessary conservation actions. In such cases, models with higher precision may be more precautionary. Notably, in population recovery contexts, predicted suitable habitats may be worthwhile of resource allocation even if they are not currently occupied.

Analytical approaches applied to telemetry data commonly focus on modeling animal movement and space use at

broader temporal scales (Brownscombe et al. 2019; Whoriskey et al. 2019). Using kernel density estimation methods to estimate broader space use outside of acoustic receiver stations, Midwood et al. (2019) found similar patterns of largemouth bass space use in Toronto Harbour using the same dataset applied here. However, a major notable difference is in movement pathways, where total detections and residency were low, but occupancy at the daily level was frequent. Hence, the occupancy modeling approach applied here identified these deeper-water habitats as important and predicted largemouth bass presence there. In terms of management applications, movement pathways are an essential component of critical habitat (Rosenfeld and Hatfield 2006) and may also be well captured using network methods (Finn et al. 2014; Jacoby and Freeman 2016; Whoriskey et al. 2019).

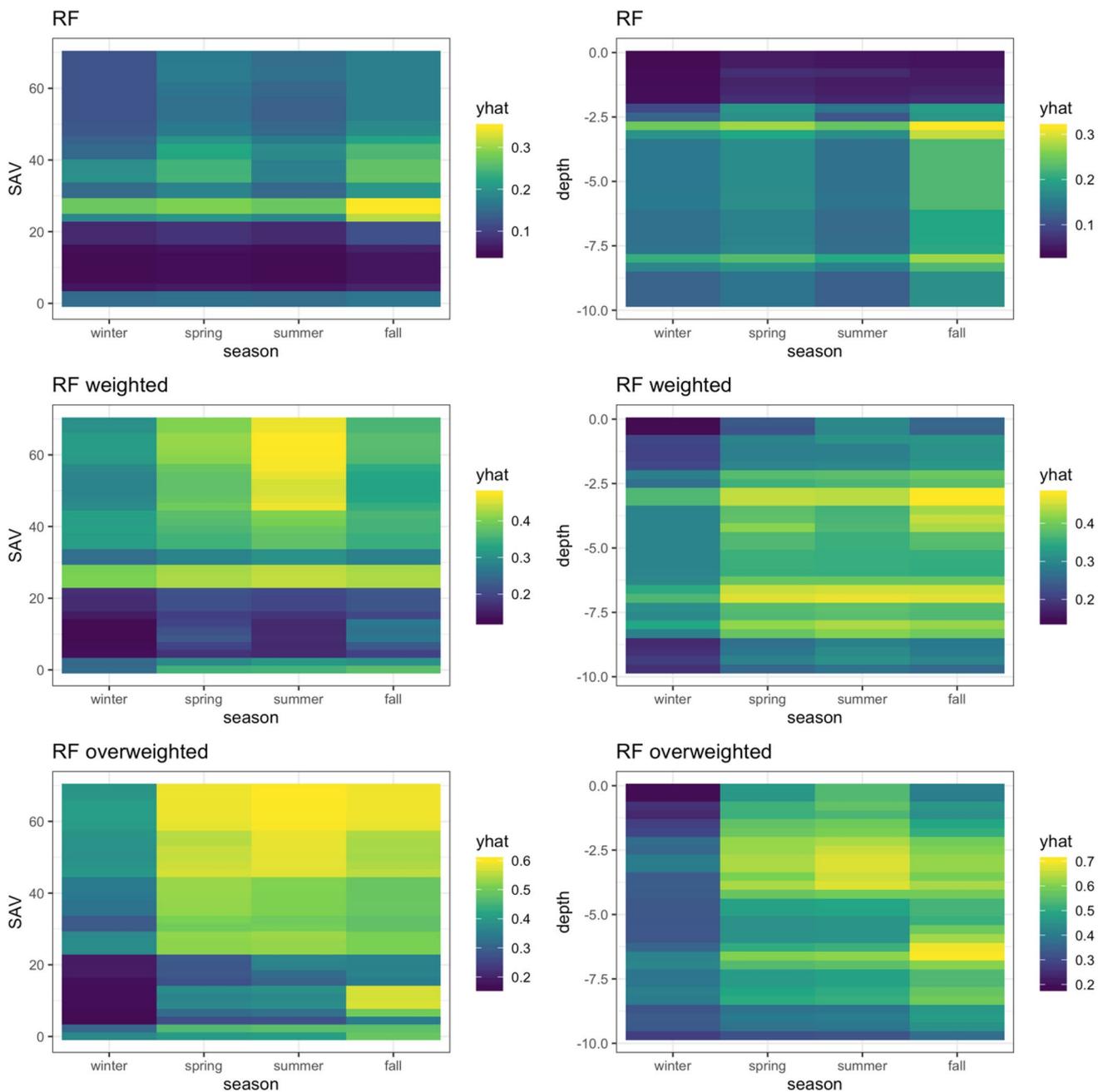


Fig. 5 Partial dependency plots of marginal predicted values (\hat{y}) from random forests models of largemouth bass site occupancy showing interactions between season, submerged aquatic vegetation (%), and water depth (m)

There is a wide range of applications of species occupancy models, of which some cursory applications for generating species distribution and habitat associations are illustrated here. In the context of fish habitat management, habitat suitability indices (HSI) play an important role for understanding fish-habitat relationships for managing habitat protection and restoration (Ahmadi-Nedushan et al. 2006; de Kerckhove et al. 2008). By focusing on presence/absence of largemouth bass relative to habitat conditions

and seasons, we generated a type II HSI, a measure of habitat use, rather than a type III HSI, which represents habitat selection relative to availability (de Kerckhove et al. 2008). The advantages of one approach over the other appear to vary depending on context (Long et al. 2009). There are methods readily available to calculate habitat selection using acoustic telemetry data, which requires data on the habitat availability in the system (e.g., Griffin et al. 2020).

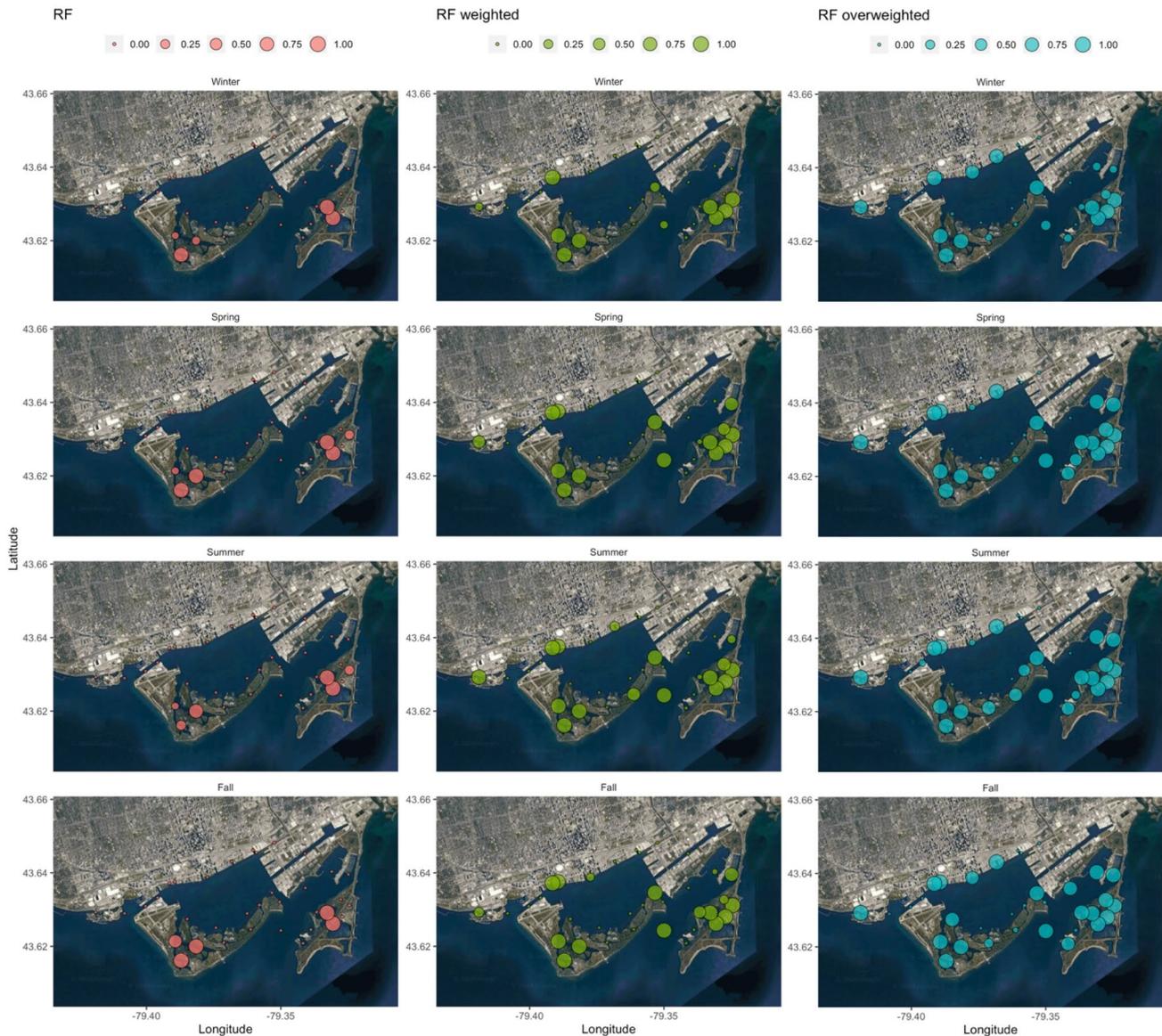


Fig. 6 Predicted largemouth bass occupancy probabilities amongst seasons with random forests models with three different weighting structures including unweighted (RF; red), weighted for class balance

(RF weighted; green) and overweighted with a focus on presence accuracy (RF overweighted; blue)

With our dataset we were able to determine that largemouth bass occupied moderately shallow, high SAV habitats with low wind exposure most frequently, followed by deep, low SAV habitats, with lower use of intermediate habitats. These moderately shallow-high SAV-low exposure environments were sheltered embayments in Toronto Harbour (Fig. 6), which were especially important largemouth bass habitats in the summer season (Fig. 5). Largemouth bass presence was markedly lower in the winter season, which, based on occupancy alone, might lead to the conclusion they are exiting Toronto Harbour. However, patterns of movement suggest they remain in the Harbour, but are just

rarely detected by the receiver configuration (Midwood et al. 2019). This highlights the need to combine a range of movement and occupancy approaches to understand habitat use at multiple scales.

Summary

We applied several modelling techniques with a range of tuning parameters to assess their impact on derived habitat and species distribution models using a long-term fish tracking data set. The findings highlight the potential for telemetry data to contribute to the development

of comprehensive models of habitat and distribution with applicability to environmental management and conservation. However, they also emphasize the importance of appropriate model tuning for effective application. Although there are many model fit metrics available, we suggest the primary concern for species occupancy applications should be, in addition to accuracy, model sensitivity and precision, with which there is a general trade-off between true positive and predictive positive accuracy. This extends beyond model selection (e.g., AUC), to model tuning to select the optimal trade-off for the application. GLMMs produced similar quality models to RFs in this simple application, although RF and other more complex algorithms may be more robust to complex datasets. In future applications of telemetry to model habitat, researchers may consider integrating measures of detection efficiency to account for false absences, as well as the potential for model weighting and temporal averaging to correct for this issue. Our findings highlight the need to consider more complex aspects of model fit than overall accuracy in many applications modelling animal distributions and habitat associations.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s00027-021-00797-5>.

Acknowledgements Funding for the Toronto and Region Area of Concern telemetry project was provided by the Natural Sciences and Engineering Research Council of Canada (Strategic and Discovery Grants to Mathew G. Wells and Steven J. Cooke) with additional support from the Great Lakes Sustainability Fund administered by Environment and Climate Change Canada (formerly Environment Canada), and the Great Lakes Fishery Commission. We greatly appreciate the help with implementation and with data management from collaborators on this project including the Toronto and Region Conservation Authority (Rick Porttiss, Gord MacPherson, Brian Graham, Adam Weir, Pete Shuttleworth, Bradley Bloemendal, Meg St. John, and Ross Davidson among others), Carleton University (Nicholas Lapointe, Lee Gutowsky, Maxime Veilleux, Liset Cruz-Font, Tyler Peat, Andrew Rous, and Robert Lennox), University of Toronto Scarborough (Mathew Wells and Bogdan Hlevca), and Fisheries and Oceans Canada (Susan Doka, Christine Boston, David Reddick, and Erin Budgett).

Author contributions JWB developed the conceptual ideas, analyzed the data and wrote the manuscript. JDM and SJC contributed to project development, funding, conceptualization, and manuscript editing.

Funding Funding for the Toronto and Region Area of Concern telemetry project was provided by the Natural Sciences and Engineering Research Council of Canada (Strategic and Discovery Grants to Mathew G. Wells and Steven J. Cooke) with additional support from the Great Lakes Sustainability Fund administered by Environment and Climate Change Canada (formerly Environment Canada), and the Great Lakes Fishery Commission.

Availability of data and materials Upon acceptance, data for this manuscript will be made available through Zenodo.

Code availability Upon acceptance, R code will be made available on Git.

Declarations

Conflict of interest None.

Ethics approval All procedures were conducted in accordance with the Carleton University Animal Care Committee (application 110723).

Consent to participate Not applicable.

Consent for publication All authors consent to the submission of this manuscript, it is published are under review in another outlet.

References

- Ahmadi-Nedushan B, St-Hilaire A, Bérubé M et al (2006) A review of statistical methods for the evaluation of aquatic habitat suitability for instream flow assessment. *River Res Appl* 22:503–523. <https://doi.org/10.1002/rra.918>
- Allen AM, Singh NJ (2016) Linking movement ecology with wildlife management and conservation. *Front Ecol Evol*. <https://doi.org/10.3389/fevo.2015.00155>
- Bates D, Mächler M, Bolker BM, Walker SC (2015) Fitting linear mixed-effects models using lme4. <https://doi.org/10.18637/jss.v067.i01>. arXiv: 1406.5823
- Boisclair D (2001) Fish habitat modeling: from conceptual framework to functional tools. *Can J Fish Aquat Sci*. <https://doi.org/10.1139/f00-251>
- Bolker BM, Brooks ME, Clark CJ et al (2009) Generalized linear mixed models: a practical guide for ecology and evolution. *Trends Ecol Evol* 24:127–135. <https://doi.org/10.1016/j.tree.2008.10.008>
- Bonar SA, Hubert WA (2002) Standard sampling of inland fish: benefits, challenges, and a call for action. *Fisheries* 27:10–16. [https://doi.org/10.1577/1548-8446\(2002\)027%3c0010:ssoif%3e2.0.co;2](https://doi.org/10.1577/1548-8446(2002)027%3c0010:ssoif%3e2.0.co;2)
- Breiman L (2001) Random forests. *Mach Learn* 45:5–32. <https://doi.org/10.1023/A:1010933404324>
- Brooks JL, Chapman JM, Barkley AN et al (2018) Biotelemetry informing management: case studies exploring successful integration of biotelemetry data into fisheries and habitat management. *Can J Fish Aquat Sci*. <https://doi.org/10.1139/cjfas-2017-0530>
- Brownscombe JW, Ledee E, Raby GD et al (2019) Conducting and interpreting fish telemetry studies: considerations for researchers and resource managers. *Rev Fish Biol Fish* 29:369–400
- Brownscombe JW, Griffin LP, Chapman JM et al (2020) A practical method to account for variation in detection range in acoustic telemetry arrays to accurately quantify the spatial ecology of aquatic animals. *Methods Ecol Evol* 11:82–94. <https://doi.org/10.1111/2041-210X.13322>
- Buston PM, Elith J (2011) Determinants of reproductive success in dominant pairs of clownfish: a boosted regression tree analysis. *J Anim Ecol* 80:528–538. <https://doi.org/10.1111/j.1365-2656.2011.01803.x>
- Camaclang AE, Maron M, Martin TG, Possingham HP (2015) Current practices in the identification of critical habitat for threatened species. *Conserv Biol* 29:482–492. <https://doi.org/10.1111/cobi.12428>
- Cappo MA, Harvey EB, Malcolm HC, Speare PA (1999) Potential of video techniques to monitor diversity, abundance and size of fish in studies of marine protected areas. *Video Tech to Monit Fish MPAs*, pp 455–464

- Christin S, Hervet É, Lecomte N (2019) Applications for deep learning in ecology. *Methods Ecol Evol* 10:1632–1644. <https://doi.org/10.1111/2041-210X.13256>
- Cohen J (1968) Weighted kappa: nominal scale agreement provision for scaled disagreement or partial credit. *Psychol Bull* 70:213. <https://doi.org/10.1037/h0026256>
- Crossin GT, Heupel MR, Holbrook CM et al (2017) Acoustic telemetry and fisheries management. *Ecol Appl* 27:1031–1049. <https://doi.org/10.1002/eap.1533>
- Cutler DR, Edwards TC, Beard KH et al (2007) Random forests for classification in ecology. *Ecology* 88:2783–2792. <https://doi.org/10.1890/07-0539.1>
- de Kerckhove DT, Smokorowski KE, Randall RG (2008) A primer on fish habitat models Canadian technical report of fisheries and aquatic sciences 2817. *Can Tech Rep Fish Aquat Sci* 2817:65
- Elith J, Graham CH, P. Anderson R, et al (2006) Novel methods improve prediction of species' distributions from occurrence data. *Ecography (Cop)* 29:129–151. <https://doi.org/10.1111/j.2006.0906-7590.04596.x>
- Elith J, Leathwick JR, Hastie T (2008) A working guide to boosted regression trees. *J Anim Ecol* 77:802–813. <https://doi.org/10.1111/j.1365-2656.2008.01390.x>
- Evans JS, Murphy MA, Holden ZA, Cushman SA (2011) Modeling species distribution and change using random forest. *Predictive species and habitat modeling in landscape ecology*. Springer, New York, pp 139–159
- Fielding A, Bell J (1997) A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environ Conserv* 24:38–49. <https://doi.org/10.1017/S037689290021214>
- Finn JT, Brownscombe JW, Haak CR et al (2014) Applying network methods to acoustic telemetry data: modeling the movements of tropical marine fishes. *Ecol Model* 293:139–149. <https://doi.org/10.1016/j.ecolmodel.2013.12.014>
- Friedman JH, Popescu BE (2008) Predictive learning via rule ensembles. *Ann Appl Stat* 2:916–954. <https://doi.org/10.1214/07-AOAS148>
- Goodchild GA (2004) Fish habitat is everyone's business, Canada's fish habitat management programme. *Fish Manag Ecol* 11:277–281. <https://doi.org/10.1111/j.1365-2400.2004.00409.x>
- Greenwell BM (2017) pdp: An R package for constructing partial dependence plots. *R J*. <https://doi.org/10.32614/rj-2017-016>
- Griffin LP, Smith BJ, Cherkiss MS et al (2020) Space use and relative habitat selection for immature green turtles within a Caribbean marine protected area. *Anim Biotelem* 8:1–13. <https://doi.org/10.1186/s40317-020-00209-9>
- Guillera-Arroita G, Lahoz-Monfort JJ, Elith J et al (2015) Is my species distribution model fit for purpose? Matching data and models to applications. *Glob Ecol Biogeogr* 24:276–292. <https://doi.org/10.1111/geb.12268>
- Heinrichs JA, Bender DJ, Gummer DL, Schumaker NH (2010) Assessing critical habitat: evaluating the relative contribution of habitats to population persistence. *Biol Conserv* 143:2229–2237. <https://doi.org/10.1016/j.biocon.2010.06.009>
- Hussey NE, Kessel ST, Aarestrup K et al (2015) Aquatic animal telemetry: a panoramic window into the underwater world. *Science* 348:1255642. <https://doi.org/10.1126/science.1255642>
- Jacoby DMP, Freeman R (2016) Emerging network-based tools in movement ecology. *Trends Ecol Evol* 31:301–314. <https://doi.org/10.1016/j.tree.2016.01.011>
- Jiménez-Valverde A, Lobo JM (2007) Threshold criteria for conversion of probability of species presence to either-or presence-absence. *Acta Oecologica* 31:361–369. <https://doi.org/10.1016/j.actao.2007.02.001>
- Klinard NV, Matley JK (2020) Living until proven dead: addressing mortality in acoustic telemetry research. *Rev Fish Biol Fish* 30:485–499
- Kuhn M, Wing J, Weston S et al (2019) caret: classification and regression training. *Astrophysics Source Code Library, record ascl:1505.003*
- Lamothe KA, Dextrase AJ, Drake DAR (2019) Characterizing species co-occurrence patterns of imperfectly detected stream fishes to inform species reintroduction efforts. *Conserv Biol* 33:1392–1403. <https://doi.org/10.1111/cobi.13320>
- Lawson CR, Hodgson JA, Wilson RJ, Richards SA (2014) Prevalence, thresholds and the performance of presence-absence models. *Methods Ecol Evol* 5:54–64. <https://doi.org/10.1111/2041-210X.12123>
- Levin PS, Stunz GW (2005) Habitat triage for exploited fishes: can we identify essential "Essential Fish Habitat?" *Estuar Coast Shelf Sci* 64:70–78. <https://doi.org/10.1016/j.ecss.2005.02.007>
- Liaw A, Wiener M (2002) Classification and regression by randomForest. *R News* 2:18–22. <https://doi.org/10.1177/154405910408300516>
- Liu C, Berry PM, Dawson TP, Pearson RG (2005) Selecting thresholds of occurrence in the prediction of species distributions. *Ecography (Cop)* 28:385–393. <https://doi.org/10.1111/j.0906-7590.2005.03957.x>
- Lobo JM, Jiménez-valverde A, Real R (2008) AUC: a misleading measure of the performance of predictive distribution models. *Glob Ecol Biogeogr* 17:145–151. <https://doi.org/10.1111/j.1466-8238.2007.00358.x>
- Loiselle BA, Howell CA, Graham CH et al (2003) Avoiding pitfalls of using species distribution models in conservation planning. *Conserv Biol* 17:1591–1600. <https://doi.org/10.1111/j.1523-1739.2003.00233.x>
- Long RA, Muir JD, Rachlow JL, Kie JG (2009) A comparison of two modeling approaches for evaluating wildlife-habitat relationships. *J Wildl Manag* 73:294–302. <https://doi.org/10.2193/2008-052>
- MacKenzie DI, Nichols JD, Lachman GB et al (2002) Estimating site occupancy rates when detection probabilities are less than one. *Ecology* 83:2248. <https://doi.org/10.2307/3072056>
- MacLeod CD, Mandelberg L, Schweder C et al (2008) A comparison of approaches for modelling the occurrence of marine animals. *Essential fish habitat mapping in the mediterranean*. Springer, Dordrecht, pp 21–32
- Marzluff JM, Knick ST, Millsbaugh JJ (2001) Chapter 12. High-tech behavioral ecology: modeling the distribution of animal activities to better understand wildlife space use and resource selection. *Radio tracking and animal populations*. Elsevier
- Matley JK, Faust MD, Raby GD et al (2020) Seasonal habitat-use differences among Lake Erie's walleye stocks. *J Great Lakes Res* 46:609–621. <https://doi.org/10.1016/j.jglr.2020.03.014>
- Midwood JD, Rous AM, Doka SE (2019) Acoustic telemetry in Toronto Harbour: assessing residency, habitat selection, and within-harbour movements of fishes over a five-year period. In: *Canadian technical report of fisheries and aquatic sciences*, 3331, p 174
- Miller JR, Miller JR, Hobbs RJ (2016) Habitat restoration—do we know what we're doing? *Restor Ecol* 15:382–390. <https://doi.org/10.1111/j.1526-100X.2007.00234.x>
- Molnar C (2019) Interpretable machine learning. A guide for making black box models explainable. *Lean Publishing*
- Morrissey MB, Ruxton GD (2018) Multiple regression is not multiple regressions: the meaning of multiple regression and the non-problem of collinearity. *Philos Theory Pract Biol*. <https://doi.org/10.3998/ptpbio.16039257.0010.003.10.3998/ptpbio.16039257.0010.003>

- Pearce J, Ferrier S (2000) Evaluating the predictive performance of habitat models developed using logistic regression. *Ecol Model* 133:225–245. [https://doi.org/10.1016/S0304-3800\(00\)00322-7](https://doi.org/10.1016/S0304-3800(00)00322-7)
- Pincock DG (2012) False detections: what they are and how to remove them from detection data. VEMCO Whitepaper Document DOC-004691, Amirix Systems Inc., Halifax, NS, Canada
- R Core Team (2019) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>
- Rondinini C, Wilson KA, Boitani L et al (2006) Tradeoffs of different types of species occurrence data for use in systematic conservation planning. *Ecol Lett* 9:1136–1145. <https://doi.org/10.1111/j.1461-0248.2006.00970.x>
- Rosenfeld JS, Hatfield T (2006) Information needs for assessing critical habitat of freshwater fish. *Can J Fish Aquat Sci* 63:683–698. <https://doi.org/10.1139/f05-242>
- Rountree RA, Gilmore RG, Goudey CA et al (2006) Listening to fish: applications of passive acoustics to fisheries science. *Fisheries* 31:433–446. [https://doi.org/10.1577/1548-8446\(2006\)31\[433:LTF\]2.0.CO;2](https://doi.org/10.1577/1548-8446(2006)31[433:LTF]2.0.CO;2)
- Simpfendorfer CA, Huvneers C, Steckenreuter A et al (2015) Ghosts in the data: false detections in VEMCO pulse position modulation acoustic telemetry monitoring equipment. *Anim Biotelem* 65:482–492. <https://doi.org/10.1139/F07-180>
- Termansen M, McClean CJ, Preston CD (2006) The use of genetic algorithms and Bayesian classification to model species distributions. *Ecol Model* 192:410–424. <https://doi.org/10.1016/j.ecolmodel.2005.07.009>
- Veilleux M (2014) Spatial ecology of fish in Toronto Harbour in response to aquatic habitat enhancement. Carleton University
- Velez DR, White BC, Motsinger AA et al (2007) A balanced accuracy function for epistasis modeling in imbalanced datasets using multifactor dimensionality reduction. *Genet Epidemiol* 31:306–315. <https://doi.org/10.1002/gepi.20211>
- Whoriskey K, Martins EG, Auger-Methe M et al (2019) A guide to current and emerging statistical methods for analyzing aquatic telemetry data. *Methods Ecol Evol* 10:935–948. <https://doi.org/10.1111/2041-210X.13188>
- Winton MV, Kneebone J, Zemeckis DR, Fay G (2018) A spatial point process model to estimate individual centres of activity from passive acoustic telemetry data. *Methods Ecol Evol* 9:2262–2272. <https://doi.org/10.1111/2041-210X.13080>
- Zuur AF, Ieno EN, Walker NJ et al (2009) Mixed effects models and extensions in ecology with R. *Stat Biol Heal*. <https://doi.org/10.1007/978-0-387-87458-6>
- Zuur AF, Ieno EN, Anatoly et al (2017) Beginner's guide to spatial, temporal, and spatial-temporal ecological data analysis with R-INLA. Highl Stat Ltd. ISBN: 978:1–12

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.