





Data-centric species distribution modeling: Impacts of modeler decisions in a case study of invasive European frog-bit

Sara E. Hansen¹  | Michael J. Monfils² | Rachel A. Hackett²  | Ryan T. Goebel¹  |
Anna K. Monfils¹ 

¹Central Michigan University, 2401 Biosciences Building, Mount Pleasant, Michigan 48858, USA

²Michigan Natural Features Inventory, Michigan State University, 1st Floor Constitution Hall, 525 W. Allegan St., Lansing, Michigan 48933, USA

Correspondence

Sara E. Hansen, Central Michigan University, 2401 Biosciences Building, Mount Pleasant, Michigan 48858, USA.
Email: hanse2s@cmich.edu

This article is part of the special issue “From Theory to Practice: New Innovations and Their Application in Conservation Biology.”

Abstract

Premise: Species distribution models (SDMs) are widely utilized to guide conservation decisions. The complexity of available data and SDM methodologies necessitates considerations of how data are chosen and processed for modeling to enhance model accuracy and support biological interpretations and ecological applications.

Methods: We built SDMs for the invasive aquatic plant European frog-bit using aggregated and field data that span multiple scales, data sources, and data types. We tested how model results were affected by five modeler decision points: the exclusion of (1) missing and (2) correlated data and the (3) scale (large-scale aggregated data or systematic field data), (4) source (specimens or observations), and (5) type (presence-background or presence-absence) of occurrence data.

Results: Decisions about the exclusion of missing and correlated data, as well as the scale and type of occurrence data, significantly affected metrics of model performance. The source and type of occurrence data led to differences in the importance of specific explanatory variables as drivers of species distribution and predicted probability of suitable habitat.

Discussion: Our findings relative to European frog-bit illustrate how specific data selection and processing decisions can influence the outcomes and interpretation of SDMs. Data-centric protocols that incorporate data exploration into model building can help ensure models are reproducible and can be accurately interpreted in light of biological questions.

KEYWORDS

aquatic invasive plants, data-centric, *Hydrocharis morsus-ranae*, machine learning, natural history collections, open data, reproducible research, species distribution model

Ecological models provide guidance for addressing critical conservation issues including but not limited to biodiversity loss, accelerating environmental change, and the spread of diseases (Schuwirth et al., 2019). Species distribution models, for example, have been applied in the preservation of endangered species (e.g., Belitz et al., 2019; Su et al., 2021; Charbonnel et al., 2023), predicting range shifts resulting from climate change (e.g., Bond et al., 2011; Borzée et al., 2019), and mitigating the spread and impact of invasive species (e.g., Mainali et al., 2015; Barbet-Massin et al., 2018; Lozano, 2021). Correlative species distribution

models (SDMs), also called habitat suitability models or ecological niche models, connect species occurrence data with environmental and location-specific information at the sites they are observed (Elith and Leathwick, 2009; Guillera-Aroita et al., 2015). The results of SDMs can be used to identify the drivers of historical and current species distribution (Elith and Franklin, 2013; Belitz et al., 2019) and estimate the potential distribution or, more realistically, relative probability of suitable habitat in areas within and outside of the sampled region (Elith and Leathwick, 2009). The SDM process is continuously evaluated and refined

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial-NoDerivs](https://creativecommons.org/licenses/by-nc-nd/4.0/) License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2024 The Authors. *Applications in Plant Sciences* published by Wiley Periodicals LLC on behalf of Botanical Society of America.

(Valavi et al., 2022; Hui, 2023) so that model results can be effectively applied to urgent ecological problems. Model performance is influenced by decisions made by modelers during model building (Vignali et al., 2020; Valavi et al., 2022); consequently, these decisions affect which models are chosen for application and how the results are interpreted in light of the biological questions being addressed (Zarzo-Arias et al., 2022). Real-world ecological decisions based on SDMs are entirely dependent on the input data; if the data are inappropriate for the questions being asked or data processing methodologies are not optimized, then the resulting conservation actions could be misguided. However, data selection and processing that occurs prior to model building is not always reported in SDM literature and could even be hidden in modeling programs, so modelers may not recognize which of their decisions could change model outcomes. Literature on species distribution modeling has addressed modeling methods (Elith et al., 2006; Jarnevich, 2017; Valavi et al., 2022), parameter optimization (Vignali et al., 2020), data quality and bias (Tessarolo et al., 2021; Zarzo-Arias et al., 2022), data management (Zuckerberg et al., 2010), explanatory variable selection and resolution (Austin and Van Niel, 2011), and translation of predictions to real-world issues (Guisan et al., 2013). In this study, we build upon this field of SDM evaluation and refinement by exploring the impact of five specific modeling decision points, two related to data processing and three related to data selection, on the performance and outcomes of SDMs for the invasive aquatic plant European frog-bit (*Hydrocharis morsus-ranae* L.) in the Laurentian Great Lakes region. Data-centric modeling practices, which we define here as identifying and controlling for the effects of input data features on the end results of models, can help ensure decisions at these five points are reproducible and lead to positive statistical outcomes and biological interpretability.

Occurrence data for European frog-bit provide a unique opportunity to explore the integration of data discovery with model optimization. By applying multiple types of European frog-bit occurrence data to the same broader ecological questions (What are the drivers of this species' distribution, and what areas contain suitable habitat?), we were able to examine how complexity in the data fed into models could influence their relative performance and outcomes. Invasive European frog-bit was first introduced to North America in 1932 when it was planted as an ornamental in Ottawa, Ontario, Canada (Minshall, 1940). Now, European frog-bit is primarily concentrated in the Laurentian Great Lakes region, where it has the potential to degrade wetlands by crowding out native plants and reducing light, oxygen, and nutrients in the water (Catling et al., 2003; Zhu et al., 2018). Monitoring efforts in invaded regions have produced occurrence data representing a range of geographic and ecological features. In this study, we utilized an aggregated data set of publicly accessible occurrence records across the invaded North American range (Monfils et al., 2021; Hansen et al., 2022a).

Heterogeneity across these data necessitated data exploration during the modeling process, which allowed us to identify the specific effects of features of the input data on model outcomes. We selected five decision points in the processing (1 and 2) and selection (3, 4, and 5) of input data that captured the variation in our data and could be applicable outside of this species and study area. We asked: How are model results and performance affected by (1) the method of removing missing values in explanatory data prior to modeling? (2) the threshold for excluding correlated explanatory variables prior to modeling? (3) the spatial and temporal scale of occurrence data? (4) the source of occurrence data? and (5) the type of occurrence data? At each decision point, we tested multiple options by building a set of models for each option and comparing model performance when predicting onto test data. At data selection decision points, we also compared model outcomes, specifically the contribution of variables in the final best model and predicted probability of suitable habitat from the final best model. We provide all information related to data access, data processing, model building, model evaluation, and data visualization through the Open Science Framework (Hansen et al., 2023; <https://doi.org/10.17605/OSF.IO/Y6MEQ>) to make this study fully reproducible using R Statistical Software (v4.3.0; R Core Team, 2023).

Decision Point 1: The first decision point we assessed was the method of handling missing data prior to modeling. When preparing data for SDMs, occurrence data are joined to explanatory variables, which are often continuous or near-continuous raster layers (Elith et al., 2006; Valavi et al., 2022). If the raster layers do not perfectly cover the study region of interest, then missing values will be introduced for those variables for some occurrence locations. Because SDMs cannot be built when any data are missing, many modeling algorithms can deal with this problem internally by automatically removing missing observations (rows), imputing missing values, or including default values (Chen and Guestrin, 2016; Tang and Ishwaran, 2017; Wood, 2023). However, if data are modified within the model in this way, the modeler may not be aware of them and will not be able to accurately interpret the final results (e.g., they could infer biological significance for a variable that largely consisted of imputed data that do not represent actual environmental associations). At this decision point, we removed missing data either by observation or by variable prior to modeling to avoid the automatic decision made by each modeling algorithm.

Decision Point 2: The second decision point we assessed was the threshold for excluding correlated explanatory variables. When many spatial explanatory variables are aggregated for SDMs, they are likely to have some correlation with each other due to the influence of geography and topography; including all correlated variables could increase model bias and uncertainty (Dormann et al., 2012; De Marco and Nóbrega, 2018). One way modelers can avoid this problem is by setting a correlation threshold and excluding one of any variable pair with a correlation coefficient above that value

(De Marco and Nóbrega, 2018). At this decision point, we tested two correlation thresholds, 0.7 or 0.5 (out of 1), which we chose based on their prevalence in SDM and other applications and recognition as indicators of “strong” and “moderate” correlation, respectively (Meier et al., 2010; Mukaka, 2012; Zhu and Peterson, 2017; Akoglu, 2018; Ajene et al., 2020; Tapa-Yotto et al., 2021).

Decision Point 3: The third decision point we assessed was the scale of occurrence data, which came about because of the differences between the aggregated and field data available to us. The scale of occurrence data as we define it here is the temporal range of occurrences and the size and location of the sampled area. At this decision point, we sought to compare a larger scale based on aggregated public data (Hansen et al., 2022a) with a smaller scale based on field data (Saginaw Bay; Monfils et al., 2021). By comparing these two occurrence data sets, we were also testing different methodologies for data sampling, one that was largely opportunistic (large scale) and one that was more systematic and generated for a specific research project (Saginaw Bay). The first part of this decision (Decision Point 3a) was defining the larger scale to be used, either the invaded range of North America or the state of Michigan, USA. The second part of this decision (Decision Point 3b) was comparing the selected larger scale to the smaller scale of Saginaw Bay.

Decision Point 4: The fourth decision point we assessed was the source of occurrence data, which came about because of the integration of European frog-bit data across data providers (e.g., natural history collections, targeted research data, observations; Hansen et al., 2022a) in the aggregated data set. At this decision point, we compared models trained on specimen data to models trained on observation data.

Decision Point 5: The fifth decision point we assessed was the type of occurrence data, which came about because of the availability of true absences (or non-detections) in our Saginaw Bay field data. As defined in this study, the type of occurrence data is what species absence points represent. Most SDM algorithms utilize species presences and absences or pseudo-absences, which are drawn from the background of the study region where the actual occurrence of the species is not known (Valavi et al., 2022). At this decision point, we compared models trained on presence-absence data to models trained on presence-background data.

Although we identified these five decision points based on features of the European frog-bit occurrence data available to us, these decisions are not specific to this species or study; the explosion of publicly accessible data makes these questions more relevant as data complexity continues to increase. Scientists building SDMs use species occurrence data from publicly accessible sources (Anderson et al., 2016; Feldman et al., 2021), which include natural history collections, community science platforms (e.g., iNaturalist; <https://www.inaturalist.org/>), regional projects (e.g., the Midwest Invasive Species Information Network [MISIN]; <https://www.misin.msu.edu/>), published research databases (e.g., made available as supplemental data with a scientific paper), and literature (e.g., records extracted from

older papers). Environmental variables can be obtained from continuous or near-continuous raster layers derived from meteorological stations (e.g., WorldClim; Fick and Hijmans, 2017), environmental models (e.g., SoilGrids; Poggio et al., 2021), federal projects (e.g., National Land Cover Database; Dewitz and USGS, 2021), and feature locations for specific regions (e.g., boat launch sites in Michigan; MDNR, 2023). Access to each of these data sources opens up tremendous potential for ecological applications; however, employing data sources with such variability, both in terms of the data themselves and the methods used to collect them (e.g., protocols, instruments, timing, accuracy), can create greater variability in model results and result in potential errors if features of the data and the species they represent are not well understood and accounted for (Moudrý and Devillers, 2020; Tessarolo et al., 2021; Marcer et al., 2022; Zarzo-Arias et al., 2022). Our data-centric modeling protocol embeds data discovery within model building to account for this variability.

METHODS

Study regions and distribution data set

In 2021, we aggregated European frog-bit occurrence data from public and private sources across the invaded range of North America in order to understand the historical distribution of this species and its pattern of spread (Hansen et al., 2022a). Data sources included opportunistic data collected by community scientists, specimens in natural history collections, and observations from targeted field surveys by researchers and natural resource managers. The resulting data set provides an ecologically relevant exemplar of complex data well suited to the questions addressed in the present study. Given the large geographic area represented by this data set, we focused on three specific regions: the invaded range of North America; the state of Michigan, USA; and Saginaw Bay, Michigan. North American data span two Canadian provinces and 10 U.S. states, primarily concentrated around the Laurentian Great Lakes, and represent the historical spread of European frog-bit since 1932 (Hansen et al., 2022a). Michigan was included as a study region because of our knowledge of and involvement in European frog-bit research and monitoring across the state (Cahill and Monfils, 2021). Michigan is located in the formerly glaciated (last covered 9000–10,000 years ago; Krist and Lusch, 2004) upper Midwestern United States between 41°41'N and 48°18' N and 82°7'W and 90°25'W. The climate is temperate and highly seasonal (Köppen climate classification humid continental; Peel et al., 2007), with some areas protected from extremes by the Laurentian Great Lakes (Albert, 1995). Saginaw Bay, which is a bay of Lake Huron in eastern Michigan, was included as a study region because it is an Area of Concern (EPA, 2023) with documented European frog-bit invasion and intensive field sampling efforts (Monfils et al., 2021; Hansen et al., 2022a).

Occurrence data gathering

Our response variable was European frog-bit distribution, as occurrence records in the aggregated data set. For the North American study region, we used the entire aggregated data set, excluding records with uncertainty above 1 km and a small proportion of outliers so that large unsampled regions would not be included in our analysis. For the Michigan study region, we did the same for records within the state. For the Saginaw Bay study region, we extracted field data of European frog-bit presences and absences collected by the authors along 50-m segments during a targeted project in 2020 (Monfils et al., 2021). These specific field data were excluded from the Michigan study region. Each study region was defined as a concave hull around occurrences, buffered by 1 km and clipped to 1 km around the land. Study regions were used for model training, testing, and predictions.

Explanatory data gathering

Thirteen explanatory variables were obtained from seven public sources directly through R and consisted of climate, soil, topographic, and anthropogenic factors with demonstrated or suspected relevance to European frog-bit based on literature on this species and other aquatic plants (Table 1).

Data processing

Occurrence data

Aggregated data in the North America and Michigan study regions were split into three data sets based on data source: specimens and observations combined (“all”), specimens only, and observations only. For each data set in the three study regions, we used the function *gridSample* in the *dismo* package (Hijmans et al., 2023) to randomly sample one occurrence per 1 km × 1 km grid cell to reduce duplication and potential sampling bias occurring from a higher density of observations in some areas (Rodríguez-Castañeda et al., 2012).

Background point sampling

The machine learning and regression models employed in this study (Table 2) accept presence-absence occurrence data, which may be substituted for presence-background data with background points acting as absences. We randomly sampled background points for each data set (no more than one per grid cell) equal in number to the number of presence points to achieve a prevalence of 0.5 and avoid biasing metrics of model performance (Rodríguez-Rey et al., 2019). This prevalence has demonstrated positive outcomes for many modeling applications (e.g., Barbet-Massin et al., 2012; Grimmitt et al., 2020;

Casas et al., 2022). Random sampling of the Saginaw Bay presence-absence data set using *gridSample* returned uneven numbers of presences and absences; we sampled from the more prevalent class to achieve an equal ratio.

Explanatory data

Explanatory variable raster layers were calculated for or cropped to each study region extent. Most layers were available at a resolution of 30 seconds of a degree, or approximately 1 km × 1 km. Phosphorus fertilizer application was available at a resolution of 30 minutes, which we resampled to 30 seconds using bilinear interpolation (Soultan and Safi, 2017). Land cover was available at a resolution of approximately 1 second (30 m × 30 m), which we resampled to 30 seconds using majority rule (Guisan et al., 2007). Data that represented monthly averages (solar radiation, wind speed, and water vapor pressure) were averaged over the year. Across each study region, we checked for correlation among the explanatory variables and excluded variables above each of the two thresholds we tested (0.7 and 0.5). To determine which variables to exclude, we used the function *findCorrelation* in the *caret* package (Kuhn, 2008), which identifies any pairwise correlations between variables above the given threshold and returns the variable to be excluded based on its mean correlation across all variables. We extracted values for each set of uncorrelated variables at occurrence locations and performed statistical standardization to coerce their means to 0 and their standard deviations to 1 (see Hansen et al., 2023).

We chose to actively handle missing values, either by observation or by variable, for explanatory variables that were introduced when joining explanatory and occurrence data. The observation removal method involved excluding all observations (rows) that had a missing value for any explanatory variable. The variable removal method involved excluding all explanatory variables (columns) that had at least one missing value in the given data set.

Model building

For each data set, we built 12 types of models representative of three classes: machine learning, regression, and envelope (see Table 2 for model types). We classified MaxEnt and MaxNet models in the regression class due to their similarity to Poisson point-process regression models (Renner and Warton, 2013). Reviews of SDM methods (e.g., Elith et al., 2006; Valavi et al., 2022) have discussed how different modeling algorithms perform relative to each other on the same input data. Our goal was to explore how a subset of modeling algorithms was differentially affected by the five decision points in this study. We split each input data set into four groups and ran four iterations of each model, using three groups for model training and one for model testing in each run of the model so we could later compare mean performance metrics (*k*-fold cross validation, *k*=4; Fielding and Bell, 1997; Hijmans, 2012).

TABLE 1 Explanatory variables for species distribution modeling and their significance to the species of interest, European frog-bit (*Hydrocharis morsus-ranae*).

Variable type	Variable	Source	Significance to European frog-bit
Climate	Mean temperature of warmest quarter (BIO10)	WorldClim (Fick and Hijmans, 2017)	Turion development and seed germination are dependent on higher water temperature (Richards and Blakemore, 1975; Cook and Lüönd, 1982; Burnham, 1998).
	Mean temperature of coldest quarter (BIO11)		Turions may not germinate when exposed to low temperatures while overwintering (Burnham, 1998; Adamec and Kučerová, 2013).
	Minimum temperature of coldest month (BIO6)		Exposure to frost is likely relevant for floating aquatic plants (Lozano, 2021).
	Precipitation of warmest quarter (BIO18)		Drying out of waterbodies during the growing season may inhibit growth (Catling et al., 2003; Lozano 2021).
	Solar radiation		Rosettes grow faster in full sun (Halpern, 2017), and shading has been shown to be an effective control mechanism (Zhu et al., 2014).
	Wind speed		Wind and waves may facilitate dispersal (Halpern, 2017), but EFB may also inhabit areas protected from wind and wave action (Cook and Lüönd, 1982; Catling et al., 2003).
	Water vapor pressure		Vapor pressure influences surface water evaporation and plant transpiration in wetlands, which could both affect and be affected by aquatic plant growth (Gale, 2004; Kirzhner and Zimmels, 2006).
Soil nutrients	Total nitrogen at surface of soil	SoilGrids (Poggio et al., 2021)	Floating aquatic plants are often limited by nutrients in water; excess nutrients in the soil and subsequently water could contribute to dominance of EFB and other floating plants (Scheffer et al., 2003). Mesotrophic or eutrophic conditions may support EFB populations (Cook and Lüönd, 1982; Catling et al., 2003).
	Phosphorus fertilizer application	Global Gridded Soil Phosphorus (Yang et al., 2014)	
Topography	Land cover	National Land Cover Database 2019 (Dewitz and USGS, 2021)	European frog-bit is often found within and along wetlands (Cook and Lüönd, 1982; Catling et al., 2003).
	Elevation	Shuttle Radar Topographic Mission Digital Elevation Database (Jarvis et al., 2018)	Elevation affects factors related to plant growth (Gale, 2004) and has been significant in other plant species SDMs (Jarnevič and Reynolds, 2011; Koncki and Aronson, 2015).
Anthropogenic	Population density (estimate)	Gridded Population of the World (CIESIN, 2018)	If human-mediated spread is important for EFB's distribution, then it may establish more frequently in areas of higher population density (Rodríguez-Rey et al., 2019).
	Distance from nearest boat launch	Calculated from Michigan Public Boating Access Sites (MDNR, 2023)	Aquatic plants are likely to spread between waterbodies on boats and equipment (Catling et al., 2003; Rodríguez-Rey et al., 2019).

Note: EFB = European frog-bit; SDM = species distribution model.

We built a set of models for all possible combinations of decisions, keeping modeling parameters the same across data sets so we could directly compare outcomes for each specific decision point. For one data set (Saginaw Bay presence-absence), GAM and BRT required adjustments only for the purpose of allowing these models to be built (see Hansen et al., 2023).

Model assessment

To compare the performance of each model, we generated predicted habitat suitability (between 0 and 1) from each trained model and compared it against the actual occurrence value (0 for absence/background or 1 for presence) in the corresponding test data. We generated three metrics to compare

TABLE 2 Species distribution model algorithms used in this study. Machine learning and regression models were built using presence-background data at the Michigan scale and presence-background and presence-absence data at the Saginaw Bay scale. Envelope models were built using presences only. Refer to these abbreviations for all following tables and figures.

Model class	Model name	Abbreviation	References
Machine learning	Artificial neural network	ANN	Pearson et al. (2002)
	Boosted regression trees (or gradient boosting machine)	BRT	Elith et al. (2008); Yu et al. (2020)
	Conditional inference forest	Cforest	Copenhaver-Parry et al. (2016)
	Random forest	RF	Valavi et al. (2021, 2022)
	eXtreme gradient boosting machine	XGBoost	Chen and Guestrin (2016); Muñoz-Mas et al. (2019); Valavi et al. (2022)
Regression	Generalized additive model	GAM	Guisan et al. (2002)
	Generalized linear model	GLM	Guisan et al. (2002)
	Multivariate adaptive regression splines	MARS	Leathwick et al. (2006)
	Maximum entropy	MaxEnt	Phillips et al. (2006, 2017)
	Maximum entropy using glmnet	MaxNet	Phillips et al. (2017)
Envelope	BIOCLIM	BIOCLIM	Busby (1991); Hijmans and Graham (2006); Hijmans et al. (2023)
	DOMAIN	DOMAIN	Carpenter et al. (1993); Hijmans and Graham (2006)

across models: mean absolute error, sensitivity, and specificity (for presence-absence and presence-background models). Mean absolute error (MAE) is a threshold-independent measure of the mean difference between predicted and actual values (Willmott and Matsuura, 2005). Sensitivity is a threshold-dependent measure of the ability of a model to detect true positives (i.e., predict presence where there is a presence). When sensitivity is high, the true positive rate is high and the false negative rate is low. Specificity is a threshold-dependent measure of the ability of a model to detect true negatives (i.e., predict absence where there is an absence [or background point]). When specificity is high, the true negative rate is high and the false positive rate is low (Yerushalmy, 1947). We used a threshold of 0.5 for calculating sensitivity and specificity, which provides an intuitive, commonly used null threshold representing equal chance of presence and absence (Wilkinson et al., 2021). We did not include any area under the curve (AUC) metrics in our analysis, as they are known to be strongly influenced by the size of the study region and thus would not be appropriate for comparing across study regions and would double-count sensitivity and specificity with lower interpretability (Lobo et al., 2008; Jiménez and Soberón, 2020). We include optional code in our protocol for running all hypothesis tests for the area under the receiver operating characteristic curve (AUC ROC; see Hansen et al., 2023). We used Wilcoxon rank-sum (also known as Mann–Whitney U) tests to assess differences in model evaluation metrics between groups for our five decision points. Models sometimes failed to build or fully converge. If more than one model failed in a given group of four, we excluded that group. As each decision was assessed, we controlled for the remaining four by grouping or filtering our results data appropriately.

To compare the outcomes of each model based on the three data selection decisions (scale, data source, and data type), we generated one high-performing model (XGBoost) from the entirety of each data set. We compared the relative contribution of each explanatory variable and predicted probability of suitable habitat across models for each study region.

Comparison of North America and Michigan study regions

Occurrence data within the North America and Michigan study regions were not independent, due to the inclusion of Michigan in both. Our goal was to select just one of these regions (Decision Point 3a) to move forward with assessment of the remaining decision points. Given evidence of both positive and negative effects of geographic partitioning of occurrence data (Osborne and Suárez-Seoane, 2002; Gonzalez et al., 2011; Qiao et al., 2019) and considerations for modeling species occurrences over long time periods (Barve et al., 2011), we wanted to directly compare the two large-scale study regions to make an informed decision about which one was most appropriate for further analysis. We built models using North America and Michigan occurrence data as outlined above, excluding the explanatory variables distance from the nearest boat launch and land cover, which were available at the state and country level, respectively.

We found significant differences in MAE for all models, sensitivity for nine out of 12 models, and specificity for four out of 10 models, with higher performance observed for the Michigan training region (Appendices S1, S2). These findings are similar to those of El-Gabbas and Dormann (2018), who

found that range-wide data did not improve models built using regional data. Reasons for the lower predictive performance of North America models may be the geographic heterogeneity and dispersal barriers (e.g., Lake Huron and Lake Erie) of the North American invaded region compared with Michigan (Barve et al., 2011) and background points that were less likely to represent true absences than the Michigan background points, which we were more confident about given our involvement in targeted European frog-bit monitoring in Michigan specifically.

We mapped predicted probability from final XGBoost models trained with North America or Michigan occurrences onto the Michigan study region. The North America models predicted marginally higher probability of suitable habitat on average (5.5%). The median difference in predictions was <0.1%, due to the fact each model predicted higher probability of suitable habitat in some parts of the Michigan study region (Appendices S3, S4). Because neither model predicted consistently higher values across the entire study region, and prediction values were similar overall (Wilcoxon effect size = 0.214), we concluded that neither model was under- or overestimating probability of suitable habitat by a large enough margin to realistically influence model interpretation and application. We note that for other species and study regions, differences based on the training region will likely not be the same as what we observed in this study. Data-centric criteria for delineating a study region may include relative data availability across a species' range, bias in either presence or absence data, and relative performance and outcomes as we assessed here.

Given these findings, Michigan was selected as the final large-scale study region representing observations over a long time period (1996–2021) with a large area of interest (approximately 15,000 km²) and high variation in environmental gradients (Hansen et al., 2022a). The Saginaw Bay study region represented a single year (2020) of a field research study covering an area of approximately 250 km² of documented European frog-bit occurrence resulting in less variation in environmental gradients (Monfils et al., 2021; Appendix S5). Further analyses at the Michigan and Saginaw Bay scales (Decision Points 1, 2, 3b, 4, and 5) were based on models built from all available occurrence data and explanatory variables. Sample sizes for Michigan “all,” specimen, and observation data were 922, 68, and 912, respectively (Appendix S6). Sample sizes for Saginaw Bay presence-absence and presence-background data were 100 and 200, respectively.

RESULTS

Model performance: Data processing decisions

Decision Point 1: Method of removing missing values in explanatory data prior to modeling

We assessed how the method of removing missing values in explanatory data, either by observation or by variable, prior to modeling influenced model performance. All data sets had

missing values for at least one explanatory variable. We found that the method of removing missing values led to significant differences in MAE in 42 of 89 comparisons (47%), controlling for explanatory variable threshold, scale, data source, and data type. The observation removal method was favored more often than the variable removal method (Table 3, Appendix S7). Significant differences in MAE were not restricted to any particular model class. Envelope models favored the variable removal method, while machine learning and regression models typically favored the observation removal method. We controlled for the method of removing missing values by retaining only the better-performing group in each case, regardless of statistical significance.

Decision Point 2: Threshold for excluding correlated explanatory variables

We assessed how the threshold for excluding correlated explanatory variables influenced model performance, testing the cutoff values of 0.7 and 0.5. At the Saginaw Bay scale, the same variables were retained at both thresholds. At the Michigan scale, we found that the threshold led to significant differences in MAE for presence-only envelope models (Table 4, Appendix S8). We controlled for the correlation threshold by retaining only the better-performing group in each case, regardless of statistical significance.

Model performance: Model type and class

After controlling for the method of removing missing values and threshold for excluding correlated variables, we tested the relative performance of each model type and class. The machine learning and regression classes performed significantly better than the envelope class ($P < 0.001$). We did not observe a significant difference in MAE between the machine learning and regression classes ($P = 0.117$ after Bonferroni correction for multiple comparisons; see Hansen et al., 2023). Pairwise comparisons between machine learning models were significant in two out of 10 cases; pairwise comparisons between regression models were significant in six out of 10 cases (Figure 1). The regression model GAM performed similarly to four out of five machine learning models, likely explaining the absence of a significant difference in MAE between these classes. The machine learning and regression model classes were not significantly different in terms of sensitivity or specificity (Figure 2).

Model performance: Data selection decisions

Decision Point 3a: Selection of large-scale study region

Occurrence data within the North America and Michigan study regions were not independent, so only one of these regions could be included in analyses of Decision Points 1, 2, 3b, 4, and 5. We selected Michigan as our large-scale study region after observing that Michigan models outperformed North America models (Appendices S1, S2).

TABLE 3 Comparisons of the method of removing missing data prior to modeling (by observation or by variable), grouped by the threshold for excluding correlated explanatory variables and scale, data source, and data type. Saginaw Bay data sets were equivalent at the 0.7 and 0.5 correlation thresholds. The method with the lower mean absolute error is given for each group; shaded values are statistically significant ($P < 0.05$). Refer to Table 2 for model abbreviations.

Presence-absence or presence-background models									
		Michigan						Saginaw Bay	
		0.7 Correlation threshold			0.5 Correlation threshold			Both correlation thresholds	
		Presence-background			Presence-background			All data (observations)	
Model class	Model type	All data	Specimen	Observation	All data	Specimen	Observation	Presence-background	Presence-absence
Machine learning	ANN	NA	Var	Obs	Obs	Var	NA	NA	Obs
	BRT	Obs	Obs	Obs	Obs	NA	Obs	Obs	NA
	Cforest	Obs	Obs	Obs	Obs	Obs	Obs	Obs	Var
	RF	Obs	Var	Obs	Obs	Var	Obs	Obs	Var
	XGBoost	Obs	Var	Obs	Obs	Var	Obs	Obs	Var
Regression	GAM	Obs	NA	Obs	Obs	Var	Obs	Var	NA
	GLM	Obs	Var	Obs	Obs	Var	Obs	Obs	Var
	MARS	Obs	Var	Obs	Obs	Obs	Obs	Obs	Obs
	MaxEnt	Obs	Var	Obs	Obs	Var	Obs	Obs	Var
	MaxNet	Obs	Var	Obs	Obs	Var	Obs	Obs	Obs

Presence-only models									
		Michigan						Saginaw Bay	
		0.7 Correlation threshold			0.5 Correlation threshold			Both correlation thresholds	
		All data	Specimen	Observation	All data	Specimen	Observation	All data (observations)	
Envelope	BIOCLIM	Var	Obs	Var	Var	Var	Var	Var	
	DOMAIN	Var	Obs	Var	Var	Var	Var	Var	

Note: Var = exclusion of explanatory variables (columns) with any missing values; Obs = exclusion of observation (rows) with any missing values; NA = fewer than three models (out of four) could be built in a given group and a comparison could not be made.

Decision Point 3b: Scale of occurrence data

We assessed how the scale of occurrence data, either Michigan or Saginaw Bay, influenced model performance. We used all available data at each scale and included only presence-background or presence-only models in our analysis, which controlled for data source and data type. We found significant differences in MAE for 11 out of 12 models, sensitivity for two out of 12 models, and specificity for all models, with higher performance observed at the Michigan scale (Figure 3, Appendix S9).

Decision Point 4: Source of occurrence data (Michigan scale only)

We assessed how the source of occurrence data, either specimens or observations, influenced model performance. These analyses could only be carried out for the Michigan scale using presence-background or presence-only data, which controlled for scale and data type. We found significant differences in MAE for seven out of 12 models, sensitivity for three out of 12 models, and sensitivity for two out of 10 models, with higher performance observed for observation-based data (Figure 4, Appendix S10).

Decision Point 5: Type of occurrence data (Saginaw Bay scale only)

We assessed how the type of occurrence data, either presence-background or presence-absence, influenced model performance. These analyses could only be carried out for the Saginaw Bay scale, which controlled for scale and data source. We found significant differences for MAE for four out of 10 models, sensitivity for three out of 10 models, and specificity for two out of 10 models, with higher performance observed for presence-background data (Figure 5, Appendix S11).

Model outcomes: Data selection decisions

Contribution of each variable to the final model

We assessed how the relative contribution of each explanatory variable was affected by scale, data source, and data type in one high-performing model, XGBoost. Distance from the nearest boat launch contributed greater than 5% across all models. Elevation and phosphorus fertilizer application contributed greater than 41% and 9%,

TABLE 4 Comparison of each threshold (0.5 or 0.7) for excluding correlated explanatory variables prior to modeling, grouped by data source. Saginaw Bay data sets were equivalent at each threshold; table refers to Michigan scale only. Machine learning and regression models utilized presence-background data and envelope models utilized presence-only data. The threshold with the lower mean absolute error is given for each group; shaded values are statistically significant ($P < 0.05$). Explanatory variables with correlations above each threshold were excluded prior to modeling using *caret::findCorrelation* (Kuhn, 2008). Refer to Table 2 for model abbreviations.

Model class	Model type	All data	Specimen	Observation
Machine learning	ANN	0.5	0.5	0.5
	BRT	0.7	0.7	0.7
	Cforest	0.5	0.5	0.5
	RF	0.7	0.5	0.7
	XGBoost	0.7	0.7	0.7
Regression	GAM	0.7	NA	0.7
	GLM	0.7	0.7	0.5
	MARS	0.7	0.7	0.7
	MaxEnt	0.7	0.7	0.7
	MaxNet	0.7	0.7	0.7
Envelope	BIOCLIM	0.5	0.5	0.5
	DOMAIN	0.5	0.5	0.5

Note: NA = fewer than three models (out of four) could be built in a given group and a comparison could not be made.

respectively, in all Michigan models. Population density and WorldClim BIO10 (mean temperature of warmest quarter) contributed greater than 24% and 8%, respectively, in both Saginaw Bay models (Figure 6).

Predictive mapping of probability of suitable habitat

We mapped predicted probability of suitable habitat from each Michigan and Saginaw Bay XGBoost model across their respective study regions (Appendices S12–S15). We found average differences between 2% and 12% when comparing between data sources (specimens, observations, or both) at the Michigan scale and an average difference of 25% when comparing between data types (presence-absence or presence-background) at the Saginaw Bay scale.

DISCUSSION

Data processing decisions

Decision Point 1: Removing missing values before modeling retained interpretability

Significant differences in MAE based on the method for removing missing values were found most often for the

largest data sets (Michigan all data, $n = 922$; Michigan observation data, $n = 912$) and usually favored the observation removal method. This is likely explained by the general principle that models with more variables have greater fit to the input data; however, with greater dependence on the input data, these models may be less accurate outside the input data set (Coelho et al., 2019). Another possible explanation is the distribution of non-missing explanatory variables for observations with missing values. For example, the land cover class “Open Water” was more prevalent among observations with missing values. Because open water does not represent “ideal” habitat for European frog-bit (Cook and Löönd, 1982; Catling et al., 2003), models including those observations could be less biologically relevant. While land cover in our models was resampled to a coarser resolution, thus losing some of its heterogeneity, the presence of land cover as a significant contributor in the Saginaw Bay presence-background model suggests this variable was still meaningful.

Identifying missing values and deciding how to handle them before modeling gives the modeler explicit control over which data are retained and allows them to interpret model results without losing meaning when model algorithms delete or impute data automatically. Considerations when choosing how to address missing values may include retaining a minimum sample size and minimum ratio of observations to explanatory variables, the distribution of variables with and without missing values relative to each other, and the biological significance of each variable if this is known.

Decision Point 2: Both exclusionary thresholds retained biologically significant variables

Significant differences in MAE based on the threshold for excluding correlated variables at the Michigan scale were found for envelope models but not machine learning or regression models. Variables excluded at the stricter threshold were solar radiation, WorldClim BIO18, and wind speed. None of these variables contributed significantly to final Michigan XGBoost models trained on all data or observation data, which likely explains the similar performance at each threshold. While these variables did contribute to the final Michigan specimen XGBoost models, their collective contribution was approximately 36%, which may not have been high enough to significantly alter model performance in their absence.

It is likely given these findings that the subset of variables that are actually the greatest contributors to European frog-bit occurrence were retained for both correlation thresholds, leading to similar results; however, this will not be the case for all species and SDM applications. The choice of correlation threshold is up to the modeler and their goals. In some cases, such as when SDM predictions will be extrapolated outside of the training region, a stricter threshold may be preferable for producing more parsimonious models more likely to retain predictive power outside the training data (Daganzo et al., 2012; Petitpierre et al., 2017; Gori et al., 2024). Another important consideration when excluding or combining explanatory

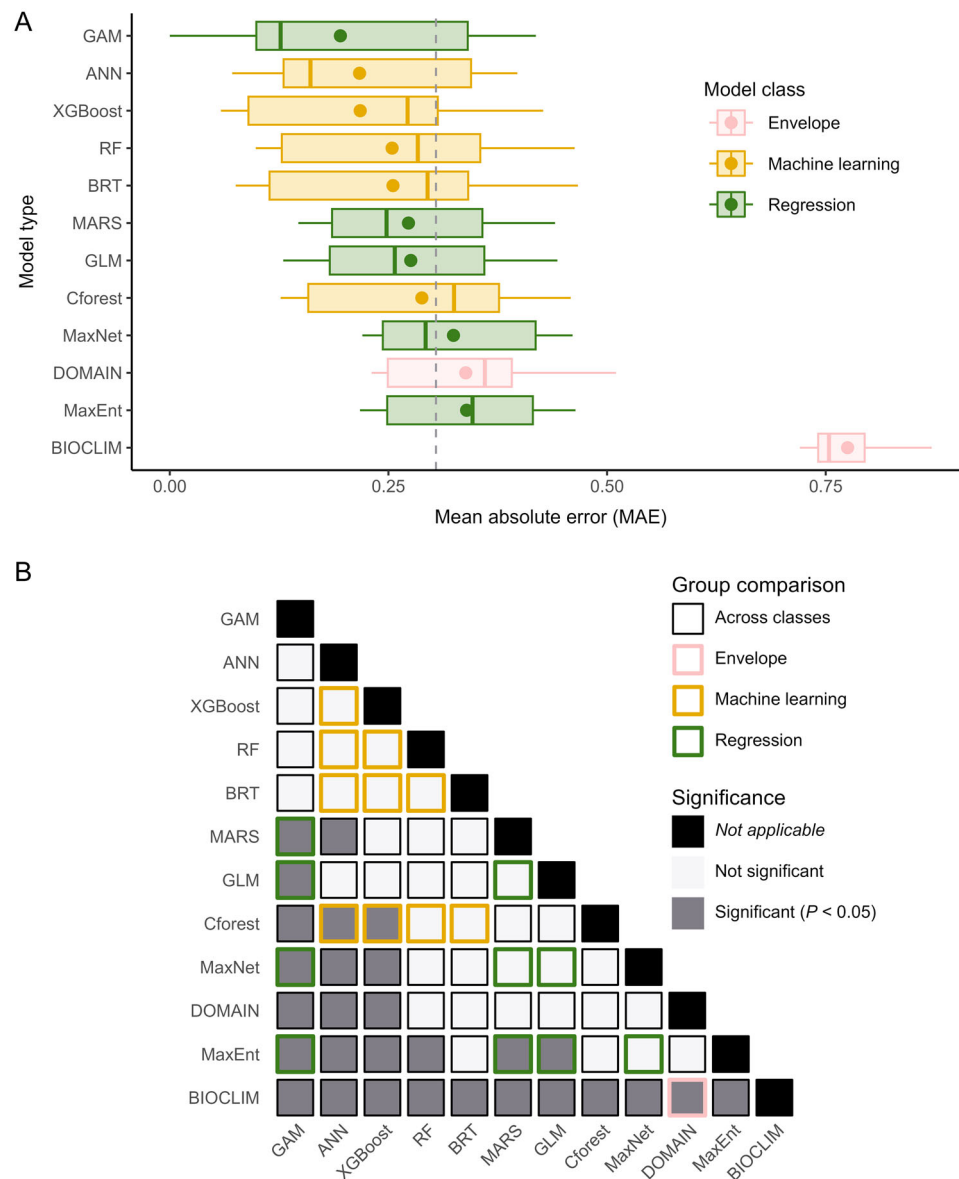


FIGURE 1 Comparison of mean absolute error (MAE) by model type and class, across all scales, data sources, and data types. (A) Mean and median MAE for each model type. (B) Pairwise comparisons of MAE for each model type. In each boxplot, the vertical line indicates the median and the dot indicates the mean; the dashed vertical gray line indicates mean MAE across all models. Refer to Table 2 for model abbreviations.

variables is their biological significance to the species being studied. Statistical selection of variables may be complemented by reviews of literature and expert opinion about the species' biology (Zarzo-Arias et al., 2022).

Data selection decisions

Decision Point 3a: Regional models demonstrated high predictive performance

Possible explanations for the lower predictive performance of range-wide North America models compared with regional Michigan models are geographic heterogeneity, dispersal barriers, and background points that may not represent true

absences in the North American invaded region. A data-centric method for selecting training regions that incorporates geography across a species' known range, data availability, data bias, and model outcomes can help ensure the chosen region is appropriate for the species being studied.

Decision Point 3b: Scale is more meaningful when selected based on the biological question

Significant differences in model performance metrics based on the scale of occurrence data indicated higher performance for Michigan models compared with Saginaw Bay models. This finding is most likely explained by the larger geographic area of the Michigan study region, which created greater geographic separation between presence and background points and

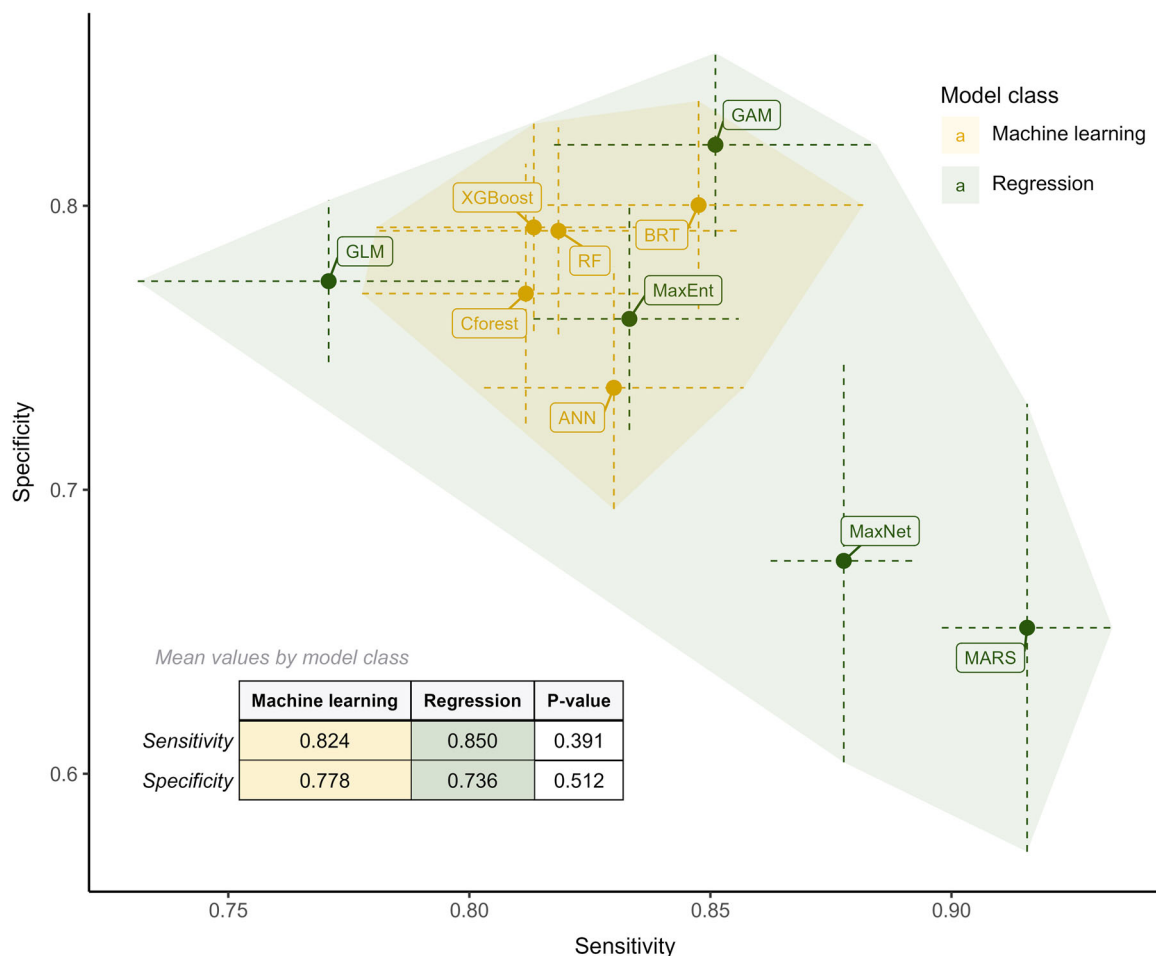


FIGURE 2 Sensitivity (true positive rate) and specificity (true negative rate) for each machine learning and regression model across all scales, data sources, and data types. Refer to Table 2 for model abbreviations.

greater variation in explanatory variables than in the Saginaw Bay study region. This explanation is supported by the equal performance of Michigan and Saginaw Bay models across most model types and metrics when the Michigan study region was redefined to be the same as the Saginaw Bay study region (see Hansen et al., 2023).

Biologically informed SDM building would base the selected scale on the goals of the model, the questions being asked, and the study species. Large-scale opportunistic data outperformed small-scale field data only when the size of the study region also differed, giving the large-scale models an advantage in separating presence and background points. Both opportunistic aggregated data and systematic field data have the ability to build high-performing models, so it is up to the modeler to make an active choice in defining the scale and nature of data to be applied to avoid overinterpretation.

Decision Point 4: Multiple sources of publicly accessible data built strong models

Significant differences in model performance metrics based on the source of species occurrence data at the Michigan scale indicated higher performance for models built with observation

data compared with specimen data. However, many models were not affected by the source of occurrence data. Out of the high-performing models ANN, GAM, and XGBoost, none differed in sensitivity and specificity based on data source and only XGBoost differed in MAE. The success of these models when trained on only specimen data despite a much smaller sample size than observation data ($n=68$ and $n=912$, respectively) is most likely explained by the strength of these modeling algorithms and attributes of the specific specimen data utilized. Based on their geographic distribution, they can be considered a representative sample of observations (Appendix S6). The largest contributor of specimen data is a known expert in the field who was actively monitoring several areas of the state and creating specimens at regular intervals (Hansen et al., 2022a), so their specimens are likely accurate representations of European frog-bit occurrence in recent years.

In our final XGBoost models for the Michigan scale, the average predicted probability of suitable habitat was 12% higher in specimen-only models than in observation-only models. Although the difference in predicted values varied across the study region, specimen-only models tended to predict higher probabilities inland (Appendices S12, S13),

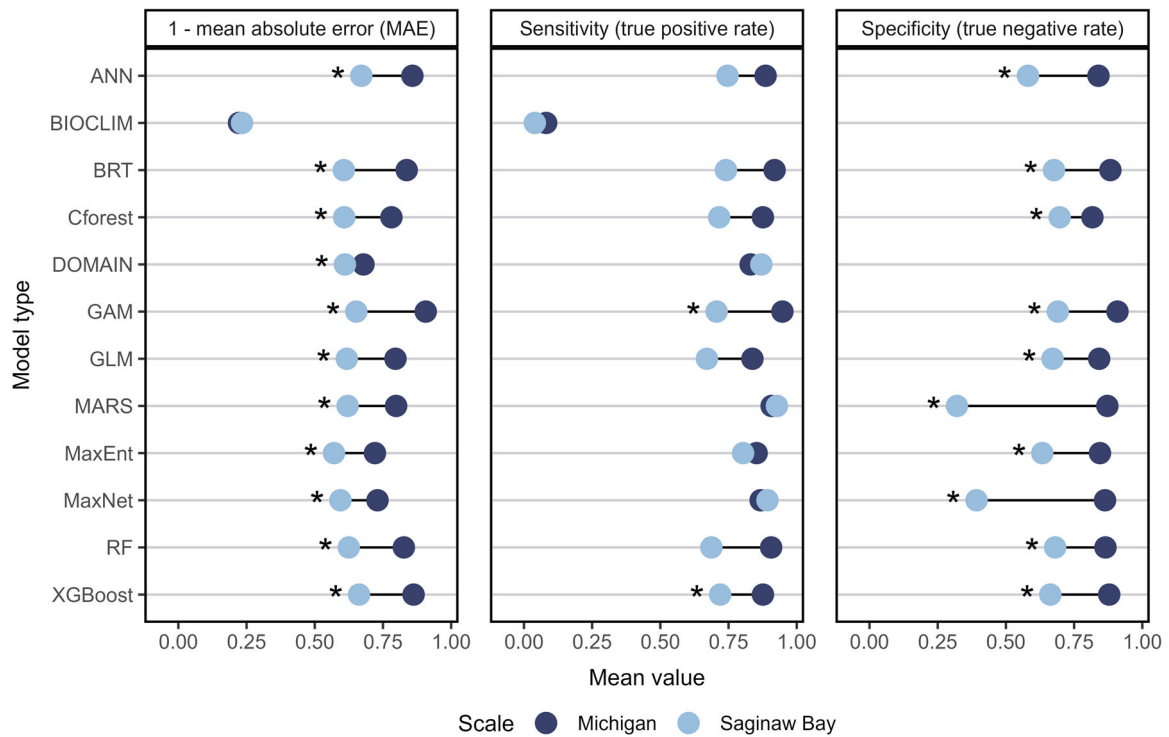


FIGURE 3 Mean values by model type for the inverse of mean absolute error, sensitivity, and specificity at the Michigan and Saginaw Bay scales, using all available presence-background or presence-only data. Statistical significance ($P < 0.05$) is indicated by asterisks. Refer to Table 2 for model abbreviations.

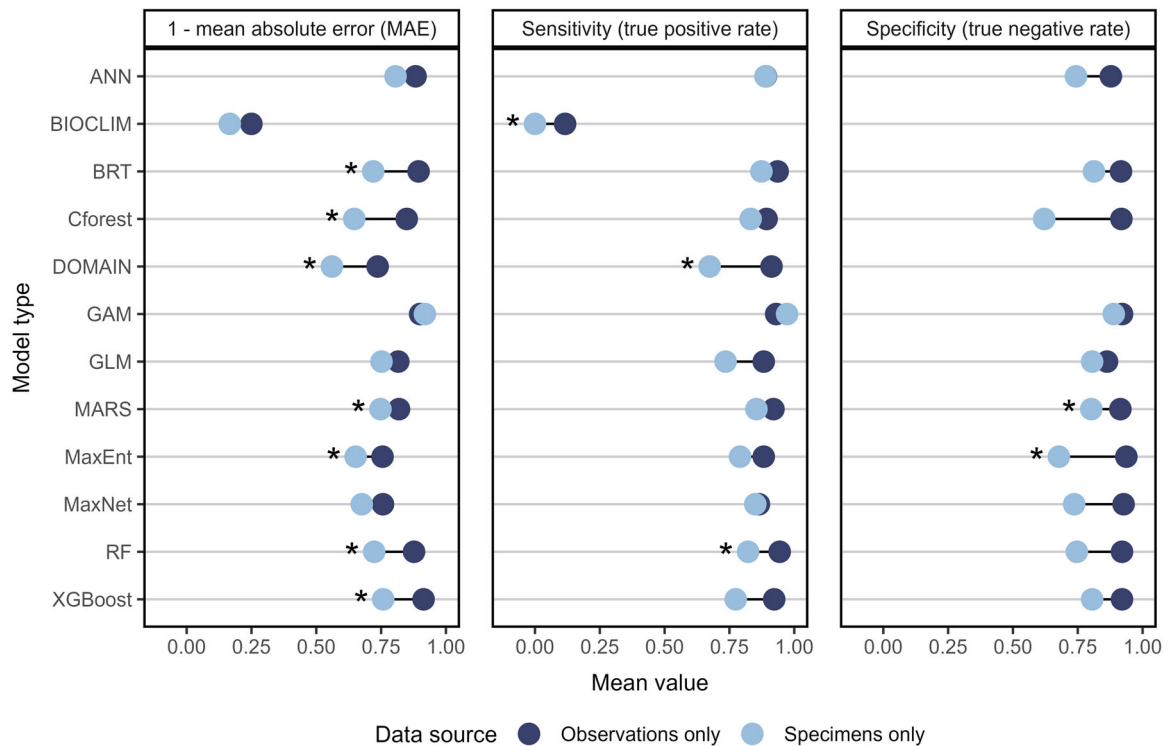


FIGURE 4 Mean values by model type for the inverse of mean absolute error, sensitivity, and specificity using observation-based and specimen-based data at the Michigan scale. Statistical significance ($P < 0.05$) is indicated by asterisks. Refer to Table 2 for model abbreviations.

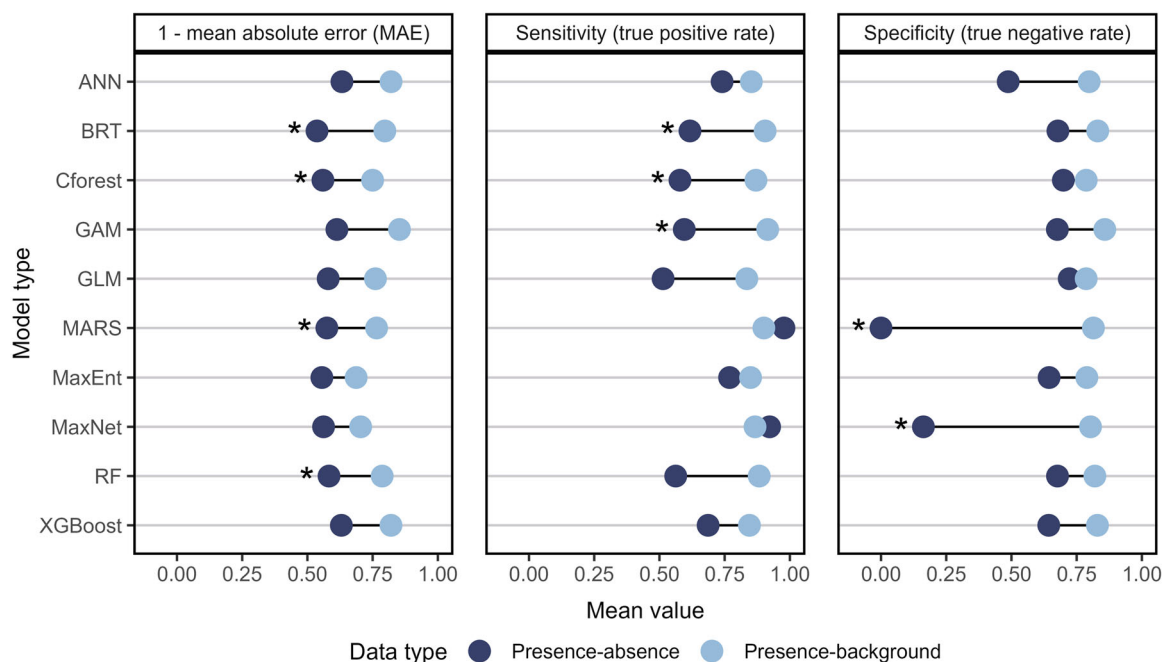


FIGURE 5 Mean values by model type for the inverse of mean absolute error, sensitivity, and specificity using presence-background and presence-absence data at the Saginaw Bay scale. Statistical significance ($P < 0.05$) is indicated by asterisks. Refer to Table 2 for model abbreviations.

which would likely influence decisions made based on the prediction maps. Specimen-only models also returned more explanatory variables contributing more than 5% (Figure 6).

In our study, specimens from natural history collections performed as well as publicly accessible observation records for several model types, including high-performing models. This may be true for other species and study regions, as long as attributes of the specimens (e.g., collectors, temporal range, spatial range) are well understood and compatible with the goals of the SDM. Regional collections in particular may prove valuable given the expertise of collectors contributing to them (Monfils et al., 2020). However, we note that different data sources will lead to different model predictions in some parts of the study region and subsequent conservation decisions. Considering both the availability and appropriateness of each data source prior to modeling will allow for clearer interpretations of models.

Decision Point 5: Presence-background models performed well when background points were drawn from likely absences

Significant differences in model performance metrics based on the type of occurrence data at the Saginaw Bay scale indicated higher performance for models built with presence-background data compared with presence-absence data. This finding could be explained by the larger geographic range of background points compared with absence points, creating clearer separation and greater variability in explanatory variables in the presence-background models than the presence-absence models, which included only points surveyed at or near the coast.

In our final XGBoost models for the Saginaw Bay scale, the average predicted probability of suitable habitat was 25%

greater in the presence-absence model than the presence-background model, with consistently higher predictions in many parts of the study region by the presence-absence model (Appendices S14, S15). The presence-background model indicated more variables contributing greater than 5% (Figure 6), likely because of the expanded range of background points compared with absence points.

In our study, we had access to presence-absence data and were able to derive background data to compare true absences and pseudo-absences, and we found that presence-absence data never performed better than presence-background data. Many SDM studies have pointed out that background points are imperfect substitutes for absences (Barbet-Massin et al., 2012; Guillera-Aroita et al., 2015). However, background points in this study were drawn from areas where European frog-bit was likely absent, based on expert opinion. If this is the case, then using background points allows modelers to extend the range of “true” absences in their models, likely leading to stronger models. When appropriate, the use of background points can also decrease research costs by not requiring sampling to confirm the absence of the species being modeled.

Performance based on model class and type

Our assessment of model performance based on model class and type aligned with findings in similar studies (Elith et al., 2006; Konowalik and Nosol, 2021), with machine learning and regression models outperforming presence-only envelope models. Top performers included the regression model GAM and the machine learning models ANN and

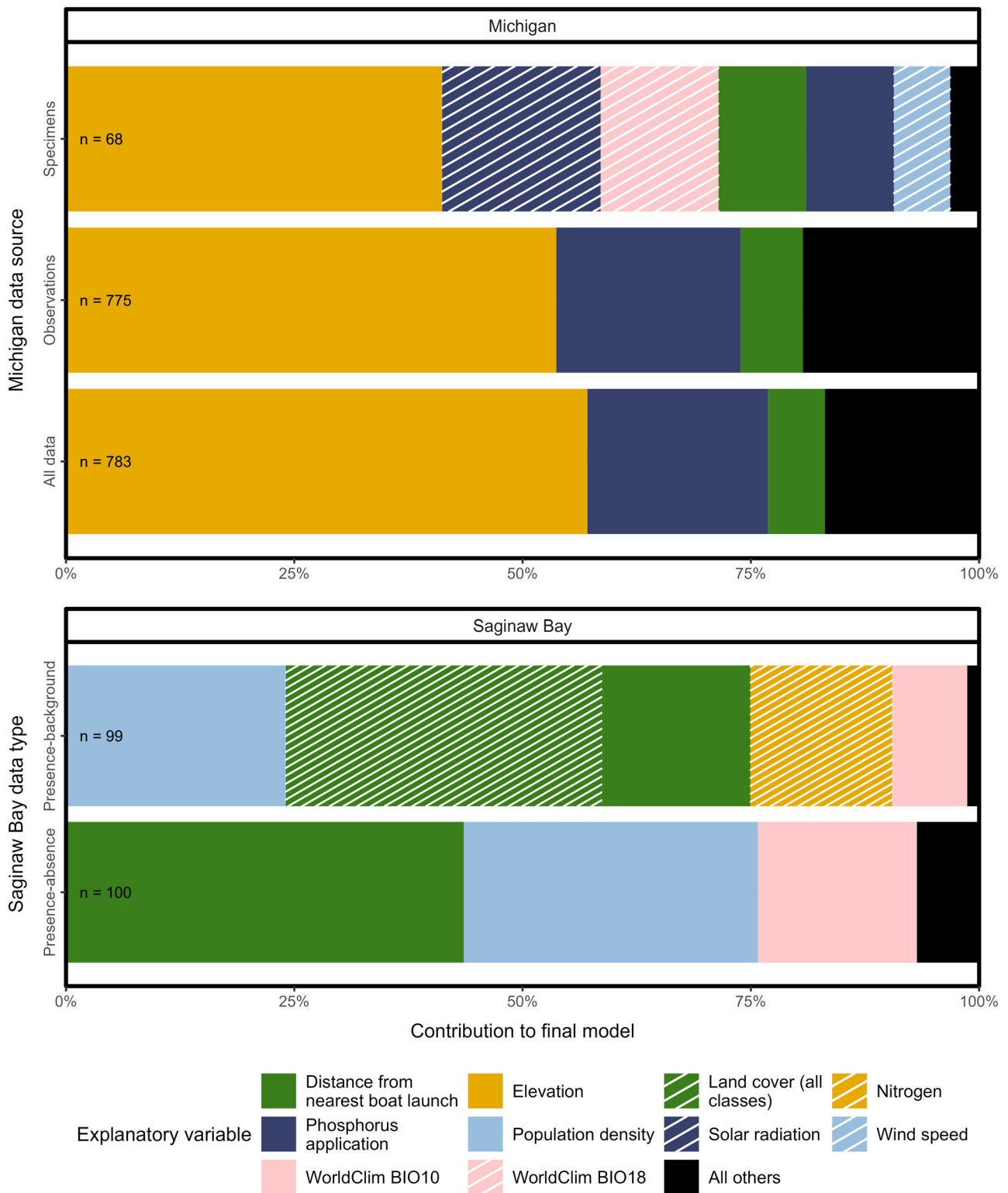


FIGURE 6 Contribution of each explanatory variable (as a percentage) in the final eXtreme gradient boosting machine (XGBoost) models for each scale, data source (Michigan), and data type (Saginaw Bay). Variables that contributed less than 5% in a given model are grouped into the “All others” category.

XGBoost. The effects of model tuning have been addressed in previous studies (e.g., Phillips and Dudík, 2008; Radosavljevic and Anderson, 2014; Vignali et al., 2020) and were outside the scope of this work, which was focused on relative performance

of input data given equal modeling parameters. We found that different modeling algorithms were more or less affected by our five decision points, but even high-performing models differed in their performance based on some data attributes.

In our study, machine learning models were among the best-performing algorithms. However, we note that many machine learning models will perform well regardless of the input data, and biological meaning cannot be inferred from data selected without a strong biological basis. Data-centric criteria for selecting a model type include how likely it is to be affected by modeling decisions and whether it is possible to optimize the model based on a particular data selection or processing decision.

Limitations of invasive species distribution modeling

Critical evaluations of invasive species distribution modeling have noted limitations that arise from non-equilibrium and range shifts when species move into an area they are not native to, which leads to a smaller realized distribution compared to the species' potential distribution (Václavík and Meentemeyer, 2012; Pili et al., 2020; Cho et al., 2022). These limitations may be more relevant for but are not exclusive to invasive species, nor do they necessarily reduce the performance capacity of invasive species models compared to native species models (Menuz et al., 2015). Given the potential for invasive species to significantly harm ecosystems, SDMs are still a useful, widely accepted tool for assessing current distributions (Dutra Silva et al., 2021; Cho et al., 2022). We were interested in assessing relative, not absolute, model performance and outcomes based on how data are selected and processed, which is not unique to any one species. Because of its ecological significance and the heterogeneity of available occurrence data, European frog-bit provided a useful case study that represents the real-world use of SDMs and the influence of data complexity on model results. If high-priority invasive species lead to more data being generated (e.g., from targeted monitoring efforts and community science projects), then there is an even greater need for special attention to their data characteristics during modeling. We aimed to reduce uncertainty and improve the accuracy of our models by following guidance from other modelers given known strengths and limitations of invasive SDMs. Václavík and Meentemeyer (2012) note extrapolation outside of the training region as a particular concern for invasive species models. We addressed this by avoiding extrapolation and focused on training, testing, and predicting onto the same study region. The predictive capacity of invasive SDMs can be improved by using anthropogenic explanatory variables in addition to climate variables (Beans et al., 2012; Rodríguez-Rey et al., 2019) and incorporating measures of dispersal pressure (Menuz et al., 2015). Population density and distance from the nearest boat launch fulfilled these roles in our models.

Conclusions

We demonstrated that conscious choices made by modelers have the power to significantly change the performance and

outcomes of SDMs. Our findings provide support for the importance of applying data-centric modeling protocols that include data discovery to improve model fit, interpretation of results, and the conservation decisions based on them. The results of our European frog-bit models in the Laurentian Great Lakes region serve as a case study of the interaction of data complexity with model results, which is relevant for any modeling application utilizing heterogeneous data (e.g., from multiple data sources, collected using multiple methodologies, available at multiple scales). We found that features of input data and how they are processed, including the spatial and temporal scale of occurrence data, source of occurrence data, type of occurrence data, method of handling missing values, and threshold for excluding correlated variables are just as important (or more) than how data are fit into a model. Even high-performing modeling algorithms differed in their performance based on these data attributes. Integrating data discovery into the modeling process can help modelers understand what attributes are present in their data that could influence model outcomes and how these attributes can be adjusted and accounted for to optimize model performance and biological relevance. Additionally, reporting these decisions along with SDM results can increase the transparency and reproducibility of their models.

The general principle that decisions made by modelers influence how models are interpreted and applied reaches beyond SDMs. An advantage of SDMs is the ability to integrate data collected at different time points using different sampling methodologies, but other ecological questions require data collected at the time a species is observed. Emerging data resources such as (Digital) Extended Specimens, which capture a snapshot of an organism's ecological context through linkages with other data types, will likely prove useful for this reason (Webster, 2017; Lendemer et al., 2020; Hardisty et al., 2022). By acting as digital proxies for physical specimens, they can also be augmented with new information as the specimens are resampled (Pyke and Ehrlich, 2010). Aggregated occurrence data, Digital Extended Specimens, and data resources that have yet to emerge will undoubtedly make data availability less of a constraint while introducing further complexity into the process of ecological modeling. Data-centric ecological modelers will welcome new sources of biodiversity data as they become available, while recognizing and harnessing variation in these data to answer new biological questions. Modeling endeavors will be enriched by careful attention to the strengths, limitations, and potentially confounding features of these data and a commitment to a continual revision of our methodologies.

AUTHOR CONTRIBUTIONS

This project was conceptualized by S.E.H., A.K.M., and M.J.M. Data collection was conducted by R.A.H. and R.T.G. Data curation was conducted by R.A.H., R.T.G., and S.E.H. Data analysis and visualization were conducted by S.E.H., with assistance from A.K.M., M.J.M., and R.A.H.

Manuscript preparation was initiated by S.E.H. with contributions from all authors; all authors approved the final version of the manuscript.

ACKNOWLEDGMENTS

The authors thank the reviewers and editors for providing their time and expertise in reviewing this manuscript. Support for this project was provided by the Earth and Ecosystem Science PhD program at Central Michigan University and the United States Environmental Protection Agency Great Lakes Restoration Initiative through the Michigan Department of Environment, Great Lakes, and Energy (award numbers GLOOE02463 and F18AP00846). Additional funding was provided by the U.S. National Science Foundation (award number DBI-1730526) to develop educational materials for biodiversity data literacy (https://qubeshub.org/community/groups/blue_data/blueresources). This is contribution 196 of the Central Michigan University Institute for Great Lakes Research.

OPEN RESEARCH BADGES



This article has earned an Open Data badge for making publicly available the digitally shareable data necessary to reproduce the reported results. The data are available at <https://doi.org/10.17605/OSF.IO/Y6MEQ>.

DATA AVAILABILITY STATEMENT

European frog-bit occurrence data used in this study are available through the Global Biodiversity Information Facility (<https://doi.org/10.15468/jsab2y>; Hansen et al., 2022a, 2022b). All R code for data access, analysis, and visualization are included in the Open Science Framework repository (<https://doi.org/10.17605/OSF.IO/Y6MEQ>; Hansen et al., 2023).

ORCID

Sara E. Hansen <http://orcid.org/0000-0003-3397-3573>
 Rachel A. Hackett <http://orcid.org/0000-0002-1075-3693>
 Ryan T. Goebel <http://orcid.org/0000-0003-1471-850X>
 Anna K. Monfils <http://orcid.org/0000-0003-3553-6430>

REFERENCES

- Adamec, L., and A. Kučerová. 2013. Overwintering temperatures affect freezing temperatures of turions of aquatic plants. *Flora—Morphology, Distribution, Functional Ecology of Plants* 208: 497–501. <https://doi.org/10.1016/j.flora.2013.07.009>
- Ajene, I. J., F. Khamis, B. van Asch, G. Pietersen, B. A. Rasowo, S. Ekesi, and S. Mohammed. 2020. Habitat suitability and distribution potential of *Liberibacter* species (“*Candidatus Liberibacter asiaticus*” and “*Candidatus Liberibacter africanus*”) associated with citrus greening disease. *Diversity and Distributions* 26: 575–588. <https://doi.org/10.1111/ddi.13051>
- Akoglu, H. 2018. User's guide to correlation coefficients. *Turkish Journal of Emergency Medicine* 18: 91–93. <https://doi.org/10.1016/j.tjem.2018.08.001>
- Albert, D. A. 1995. Regional landscape ecosystems of Michigan, Minnesota and Wisconsin: A working map and classification. North Central Forest Experiment Station General Technical Report NC178. United States Forest Service, St. Paul, Minnesota, USA.
- Anderson, R. P., M. Araújo, A. Guisan, J. M. Lobo, E. Martínez-Meyer, A. T. Peterson, and J. Soberón. 2016. Final report of the Task Group on GBIF Data Fitness for Use in Distribution Modelling. Global Biodiversity Information Facility, Copenhagen, Denmark.
- Austin, M. P., and K. P. Van Niel. 2011. Improving species distribution models for climate change studies: Variable selection and scale. *Journal of Biogeography* 38: 1–8. <https://doi.org/10.1111/j.1365-2699.2010.02416.x>
- Barbet-Massin, M., F. Jiguet, C. H. Albert, and W. Thuiller. 2012. Selecting pseudo-absences for species distribution models: How, where and how many? *Methods in Ecology and Evolution* 3: 327–338. <https://doi.org/10.1111/j.2041-210X.2011.00172.x>
- Barbet-Massin, M., Q. Rome, C. Villemant, and F. Courchamp. 2018. Can species distribution models really predict the expansion of invasive species? *PLoS ONE* 13: e0193085. <https://doi.org/10.1371/journal.pone.0193085>
- Barve, N., V. Barve, A. Jiménez-Valverde, A. Lira-Noriega, S. P. Maher, A. T. Peterson, J. Soberón, and F. Villalobos. 2011. The crucial role of the accessible area in ecological niche modeling and species distribution modeling. *Ecological Modelling* 222: 1810–1819. <https://doi.org/10.1016/j.ecolmodel.2011.02.011>
- Beans, C. M., F. F. Kilkenny, and L. F. Galloway. 2012. Climate suitability and human influences combined explain the range expansion of an invasive horticultural plant. *Biological Invasions* 14: 2067–2078. <https://doi.org/10.1007/s10530-012-0214-0>
- Belitz, M. W., M. J. Monfils, D. L. Cuthrell, and A. K. Monfils. 2019. Landscape-level environmental stressors contributing to the decline of Poweshiek skipperling (*Oarisma poweshiek*). *Insect Conservation and Diversity* 13: 187–200. <https://doi.org/10.1111/icad.12399>
- Bond, N., J. Thomson, P. Reich, and J. Stein. 2011. Using species distribution models to infer potential climate change-induced range shifts of freshwater fish in south-eastern Australia. *Marine and Freshwater Research* 62: 1043–1061. <https://doi.org/10.1071/MF10286>
- Borzée, A., D. Andersen, J. Groffen, H.-T. Kim, Y. Bae, and Y. Jang. 2019. Climate change-based models predict range shifts in the distribution of the only Asian plethodontid salamander: *Karsenia koreana*. *Scientific Reports* 9: 11838. <https://doi.org/10.1038/s41598-019-48310-1>
- Burnham, J. 1998. The contribution of seeds and turions towards population growth and persistence of *Hydrocharis morsus-ranae* L. Master's thesis, University of Guelph, Guelph, Ontario, Canada. Website: <https://hdl.handle.net/10214/20162>
- Busby, J. R. 1991. BIOCLIM – a bioclimate analysis and prediction system. *Plant Protection Quarterly* 6: 8–9.
- Cahill, B. C., and A. K. Monfils. 2021. Michigan's European frog-bit (*Hydrocharis morsus-ranae* L.) adaptive management framework strategic plan. Technical report. Central Michigan University, Michigan, USA. <https://doi.org/10.13140/RG.2.2.16640.35849>
- Carpenter, G., A. N. Gillison, and J. Winter. 1993. DOMAIN: A flexible modelling procedure for mapping potential distributions of plants and animals. *Biodiversity and Conservation* 2: 667–680. <https://doi.org/10.1007/BF00051966>
- Casas, E., L. Martín-García, P. Hernández-Leal, and M. Arbelo. 2022. Species distribution models at regional scale: *Cymodocea nodosa* seagrasses. *Remote Sensing* 14: 4334. <https://doi.org/10.3390/rs14174334>
- Catling, P. M., G. Mitrow, E. Haber, U. Posluszny, and W. A. Charlton. 2003. The biology of Canadian weeds. 124. *Hydrocharis morsus-ranae* L. *Canadian Journal of Plant Science* 83: 1001–1016. <https://doi.org/10.4141/P02-033>
- Center for International Earth Science Information Network (CIESIN). 2018. Gridded Population of the World (version 4.11, December 2018). NASA Socioeconomic Data and Applications Center (SEDAC) data release. <https://doi.org/10.7927/H4JW8BX5>
- Charbonnel, A., P. Lambert, G. Lassalle, E. Quinton, A. Guisan, L. Mas, G. Paquignon, et al. 2023. Developing species distribution models for critically endangered species using participatory data: The European sturgeon marine habitat suitability. *Estuarine, Coastal and Shelf Science* 280: 108136. <https://doi.org/10.1016/j.ecss.2022.108136>

- Chen, T., and C. Guestrin. 2016. XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 2016: 785–794. <https://doi.org/10.1145/2939672.2939785>
- Cho, K. H., J.-S. Park, J. H. Kim, Y. S. Kwon, and D.-H. Lee. 2022. Modeling the distribution of invasive species (*Ambrosia* spp.) using regression kriging and Maxent. *Frontiers in Ecology and Evolution* 10: 1036816. <https://doi.org/10.3389/fevo.2022.1036816>
- Coelho, M. T. P., J. A. Diniz-Filho, and T. F. Rangel. 2019. A parsimonious view of the parsimony principle in ecology and evolution. *Ecography* 42: 968–976. <https://doi.org/10.1111/ecog.04228>
- Cook, C. K., and R. Löönd. 1982. A revision of the genus *Hydrocharis* (Hydrocharitaceae). *Aquatic Botany* 14: 177–204. [https://doi.org/10.1016/0304-3770\(82\)90097-3](https://doi.org/10.1016/0304-3770(82)90097-3)
- Copenhaver-Parry, P. E., S. E. Albeke, and D. B. Tinker. 2016. Do community-level models account for the effects of biotic interactions? A comparison of community-level and species distribution modeling of Rocky Mountain conifers. *Plant Ecology* 217: 533–547. <https://doi.org/10.1007/s11258-016-0598-5>
- Daganzo, C. F., V. V. Gayah, and E. J. Gonzales. 2012. The potential of parsimonious models for understanding large scale transportation systems and answering big picture questions. *EURO Journal on Transportation and Logistics* 1: 47–65. <https://doi.org/10.1007/s13676-012-0003-z>
- De Marco Jr., P., and C. C. Nóbrega. 2018. Evaluating collinearity effects on species distribution models: An approach based on virtual species simulation. *PLoS ONE* 13: e0202403. <https://doi.org/10.1371/journal.pone.0202403>
- Dewitz, J., and United States Geological Survey (USGS). 2021. National Land Cover Database (NLCD) 2019 products (version 2.0, June 2021). USGS data release. <https://doi.org/10.5066/P9KZCM54>
- Dormann, C. F., J. Elith, S. Bacher, C. Buchmann, G. Carl, G. Carré, J. R. García Marquéz, et al. 2012. Collinearity: A review of methods to deal with it and a simulation study evaluating their performance. *Ecography* 36: 27–46. <https://doi.org/10.1111/j.1600-0587.2012.07348.x>
- Dutra Silva, L., R. Bento Elias, and L. Silva. 2021. Modelling invasive alien plant distribution: A literature review of concepts and bibliometric analysis. *Environmental Modelling and Software* 145: 105203. <https://doi.org/10.1016/j.envsoft.2021.105203>
- El-Gabbas, A., and C. F. Dormann. 2018. Wrong, but useful: Regional species distribution models may not be improved by range-wide data under biased sampling. *Ecology and Evolution* 8: 2196–2206. <https://doi.org/10.1002/ece3.3834>
- Elith, J., C. H. Graham, R. P. Anderson, M. Dudík, S. Ferrier, A. Guisan, R. J. Hijmans, et al. 2006. Novel methods improve prediction of species' distributions from occurrence data. *Ecography* 29: 129. <https://doi.org/10.1111/j.2006.0906-7590.04596.x>
- Elith, J., J. R. Leathwick, and T. Hastie. 2008. A working guide to boosted regression trees. *Journal of Animal Ecology* 77: 802–813. <https://doi.org/10.1111/j.1365-2656.2008.01390.x>
- Elith, J., and J. R. Leathwick. 2009. Species distribution models: Ecological explanation and prediction across space and time. *Annual Review of Ecology, Evolution, and Systematics* 40: 677–697. <https://doi.org/10.1146/annurev.ecolsys.110308.120159>
- Elith, J., and J. Franklin. 2013. Species distribution modeling. In S. A. Levin [ed.], *Encyclopedia of biodiversity*, 2nd ed., 692–705. Academic Press, Cambridge, Massachusetts, USA.
- Environmental Protection Agency (EPA). 2023. Saginaw River and Bay AOC. Website <https://www.epa.gov/great-lakes-aocs/saginaw-river-and-bay-aoc> [accessed 2 November 2023].
- Feldman, M. J., L. Imbeau, P. Marchand, M. J. Mazerolle, M. Darveau, and N. J. Fenton. 2021. Trends and gaps in the use of citizen science derived data as input for species distribution models: A quantitative review. *PLoS ONE* 16: e0234587. <https://doi.org/10.1371/journal.pone.0234587>
- Fick, S. E., and R. J. Hijmans. 2017. WorldClim 2: New 1-km spatial resolution climate surfaces for global land areas. *International Journal of Climatology* 37: 4302–4315. <https://doi.org/10.1002/joc.5086>
- Fielding, A. H., and J. F. Bell. 1997. A review of the methods for the assessment of prediction errors in conservation presence/absence models. *Environmental Conservation* 24: 38–49. <https://doi.org/10.1017/S0376892997000088>
- Gale, J. 2004. Plant and altitude—Revisited. *Annals of Botany* 94: 199. <https://doi.org/10.1093/aob/mch143>
- Gonzalez, S. C., J. A. Soto-Centeno, and D. L. Reed. 2011. Population distribution models: Species distributions are better modeled using biologically relevant data partitions. *BMC Ecology* 11: 20. <https://doi.org/10.1186/1472-6785-11-20>
- Gori, M., A. Betti, and S. Melacci. 2024. Machine Learning, 2nd ed. Morgan Kaufmann, Burlington, Massachusetts, USA.
- Grimmett, L., R. Whited, and A. Horta. 2020. Presence-only species distribution models are sensitive to sample prevalence: Evaluating models using spatial prediction stability and accuracy metrics. *Ecological Modelling* 431: 109194. <https://doi.org/10.1016/j.ecolmodel.2020.109194>
- Guillera-Arroita, G., J. J. Lahoz-Monfort, J. Elith, A. Gordon, H. Kujala, P. E. Lentini, M. A. McCarthy, et al. 2015. Is my species distribution model fit for purpose? Matching data and models to applications. *Global Ecology and Biogeography* 24: 276–292. <https://doi.org/10.1111/geb.12268>
- Guisan, A., T. C. Edwards, and T. Hastie. 2002. Generalized linear and generalized additive models in studies of species distributions: Setting the scene. *Ecological Modelling* 157: 89–100. [https://doi.org/10.1016/S0304-3800\(02\)00204-1](https://doi.org/10.1016/S0304-3800(02)00204-1)
- Guisan, A., C. H. Graham, J. Elith, F. Huettmann, and the NCEAS Species Distribution Modelling Group. 2007. Sensitivity of predictive species distribution models to change in grain size. *Diversity and Distributions* 13: 332–340. <https://doi.org/10.1111/j.1472-4642.2007.00342.x>
- Guisan, A., R. Tingley, J. B. Baumgartner, I. Naujokaitis-Lewis, P. R. Sutcliffe, A. I. T. Tulloch, T. J. Regan, et al. 2013. Predicting species distributions for conservation decisions. *Ecology Letters* 16: 1424–1435. <https://doi.org/10.1111/ele.12189>
- Halpern, A. D. 2017. *Hydrocharis morsus-ranae* L. in the Upper St. Lawrence River in New York: Its success within heterogeneous wetland habitat and potential management approaches. Ph.D. dissertation, State University of New York College of Environmental Science and Forestry, Syracuse, New York, USA. <https://experts.esf.edu/esploro/outputs/doctoral/AD-Halpern-Hydrocharis-morsus-ranae-L-in/99872704204826>
- Hansen, S. E., B. C. Cahill, R. A. Hackett, M. J. Monfils, R. T. Goebel, S. Asencio, and A. Monfils. 2022a. Aggregated occurrence records of invasive European frog-bit (*Hydrocharis morsus-ranae* L.) across North America. *Biodiversity Data Journal* 10: e77492. <https://doi.org/10.3897/BDJ.10.e77492>
- Hansen, S. E., B. C. Cahill, R. A. Hackett, M. J. Monfils, R. T. Goebel, S. Asencio, and A. Monfils. 2022b. Aggregated occurrence records of invasive European frog-bit (*Hydrocharis morsus-ranae* L.) across North America. Occurrence dataset at GBIF: <https://doi.org/10.15468/jsab2y> [accessed 2 February 2024].
- Hansen, S. E., M. J. Monfils, R. A. Hackett, R. T. Goebel, and A. K. Monfils. 2023. Data-centric species distribution modeling. Website osf.io/y6meq [accessed 30 September 2023].
- Hardisty, A. R., E. R. Ellwood, G. Nelson, B. Zimkus, J. Buschbom, W. Addink, R. K. Rabeler, et al. 2022. Digital Extended Specimens: Enabling an extensible network of biodiversity data records as integrated digital objects on the Internet. *BioScience* 72: 978–987. <https://doi.org/10.1093/biosci/biac060>
- Hijmans, R. J. 2012. Cross-validation of species distribution models: Removing spatial sorting bias and calibration with a null model. *Ecology* 93: 679–688. <https://doi.org/10.1890/11-0826.1>
- Hijmans, R. J., and C. H. Graham. 2006. The ability of climate envelope models to predict the effect of climate change on species distribution. *Global Change Biology* 12: 2272–2281. <https://doi.org/10.1111/j.1365-2486.2006.01256.x>
- Hijmans, R. J., S. Phillips, J. Leathwick, and J. Elith. 2023. dismo: Species Distribution Modeling. R package version 1.3-14. <https://CRAN.R-project.org/package=dismo>

- Hui, C. 2023. The dos and don'ts for predicting invasion dynamics with species distribution models. *Biological Invasions* 25: 947–953. <https://doi.org/10.1007/s10530-022-02976-3>
- Jarnevich, C. S., and L. V. Reynolds. 2011. Challenges of predicting the potential distribution of a slow-spreading invader: A habitat suitability map for an invasive riparian tree. *Biological Invasions* 13: 153–163. <https://doi.org/10.1007/s10530-010-9798-4>
- Jarnevich, C. S., M. Talbert, J. Morisette, C. Aldrige, C. S. Brown, S. Kumar, D. Manier, et al. 2017. Minimizing effects of methodological decisions on interpretation and prediction in species distribution studies: An example with background selection. *Ecological Modelling* 363: 48–56. <https://doi.org/10.1016/j.ecolmodel.2017.08.017>
- Jarvis, A., H. I. Reuter, A. Nelson, and E. Geuevara. 2018. Shuttle Radar Topography Mission (SRTM) 90m DEM Digital Elevation Database (version 4, November 2018). CGIAR-CSI data release. <https://srtm.csi.cgiar.org/>
- Jiménez, L., and J. Soberón. 2020. Leaving the area under the receiving operating characteristic curve behind: An evaluation method for species distribution modelling applications based on presence-only data. *Methods in Ecology and Evolution* 11: 1571–1586. <https://doi.org/10.1111/2041-210X.13479>
- Kirzchner, F., and Y. Zimmels. 2006. Water vapor and air transport through ponds with floating aquatic plants. *Water Environment Research* 78: 880–886. <https://doi.org/10.2175/106143005X73127>
- Koncki, N. G., and M. F. J. Aronson. 2015. Invasion risk in a warmer world: Modeling range expansion and habitat preferences of three nonnative aquatic invasive plants. *Invasive Plant Science and Management* 8: 436–449. <https://doi.org/10.1614/IPSM-D-15-00020.1>
- Konowalik, K., and A. Nosol. 2021. Evaluation metrics and validation of presence-only species distribution models based on distributional maps with varying coverage. *Scientific Reports* 11: 1482. <https://doi.org/10.1038/s41598-020-80062-1>
- Krist Jr., F. J., and D. P. Lusch. 2004. Glacial history of Michigan, U.S.A.: A regional perspective. In J. Ehlers and P. L. Gibbard [eds.], *Quaternary glaciations—Extent and chronology, Part II*, 111–117. Elsevier Science, Amsterdam, the Netherlands.
- Kuhn, M. 2008. Building predictive models in R using the caret package. *Journal of Statistical Software* 28: 1–26. <https://doi.org/10.18637/jss.v028.i05>
- Leathwick, J. R., J. Elith, and T. Hastie. 2006. Comparative performance of generalized additive models and multivariate adaptive regression splines for statistical modelling of species distributions. *Ecological Modelling* 199: 188–196. <https://doi.org/10.1016/j.ecolmodel.2006.05.022>
- Lendemer, J., B. Thiers, A. K. Monfils, J. Zaspel, E. R. Ellwood, A. Bentley, K. LeVan, et al. 2020. The Extended Specimen Network: A strategy to enhance US biodiversity collections, promote research and education. *BioScience* 70: 23–30. <https://doi.org/10.1093/biosci/biz140>
- Lobo, J. M., A. Jiménez-Valverde, and R. Real. 2008. AUC: A misleading measure of the performance of predictive distribution models. *Global Ecology and Biogeography* 17: 145–151. <https://doi.org/10.1111/j.1466-8238.2007.00358.x>
- Lozano, V. 2021. Distribution of five aquatic plants native to South America and invasive elsewhere under current climate. *Ecologies* 2: 27–42. <https://doi.org/10.3390/ecologies2010003>
- Mainali, K. P., D. L. Warren, K. Dhileepan, A. McConnachie, L. Strathie, G. Hassan, D. Karki, et al. 2015. Predicting future expansion of invasive species: Comparing and improving methodologies for species distribution modeling. *Global Change Biology* 21: 4464–4480. <https://doi.org/10.1111/gcb.13038>
- Marcet, A., A. D. Chapman, J. R. Wieczorek, F. X. Picó, F. Uribe, J. Waller, and A. H. Ariño. 2022. Uncertainty matters: Ascertaining where specimens in natural history collections come from and its implications for predicting species distributions. *Ecography* 2022: e06025. <https://doi.org/10.1111/ecog.06025>
- Meier, E. S., F. Kienast, P. B. Pearman, J.-C. Svenning, W. Thuiller, M. B. Araújo, A. Guisan, and N. E. Zimmermann. 2010. Biotic and abiotic variables show little redundancy in explaining tree species distributions. *Ecography* 33: 1038–1048. <https://doi.org/10.1111/j.1600-0587.2010.06229.x>
- Menuz, D. R., K. M. Kettenring, C. P. Hawkins, and D. R. Cutler. 2015. Non-equilibrium in plant distribution models – only an issue for introduced or dispersal limited species? *Ecography* 38: 231–240. <https://doi.org/10.1111/ecog.00928>
- Michigan Department of Natural Resources (MDNR). 2023. Michigan Public Boating Access Sites. MDNR data release. Michigan Department of Natural Resources, Lansing, Michigan, USA. Data set https://gis-midnr.opendata.arcgis.com/datasets/e33baefa170b46928869155f3627de92_1/about [accessed 24 October 2023].
- Minshall, W. H. 1940. Frog-bit *Hydrocharis morsus-ranae* L. at Ottawa. *Canadian Field-Naturalist* 54: 44–45.
- Monfils, A. K., E. R. Krimmel, J. M. Bates, J. E. Bauer, M. W. Belitz, B. C. Cahill, A. M. Caywood, et al. 2020. Regional collections are an essential component of biodiversity research infrastructure. *BioScience* 70: 1045–1047. <https://doi.org/10.1093/biosci/biaa102>
- Monfils, A. K., M. J. Monfils, B. C. Cahill, R. T. Goebel, S. E. Hansen, R. A. Hackett, and Y. Lee. 2021. Final report on project #2019-EFB1 of grant #00E02463 European frog-bit: Enhancing control and assessing impacts and management. Technical report. Central Michigan University, Michigan, USA. <https://doi.org/10.13140/RG.2.2.36773.01766>
- Moudry, V., and R. Devillers. 2020. Quality and usability challenges of global marine biodiversity databases: An example for marine mammal data. *Ecological Informatics* 56: 101051. <https://doi.org/10.1016/j.ecoinf.2020.101051>
- Mukaka, M. M. 2012. Statistics Corner: A guide to appropriate use of correlation coefficient in medical research. *Malawi Medical Journal* 24: 69–71.
- Muñoz-Mas, R., E. Gil-Martínez, F. J. Oliva-Paterna, E. J. Belda, and F. Martínez-Capel. 2019. Tree-based ensembles unveil the micro-habitat suitability for the invasive bleak (*Alburnus alburnus* L.) and pumpkinseed (*Lepomis gibbosus* L.): Introducing XGBoost to eco-informatics. *Ecological Informatics* 53: 100974. <https://doi.org/10.1016/j.ecoinf.2019.100974>
- Osborne, P. E., and S. Suárez-Seoane. 2002. Should data be partitioned spatially before building large-scale distribution models? *Ecological Modelling* 157: 249–259. [https://doi.org/10.1016/S0304-3800\(02\)00198-9](https://doi.org/10.1016/S0304-3800(02)00198-9)
- Pearson, R. G., T. P. Dawson, P. M. Perry, and P. A. Harrison. 2002. SPECIES: A spatial evaluation of climate impact on the envelope of species. *Ecological Modelling* 154: 289–300. [https://doi.org/10.1016/S0304-3800\(02\)00056-X](https://doi.org/10.1016/S0304-3800(02)00056-X)
- Peel, M. C., B. L. Finlayson, and T. A. McMahon. 2007. Updated world map of the Köppen-Geiger climate classification. *Hydrology and Earth System Sciences* 11: 1633–1644. <https://doi.org/10.5194/hess-11-1633-2007>
- Petitpierre, B., O. Broennimann, C. Kueffer, C. Daehler, and A. Guisan. 2017. Selecting predictors to maximize the transferability of species distribution models: Lessons from cross-continental plant invasions. *Global Ecology and Biogeography* 26: 275–287. <https://doi.org/10.1111/geb.12530>
- Phillips, S. J., R. P. Anderson, and R. E. Schapire. 2006. Maximum entropy modeling of species geographic distributions. *Ecological Modelling* 190(3–4): 231–259. <https://doi.org/10.1016/j.ecolmodel.2005.03.026>
- Phillips, S. J., and M. Dudík. 2008. Modeling of species distributions with Maxent: New extensions and a comprehensive evaluation. *Ecography* 31: 161–175. <https://doi.org/10.1111/j.0906-7590.2008.5203.x>
- Phillips, S. J., R. P. Anderson, M. Dudík, R. E. Schapire, and M. E. Blair. 2017. Opening the black box: An open-source release of Maxent. *Ecography* 40: 887–893. <https://doi.org/10.1111/ecog.03049>
- Pili, A. N., R. Tingley, E. Y. Sy, M. L. L. Diesmos, and A. C. Diesmos. 2020. Niche shifts and environmental non-equilibrium undermine the usefulness of ecological niche models for invasion risk assessments. *Scientific Reports* 10: 7972. <https://doi.org/10.1038/s41598-020-64568-2>
- Poggio, L., L. M. de Sousa, N. H. Batjes, G. B. M. Heuvelink, B. Kempen, E. Ribeiro, and D. Rossiter. 2021. SoilGrids 2.0: Producing soil information for the globe with quantified spatial uncertainty. *SOIL* 7: 217–240. <https://doi.org/10.5194/soil-7-217-2021>

- Pyke, G. H., and P. R. Ehrlich. 2010. Biological collections and ecological/environmental research: A review, some observations and a look to the future. *Biological Reviews* 85: 247–266. <https://doi.org/10.1111/j.1469-185X.2009.00098.x>
- Qiao, H., X. Feng, L. E. Escobar, A. T. Peterson, J. Soberón, G. Zhu, and M. Papeš. 2019. An evaluation of transferability of ecological niche models. *Ecography* 42: 521–534. <https://doi.org/10.1111/ecog.03986>
- R Core Team. 2023. R: A language and environment for statistical computing. Version 4.3.0 (2023-04-21). R Foundation for Statistical Computing, Vienna, Austria. Website <https://www.R-project.org/> [accessed 8 September 2023].
- Radosavljevic, A., and R. P. Anderson. 2014. Making better MAXENT models of species distributions: Complexity, overfitting and evaluation. *Journal of Biogeography* 41: 629–643. <https://doi.org/10.1111/jbi.12227>
- Renner, I. W., and D. I. Warton. 2013. Equivalence of MAXENT and Poisson point process models for species distribution modeling in ecology. *Biometrics* 69: 274–281. <https://doi.org/10.1111/j.1541-0420.2012.01824.x>
- Richards, A. J., and J. Blakemore. 1975. Factors affecting the germination of turions in *Hydrocharis morsus-ranae* L. *Watsonia* 10: 273–275.
- Rodríguez-Castañeda, G., A. R. Hof, R. Jansson, and L. E. Harding. 2012. Predicting the fate of biodiversity using species' distribution models: Enhancing model comparability and repeatability. *PLoS ONE* 7: e44402. <https://doi.org/10.1371/journal.pone.0044402>
- Rodríguez-Rey, M., S. Consuegra, L. Börger, and C. Garcia de Leaniz. 2019. Improving species distribution modelling of freshwater invasive species for management applications. *PLoS ONE* 14: e0217896. <https://doi.org/10.1371/journal.pone.0217896>
- Scheffer, M., S. Szabo, A. Gragnani, E. H. van Nes, S. Rinaldi, N. Kautsky, J. Norberg, et al. 2003. Floating plant dominance as a stable state. *Proceedings of the National Academy of Sciences, USA* 100: 4040–4045. <https://doi.org/10.1073/pnas.0737918100>
- Schuwirth, N., F. Borgwardt, S. Domsch, M. Friedrichs, M. Kattwinkel, D. Kneis, M. Kuemmerlen, et al. 2019. How to make ecological models useful for environmental management. *Ecological Modelling* 411: 108784. <https://doi.org/10.1016/j.ecolmodel.2019.108784>
- Soultan, A., and K. Safi. 2017. The interplay of various sources of noise on reliability of species distribution models hinges on ecological specialisation. *PLoS ONE* 12: e0187906. <https://doi.org/10.1371/journal.pone.0187906>
- Su, H., M. Bista, and M. Li. 2021. Mapping habitat suitability for Asiatic black bear and red panda in Makalu Barun National Park of Nepal from Maxent and GARP models. *Scientific Reports* 11: 14135. <https://doi.org/10.1038/s41598-021-93540-x>
- Tang, F., and H. Ishwaran. 2017. Random forest missing data algorithms. *Statistical Analysis and Data Mining* 10: 363–377. <https://doi.org/10.1002/sam.11348>
- Tepa-Yotto, G. T., H. E. Z. Tonnang, G. Goergen, S. Subramanian, E. Kimathi, E. M. Abdel-Rahman, D. Flø, et al. 2021. Global habitat suitability of *Spodoptera frugiperda* (JE Smith) (Lepidoptera, Noctuidae): Key parasitoids considered for its biological control. *Insects* 12: 273. <https://doi.org/10.3390/insects12040273>
- Tessarolo, G., J. M. Lobo, T. F. Rangel, and J. Hortal. 2021. High uncertainty in the effects of data characteristics on the performance of species distribution models. *Ecological Indicators* 121: 107147. <https://doi.org/10.1016/j.ecolind.2020.107147>
- Václavík, T., and R. K. Meentemeyer. 2012. Equilibrium or not? Modelling potential distribution of invasive species in different stages of invasion. *Diversity and Distributions* 18: 73–83. <https://doi.org/10.1111/j.1472-4642.2011.00854.x>
- Valavi, R., J. Elith, J. J. Lahoz-Monfort, and G. Guillera-Arroita. 2021. Modelling species presence-only data with random forests. *Ecography* 44: 1731–1742. <https://doi.org/10.1111/ecog.05615>
- Valavi, R., G. Guillera-Arroita, J. J. Lahoz-Monfort, and J. Elith. 2022. Predictive performance of presence-only species distribution models: A benchmark study with reproducible code. *Ecological Monographs* 92(1): e01486. <https://doi.org/10.1002/ecm.1486>
- Vignali, S., A. G. Barras, R. Arlettaz, and V. Braunisch. 2020. *SDMtune*: An R package to tune and evaluate species distribution models. *Ecology and Evolution* 10: 11488–11506. <https://doi.org/10.1002/ece3.6786>
- Webster, M. 2017. *The Extended Specimen: Emerging frontiers in collections-based ornithological research*. CRC Press, Boca Raton, Florida, USA.
- Wilkinson, D. P., N. Golding, G. Guillera-Arroita, R. Tingley, and M. A. McCarthy. 2021. Defining and evaluating predictions of joint species distribution models. *Methods in Ecology and Evolution* 12: 394–404. <https://doi.org/10.1111/2041-210X.13518>
- Willmott, C. J., and K. Matsuura. 2005. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Research* 30: 79–82. <https://doi.org/10.3354/cr030079>
- Wood, S. 2023. *mgcv*: Mixed GAM Computation Vehicle with Automatic Smoothness Estimation. R package version 1.8-42. <https://cran.r-project.org/package=mgcv>
- Yang, X., W. M. Post, P. E. Thornton, and A. Jain. 2014. Global gridded soil phosphorus distribution maps at 0.5-degree resolution (version 3, May 2014). Oak Ridge National Laboratory Distribution Active Archive Center (ORNL DAAC) data release. <https://doi.org/10.3334/ORNLDAAC/1223>
- Yerushalmi, J. 1947. Statistical problems in assessing methods of medical diagnosis, with special reference to X-ray techniques. *Public Health Reports* 62: 1432–1449. <https://doi.org/10.2307/4586294>
- Yu, H., A. R. Cooper, and D. M. Infante. 2020. Improving species distribution model predictive accuracy using species abundance: Application with boosted regression trees. *Ecological Modelling* 432: 109202. <https://doi.org/10.1016/j.ecolmodel.2020.109202>
- Zarzo-Arias, A., V. Penteriani, L. Gábor, P. Šimová, F. Grattarola, and V. Moudrý. 2022. Importance of data selection and filtering in species distribution models: A case study of the Cantabrian brown bear. *Ecosphere* 13: e4284. <https://doi.org/10.1002/ecs2.4284>
- Zhu, B., M. S. Ellis, K. L. Fancher, and L. G. Rudstam. 2014. Shading as a control method for invasive European frogbit (*Hydrocharis morsus-ranae* L.). *PLoS ONE* 9: e98488. <https://doi.org/10.1371/journal.pone.0098488>
- Zhu, B., C. C. Ottaviani, R. Naddaft, Z. Dai, and D. Du. 2018. Invasive European frogbit (*Hydrocharis morsus-ranae* L.) in North America: An updated review 2003–16. *Journal of Plant Ecology* 11: 17–25. <https://doi.org/10.1093/jpe/rtx031>
- Zhu, G.-P., and A. T. Peterson. 2017. Do consensus models outperform individual models? Transferability evaluations of diverse modeling approaches for an invasive moth. *Biological Invasions* 19: 2519–2532. <https://doi.org/10.1007/s10530-017-1460-y>
- Zuckerberg, B., F. Huettmann, and J. Frair. 2010. Proper data management as a scientific foundation for reliable species distribution modeling. In C. A. Drew, Y. F. Wiersma, and F. Huettmann [eds.], *Predictive species and habitat modeling in landscape ecology*, 45–70. Springer, New York, New York, USA.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

Appendix S1. Mean values by model type for the inverse of mean absolute error, sensitivity, and specificity for the North America and Michigan models.

Appendix S2. Results of Wilcoxon rank-sum tests comparing the delineation of the large-scale training region, as either North America or Michigan.

Appendix S3. Predicted probability of suitable habitat in the Michigan study region given by the final eXtreme gradient boosting machine (XGBoost) models built with North America and Michigan occurrences.

Appendix S4. Differences based on training region in predicted probability of suitable habitat given by the final eXtreme gradient boosting machine (XGBoost) models in the Michigan study region.

Appendix S5. The Michigan and Saginaw Bay study regions based on European frog-bit observations.

Appendix S6. Distribution of specimen and observation response data for the Michigan scale. Green points represent original European frog-bit presences before occurrence thinning.

Appendix S7. Results of Wilcoxon rank-sum tests comparing the method of removing missing data prior to modeling, by either variable or observation, grouped by the threshold for excluding correlated variables and scale, data source, and data type.

Appendix S8. Results of Wilcoxon rank-sum tests comparing the threshold for excluding correlated variables for the Michigan scale, grouped by data source and data type.

Appendix S9. Results of Wilcoxon rank-sum tests comparing the scale of response data, either Michigan or Saginaw Bay.

Appendix S10. Results of Wilcoxon rank-sum tests comparing the source of response data for the Michigan scale, either specimens or observations.

Appendix S11. Results of Wilcoxon rank-sum tests comparing the type of response data for the

Saginaw Bay scale, either presence-absence or presence-background.

Appendix S12. Predicted probability of suitable habitat given by the final eXtreme gradient boosting machine (XGBoost) models in Michigan, using specimen-based and observation-based response data combined (“all data”), specimens only, and observations only.

Appendix S13. Differences based on data source in predicted probability of suitable habitat given by the final eXtreme gradient boosting machine (XGBoost) models in Michigan.

Appendix S14. Predicted probability of suitable habitat given by the final eXtreme gradient boosting machine (XGBoost) models in Saginaw Bay, using presence-background and presence-absence response data.

Appendix S15. Differences based on data type in predicted probability of suitable habitat given by the final eXtreme gradient boosting machine (XGBoost) models in Saginaw Bay.

How to cite this article: Hansen, S. E., M. J. Monfils, R. A. Hackett, R. T. Goebel, and A. K. Monfils. 2024. Data-centric species distribution modeling: Impacts of modeler decisions in a case study of invasive European frog-bit. *Applications in Plant Sciences* 12(3): e11573. <https://doi.org/10.1002/aps3.11573>