

Research Article

Cite this article: Jorgensen N and Renz M (2021) Assessing the performance and accuracy of invasive plant habitat suitability models in detecting new observations in Wisconsin. *Invasive Plant Sci. Manag* **14**: 214–222. doi: [10.1017/inp.2021.27](https://doi.org/10.1017/inp.2021.27)

Received: 18 November 2020
Revised: 15 July 2021
Accepted: 30 August 2021
First published online: 6 September 2021

Associate Editor:

Catherine Jarnevich, U.S. Geological Survey

Keywords:

Community science; ensemble modeling; invasive plants

Author for correspondence:

Mark Renz, University of Wisconsin–Madison, 1575 Linden Drive, Madison, WI 53706. (Email: mrenz@wisc.edu)

Assessing the performance and accuracy of invasive plant habitat suitability models in detecting new observations in Wisconsin

Niels Jorgensen¹  and Mark Renz²

¹Former Graduate Student, University of Wisconsin–Madison, Madison, WI, USA and ²Professor, University of Wisconsin–Madison, Madison, WI, USA

Abstract

Land managers require tools that improve understanding of suitable habitat for invasive plants and that can be incorporated into survey efforts to improve efficiency. Habitat suitability models (HSMs) contain attributes that can meet these requirements, but it is not known how well they perform, as they are rarely field-tested for accuracy. We developed ensemble HSMs in the state of Wisconsin for 15 species using five algorithms (boosted regression trees, generalized linear models, multivariate regression splines, MaxEnt, and random forests), evaluated performance, determined variables that drive suitability, and tested accuracy. All models had good model performance during the development phase (Area Under the Curve [AUC] > 0.7 and True Skills Statistic [TSS] > 0.4). While variable importance and directionality was species specific, the most important predictor variables across all of the species' models were mean winter minimum temperatures, total summer precipitation, and tree canopy cover. Post model development, we obtained 5,005 new occurrence records from community science observations for all 15 focal species to test the models' abilities to accurately predict results. Using a correct classification rate of 80%, just 8 of the 15 species correctly predicted suitable habitat ($\alpha \leq 0.05$). Exploratory analyses found the number of reporters of these new data and the total number of new occurrences reported per species contributed to increasing correct classification. Results suggest that while some models perform well on evaluation metrics, relying on these metrics alone is not sufficient and can lead to errors when utilized for surveying. We recommend any model should be tested for accuracy in the field before use to avoid this potential issue.

Introduction

Invasive plants are a pervasive problem that impact natural and managed lands. They can disrupt native plant, animal, and microbial communities (Dukes and Mooney 2004); alter ecosystem processes (DiTomaso 2000; Levine et al. 2003); be costly to control (Epanchin-Niell et al. 2010); and even impact human health (Mazza et al. 2014). While efforts to prevent invasions and their associated impacts have had success (e.g., Rothlisberger et al. 2010), establishment and spread of invasive species remains common in natural areas. Land management agencies commonly employ early detection and rapid response (EDRR) to limit these impacts (Reaser et al. 2019; Westbrooks 2004). Within this framework, early detection is a critical first step, as it minimizes spread, which prevents impacts across a large area (Maistrello et al. 2016) and allows for cost-effective management (Westbrooks 2004).

Invasive species monitoring efforts, however, are a challenge for land management organizations. Maxwell et al. (2012) estimate that 1% to 2% of managed natural lands are monitored annually for invasive plants. While increasing monitoring is the optimal solution, few have budgets to enhance these efforts, and many state and federal agencies have reduced monitoring programs (NISC 2018). As professional staff and monitoring resources are unlikely to increase, developing tools to improve efficiency of existing monitoring efforts is an attractive alternative that has been a high priority (WISC 2013). Habitat suitability models (HSMs) are one example, as they have the potential to assist with early detection of invasive species. HSMs can improve detection of target species with much greater efficiency (up to 80% detection rate; Wang et al. 2014) compared with the random searches currently employed (Crall et al. 2013). However, a general disconnect exists between researchers and land managers (Renz et al. 2009), which suggests that although these models have potential benefits to land managers, their actual implementation on the landscape is limited.

HSMs have been developed for many invasive species (e.g., Allen and Bradley 2016; Magarey et al. 2018) at a range of scales (e.g., Young et al. 2020). While predictor variables often differ depending on the focal species, climatic variables (e.g., precipitation and temperature), remotely sensed vegetation indices, landscape attributes (e.g., aspect, slope), solar irradiation,

© The Author(s), 2021. Published by Cambridge University Press on behalf of the Weed Science Society of America. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted re-use, distribution, and reproduction in any medium, provided the original work is properly cited.



Management Implications

With availability of resources for invasive plant monitoring and control efforts nationwide in decline, new tools and technologies are required to fill this gap. Habitat suitability models (HSMs) are an effective tool for improving understanding and detection of invasive plants on the landscape. HSMs use environmental and physical variables of known locations of a focal species to predict where other similar locations on the landscape exist. While HSMs are common in the scientific community, gaps between development of these tools and use by land management agencies exist. As these models are rarely evaluated for accuracy, their implementation in the field has been limited due to concerns over real-world performance. As developers rarely have resources to field validate, this study sought to determine whether HSMs accurately identify presence of community scientist observations of 15 invasive plants throughout Wisconsin. A total of 5,005 verified new occurrences collected over 2 yr from a structured network of community scientists was utilized to test the ability of these models to correctly identify suitable habitat. While model performance looked successful during development, only half of the models performed at or above our performance threshold (80% correct classification). This result underscores the importance of vigorous field testing of accuracy of these models with separate data before release to land managers to ensure accuracy.

distance to dispersal corridors (e.g., water, roads), and soil attributes are commonly used for creation of HSMs (Evangelista et al. 2008; Kumar et al. 2006; Stohlgren et al. 2003). Predictions are best when true absence data are included, but these data sets are difficult to attain. Therefore, targeted background approaches have been developed and demonstrated to be an effective alternative (Crall et al. 2013; Phillips et al. 2009). Different algorithms have been utilized in modeling suitable habitat; however, ensemble predictions utilizing multiple algorithms have shown better predictive capabilities of suitability than reliance on any single algorithm (Margery et al. 2018; Morissette et al. 2013). Updating models with new data (iterative approach) can also improve predictions over time (Cook et al. 2019; Crall et al. 2013).

While HSMs are available for a wide range of invasive species, their use by practitioners and land managers is limited. Improvement in transparency in habitat modeling development is one approach researchers are taking to address this (Araújo et al. 2019; Sofaer et al. 2019) and improve understanding and adoption of HSMs on the landscape. However, these tools are rarely field-tested for accuracy (Barbet-Massin et al. 2018; West et al. 2016). Thus, the effectiveness of these tools on the landscape is not well known, and this may reduce adoption (Funk et al. 2020; Sofaer et al. 2019). While validation can be difficult, increases in community scientist and stakeholder monitoring efforts that are publicly shared provide a resource to test the accuracy of these models and potentially increase adoption of HSMs.

As part of an outreach effort from 2014 to 2018 focusing on detection of invasive species within Wisconsin (Great Lakes Early Detection Network established by Crall et al. [2012]), we collected known invasive species occurrences for invasive plants and promoted monitoring throughout the state to stakeholders and through a community science group (WIFDN 2020). These data were used to develop HSMs for 15 invasive plants in the state of

Wisconsin using an ensemble approach. With these models, we assessed performance and the drivers of suitable habitat and, after 2 yr, determined the accuracy of HSMs using occurrences reported by community scientists and stakeholders. As the accuracy of HSMs varied among species, further analysis sought to understand what factors are influencing the ability of models in correct classification of invasive plant presence.

Materials and Methods

Predictor Variables

Fourteen candidate predictor variables (Table 1) were utilized for model development. Topographic variables (aspect, elevation, and slope) and distance to dispersal corridors (roads, waterways) were created from a digital elevation model and the Topologically Integrated Geographic Encoding and Referencing (TIGER) data sets, respectively, using ArcMap (ESRI v. 10.5, 2016, Redlands, CA). Climate predictors were obtained from the AdaptWest Project (Wang et al. 2016) and processed to create seasonal average precipitation and average summer maximum and winter minimum temperatures (1981 to 2010 climate normals). Soil attributes (percent organic matter and percent clay) were derived from the USDA gridded soil survey and were included to distinguish soil types across the study area. Finally, a 10-yr average enhanced vegetation index (EVI; Didan 2015) and percent tree canopy cover (Sexton et al. 2013), both calculated from MODIS satellite data, were used as measures of greenness to differentiate canopy density and vegetation types.

Occurrence Data

More than 100,000 occurrence records for more than 100 species of terrestrial invasive plants were obtained through the Wisconsin Department of Natural Resources and the Early Detection and Distribution Mapping System from 2000 to 2016. Occurrences were evaluated, and duplicates and occurrences with GPS error >30 m were removed. Fifteen invasive plants (Table 2) were ultimately chosen for modeling in this study.

Habitat Suitability Models

All habitat models were created using the Software for Assisted Habitat Modeling (SAHM) (Morissette et al. 2013) modules within the broader Vistrails framework (Freire and Silva 2012). SAHM was designed to assist researchers with the development of spatial distribution models by allowing the user to make decisions in an interactive workflow. For each species modeled, the presence data were complemented with pseudo-absence records by utilizing additional reported invasive plant species occurrences taken from the same surveys performed for the focal species. Phillips et al. (2009) describe this “targeted background” approach to habitat models as providing a better representation of absence data, rather than randomly assigned background points, when true absence data are not available.

The 14 predictor layers were projected, aggregated (mean values), resampled (bilinear interpolation), and clipped to the extent of the state of Wisconsin using the PARC module within SAHM. These predictors were selected based on previous research in Wisconsin (Crall et al. 2013) to identify fine-scale habitat suitability. While spatial resolutions of 30, 100, and 250 m were considered, we selected a final spatial resolution of 30 m, as this provided similar results as coarser resolutions (data not shown)

Table 1. Predictors used for model development, including data source and whether postprocessing of the data was required.^a

Predictor layer	Native resolution	Data source	Processing notes
Aspect (° from N)	30 m	USGS National Elevation Data Set	Derived from DEM
Average maximum summer temperatures (1981–2010)	1 km	AdaptWest Project	June, July, August
Average minimum winter temperatures (1981–2010)	1 km	AdaptWest Project	December, January, February
Distance from roads	30 m	TIGER Roads	Derived (Euclidean distance) from road networks
Distance from water	30 m	USGS National Hydrography Data Set	Derived (Euclidean distance) from streams, rivers, waterbodies
Fall precipitation (1981–2010)	1 km	AdaptWest Project	September, October, November
Mean annual (2000–2010) enhanced vegetation index	~230 m	NASA MODIS	Average values created using USGS Phenology Image Service Toolbox
Percent clay in soil	30 m	USDA NRCS gSSURGO	
Percent organic matter in soil	30 m	USDA NRCS gSSURGO	
Percent tree canopy cover	30 m	Global Land Cover Facility	Vegetation continuous fields layer from NASA MODIS
Slope (°)	30 m	USGS National Elevation Data Set	Derived from DEM
Spring precipitation (1981–2010)	1 km	AdaptWest Project	March, April, May
Summer precipitation (1981–2010)	1 km	AdaptWest Project	June, July, August
Winter precipitation (1981–2010)	1 km	AdaptWest Project	December, January, February

^aAll GIS-related manipulations were performed in ArcMap (v. 10.5, ESRI).

Table 2. Species modeled, including common and scientific names, number of occurrence records used to develop the models, and how many counties in the state those points came from (an indication of the spatial extent of these data).

Species common name	Species scientific name	Number of occurrence records	Number of counties
Autumn olive	<i>Elaeagnus umbellata</i> Thunb.	169	26
Bush honeysuckle spp. (Amur honeysuckle, Bell's honeysuckle, Morrow's honeysuckle, Tatarian honeysuckle)	<i>Lonicera maackii</i> (Rupr.) Herder, <i>Lonicera × bella</i> Zabel, <i>Lonicera morrowii</i> A. Gray, <i>Lonicera tatarica</i> L.	1758	44
Canada thistle	<i>Cirsium arvense</i> (L.) Scop.	902	29
Common buckthorn	<i>Rhamnus cathartica</i> L.	806	44
Common tansy	<i>Tanacetum vulgare</i> L.	2203	29
Crownvetch	<i>Securigera varia</i> (L.) Lassen	177	35
Garlic mustard	<i>Alliaria petiolata</i> (M. Bieb.) Cavara & Grande	2265	56
Hedge-parsley spp. (Japanese hedge-parsley, spreading hedge-parsley)	<i>Torilis arvensis</i> (Huds.) Link, <i>Torilis japonica</i> (Houtt.) DC.	380	24
Japanese barberry	<i>Berberis thunbergii</i> DC.	343	33
Knotweed spp. (Japanese knotweed, giant knotweed, Bohemian knotweed)	<i>Reynoutria japonica</i> Houtt., <i>Reynoutria sachalinensis</i> (F. Schmidt ex Maxim.) Nakai, <i>Reynoutria × bohemica</i> Chrtek & Chrtková	547	62
Leafy spurge	<i>Euphorbia esula</i> L.	230	43
Purple loosestrife	<i>Lythrum salicaria</i> L.	572	72
Spotted knapweed	<i>Centaurea stoebe</i> L.	1783	69
Teasel spp. (common teasel, cutleaf teasel)	<i>Dipsacus fullonum</i> L., <i>Dipsacus laciniatus</i> L.	1013	30
Wild parsnip	<i>Pastinaca sativa</i> L.	3491	71

and provided a more detailed map that could be utilized by land managers. A 10-fold cross-validation test of the training models was conducted on each individual species' models to determine whether overfitting of the data had occurred during model development. The cross-validation method randomly withholds 10% of the data and tests it against the trained model, iterating the process 10 times, with the goal of observing limited variance in the prediction. In the case that predictor layers were highly correlated ($|r| > 0.75$), based on three coefficient indices (Pearson, Spearman, and Kendall), the predictor with the greater biological importance for the specific species, and the higher percent deviance explained for a simple generalized additive model was selected (Morissette et al. 2013).

Five algorithms (boosted regression trees, generalized linear models, multivariate adaptive regression splines, maximum entropy, and random forests [RFs]) were used to produce habitat suitability predictions for the state of Wisconsin for the species listed in Table 2. Each algorithm's predictive performance was assessed using both threshold-independent (Area Under the Curve [AUC]) and threshold-dependent (sensitivity, specificity, true skills statistic [TSS]) metrics. If AUC values for the training and cross-validation test were ≥ 0.075 , model hyperparameters (see Supplementary Table S1 for more details) were adjusted to curtail overfitting. If overfitting could not be limited to meet this standard, those models were not included in analyses.

Each algorithm produced a unique probability density surface of predicted suitability for the study area. These probability surfaces were converted to a binary simplification using a threshold cut-off found by maximizing the average of the sensitivity plus specificity. Pixels considered above this threshold were given a value of 1, while unsuitable pixels were designated as 0. Each binary prediction was then summed to create an ensemble model (see Morissette et al. 2013) of suitability for a specific species, and a final model was created where at least one model deemed a pixel suitable.

Variable importance and response curves were produced by SAHM and examined for similarities across the five models. Variable importance is calculated by determining the change in AUC when predictor values are permuted between the occurrence records and background data points. The larger the change produced by the permutation, the greater the influence the predictor has on the model. The predictors selected by each of the five different modeling approaches varied, so we relativized the values to a percentage of all the variables used in the model. Marginal response curves were also produced for each predictor in the model by calculating the response when the other predictors are held constant at their mean values (Jarnevich et al. 2018; Lötter and Maitre 2014; Pearson 2007).

Community Science Outreach Campaign

In 2017, following the development of our HSMs, we developed an active outreach campaign through Wisconsin First Detector Network to report invasive plants in Wisconsin. Our campaign trained individuals to identify key invasive species we were modeling and report occurrences. Over the 2-yr time frame, 34 workshops conducted in 19 counties trained 986 people.

Testing Accuracy of Models from Community Scientist Observations

As a result of our community outreach campaign, more than 15,000 new invasive plant reports were submitted by community scientists in the state of Wisconsin. We used a subset of these reports to assess how well our models predicted presence of specific invasive plants. Reports were used only if they were verified by an expert reviewer and had a GPS error ≤ 30 m. Additionally, accuracy was only tested when we had >40 observations for each species in this new data set. This resulted in 5,005 new reports to test 15 species distribution models. Each report was run through the species-specific ensemble model to determine the sensitivity associated with the model predictions (Jorgensen et al. 2021). A χ^2 test was performed in R (R Core Team 2019) to determine whether the correct classification rate of these new occurrence records was higher than an 80% cutoff rate. This cutoff value for assessing predictive performance is commonly used in remote-sensing evaluation (Engler et al. 2013).

To determine the drivers of correct classification of the community scientist observations across the 15 species evaluated, we developed a RF model in R using the RANDOMFOREST package. RF is a nonparametric ensemble machine learning approach that stochastically builds regression trees via a bootstrap technique (i.e., leave-one-out sampling with replacement). RF models are robust to data sets including large quantities of both continuous and categorical variables compared with the number of observations (Cutler et al. 2007). The importance of predictor variables is ranked by each tree in the forest, and the tree rankings are then averaged to develop unbiased estimations of the most important predictors

where the overall model error is minimized. Twenty-six predictor variables were used to understand what is impacting the accuracy of community scientist observations. Twenty model evaluation metrics (training AUC, TSS, sensitivity, and specificity for each individual model in the species ensemble), three variables associated with the HSM (number of occurrences, pseudo-absence records, and number of counties where the species had been reported), and three variables associated with community scientist data (the number of new occurrence records to test the models, and how many reporters contributed these reports) were used in this RF analysis. The model was tuned to curtail overfitting by setting the mtry parameter to the square root of the number of predictors and the number of trees to 500 (Probst et al. 2019).

Results and Discussion

Model Characteristics and Important Variables

The ensemble binary models for all 15 species' HSMs were created and deemed appropriate based on evaluation metrics (AUC, TSS) (Figure 1; Supplementary Figures S1–S14.) Two to 14 variables (Table 3) were retained in the final models, depending upon multicollinearity or the intrinsic model algorithm variable selection method for each species. The relative importance of the predictor variables was both species and model specific. The top three predictors (Table 3) were either consistently important (i.e., within the top three at least 30% of the time), occasionally important (in the top three 10% to 30% of the time), or rarely influential on model predictions (in the top three $<10\%$ of the time). Minimum winter temperature was the most common important variable across species models (66% of the top three variables across all 15 models), which is consistent with other work modeling invasive species habitat suitability (Young et al. 2020). We also found summer precipitation (47%), summer maximum (41%) temperatures, and percent tree canopy cover (31%) consistently important across species' models. Precipitation in the fall, spring, and winter was less consistently important, ranking within the top three most important predictors in 27%, 23%, and 27% of the models, respectively. Similarly, percent clay in the soil and distance to road networks were among the top three most important predictors for 14% and 12% of the models, respectively. Crall et al. (2013) found percent clay to be the most important predictor for modeling suitable habitat of spotted knapweed (*Centaurea stoebe* L.) in Wisconsin, while it ranked within the top three important predictors 14 times in our study, two of which were *C. stoebe*.

Predictors that were rarely ranked in the top three most important were the topographic variables (aspect, slope), EVI, distance to water, and percent soil organic matter (Table 3). Wisconsin has relatively homogenous topography with only a 426-m difference between the highest and lowest point in the state (mean = 320 m), and a mean slope of just 2.7% rise (range: 0% to 59%). The small differences among these variables in Wisconsin may be a factor in the importance of this variable in model predictions. Similarly, distance to water was only ranked among the top three important predictors in wetland invasive species (*Phragmites* and purple loosestrife [*Lythrum salicaria* L.]).

In addition to differences in important predictor variables among species, directionality of the response also varied among species. For example, response to tree canopy cover varied, with 9 species responding positively, 10 negatively, and 2 with a bimodal response (data not shown). While others suggest that HSM for species with common growth forms can be combined, as they behave

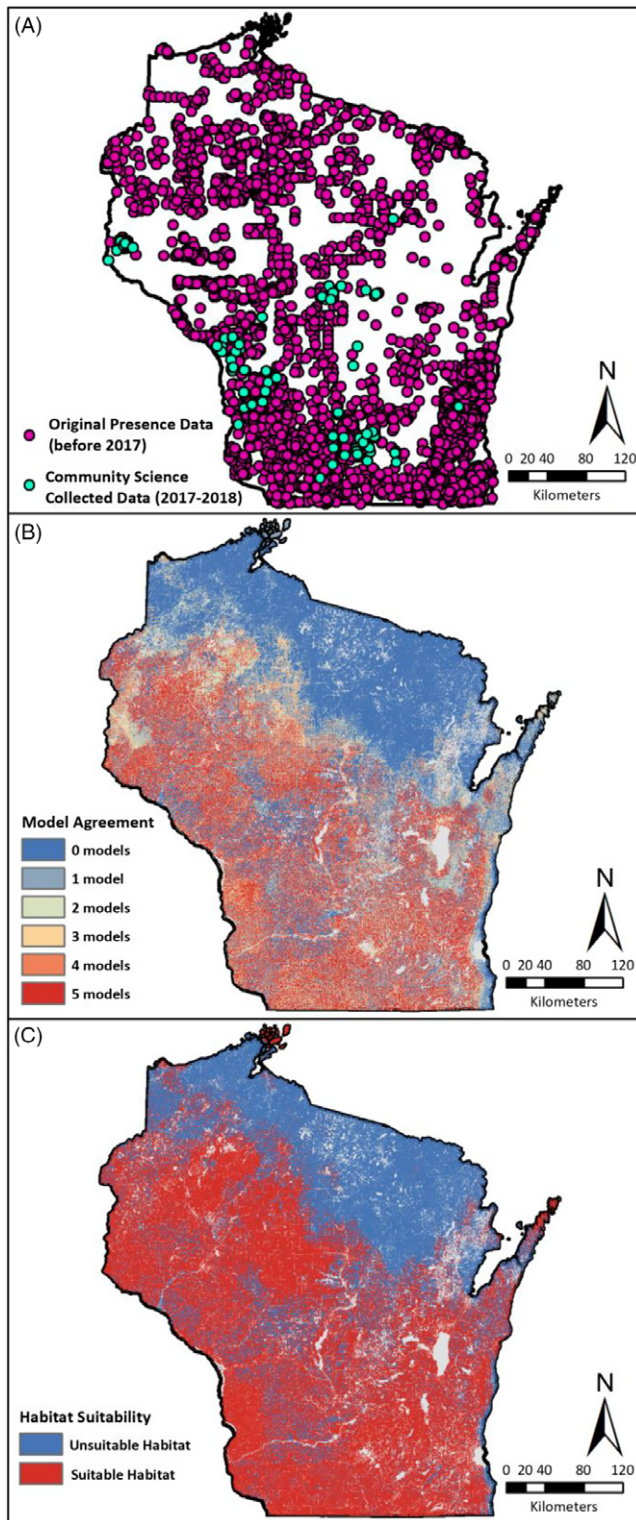


Figure 1. Occurrence records for model development and assessment (A), the binary ensemble habitat suitability model (HSM) created using five model algorithms (B), and a final binary HSM representation (C) of the ensemble model for *Pastinaca sativa* in the state of Wisconsin.

similarly (Iannoni et al. 2015), our results suggest that the drivers are species specific and agree with those of Allen and Bradley (2016), who reported that invasive plant species occupy different niches.

Habitat Suitability Model Performance

All models performed at or exceeded acceptable thresholds for both threshold dependent (TSS) and independent (AUC) evaluation metrics (Supplementary Table S2). AUCs for models were above the minimum threshold (0.9) for excellent performance 62.5% of the time and never below the acceptable threshold (0.7) (Hosmer and Lemeshow 1989). TSS, an indication of how well the models are able to accurately distinguish presence from background, was above the minimum threshold of 0.4 established by Jarnevic et al. (2018) for all but one of the retained model outputs (multivariate adaptive regression splines [MARS]: *Lonicera* spp.). Assessment of AUC and TSS values from the 10-fold cross-validation also indicated good model results, as differences between the training and mean cross-validation testing sets indicated limited overfitting with only one exception. Any models that were considered to be overfitting the training data were tuned appropriately (Supplementary Table S1).

Accuracy of HSMs in Correctly Classifying Community Scientist Observations

Over two growing seasons, an additional 6,035 new occurrence records were reported and verified by experts for the 15 modeled species by 114 reporters across 86% of Wisconsin's counties. Of these new occurrences, 1,030 were excluded from this data set, as they were in "novel" areas where models did not assign a prediction (the majority were waterbodies and roads, which were excluded in model development). This resulted in 5,005 new occurrence records.

We used these 5,005 new data points as an independent data set to evaluate HSMs for each of the 15 species. Occurrences used per species ranged from 48 to 1,291 and were distributed across 9 to 32 of Wisconsin's 72 counties (Table 4). Eight of the 15 models performed at or above an 80% correct classification of the new data points in suitable pixels, with leafy spurge (*Euphorbia esula* L.), wild parsnip (*Pastinaca sativa* L.), and *Torilis* spp. performing at or above 97% correct classification (Table 4). All three of these species' models also performed excellently for AUC and TSS values across all five of the model algorithms. However, autumn olive (*Elaeagnus umbellata* Thunb.), common tansy (*Tanacetum vulgare* L.), and the teasel species' (*Dipsacus* spp.) models performed at less than 60% correct classification of these new data despite great to excellent performance for AUC metrics and above thresholds for other model performance metrics.

Comparing *E. umbellata* and *P. sativa* emphasizes this pattern, as both species performed similarly on AUC and TSS and had several hundred new occurrence records used to validate the ensemble models (454 for *E. umbellata*, and 365 for *P. sativa*). Despite these similarities, the *P. sativa* ensemble correctly classified presence 97% of the time, compared with 56% correct classification by the *E. umbellata* ensemble model.

Simplified models that overpredict suitable habitat could be a reason for high performance on the community scientist observations. The total area deemed suitable for each species differed substantially, with species like teasels (*Dipsacus* spp.) predicting suitability at only 11% of Wisconsin compared with common buckthorn (*Rhamnus cathartica* L.) predicting suitability at 63% of the state (Table 4). Of the species that performed above our 80% threshold, *Lonicera* spp., *E. esula*, *C. stoebe*, and *P. sativa* predicted greater than 50% of Wisconsin's area suitable habitat, but crown-vetch [*Securigera varia* (L.) Lassen], garlic mustard [*Alliaria petiolata* (M. Bieb.) Cavara & Grande], *Torilis* spp.,

Table 3. Number of species models for which each input was within the top three most important variables.^a

Variable	BRT	GLM	MARS	MAXENT	RF	Overall	Percentage
Clay	2	3	3	1	5	14	13.5
Distance to roads	2	4		2	4	12	11.5
Distance to water	1	2	1	1	2	7	6.7
EVI ^b	1	1	–	–	1	3	2.9
Organic matter	–	–	–	–	1	1	1.0
Fall precip.	5	5	6	8	4	28	26.9
Spring precip.	2	6	9	5	2	24	23.1
Summer precip.	13	7	10	11	8	49	47.1
Winter precip.	8	3	6	6	5	28	26.9
Summer T_{max}	9	9	8	8	9	43	41.3
Winter T_{min}	14	11	14	16	14	69	66.3
Tree cover	6	7	6	5	8	32	30.8
Slope	–	1	–	–	–	1	1.0
Aspect	–	–	–	–	–	0	0.0

^aResults are summarized by model algorithm and as a percentage of all models (n = 104). BRT, boosted regression trees; GLM, generalized linear models; MARS, multivariate adaptive regression splines; MaxEnt, maximum entropy; RF, random forest.
^bEVI, enhanced vegetation index.

Table 4. Species used to test the accuracy of community scientist observations (n = 15) and how many of the new points (Figure 1; Supplementary Figures 1–14) were correctly classified as suitable by the habitat models.^a

Common name	Scientific name	% Correctly classified	Total points	No. of reporters	No. of counties reported in	% Area of state suitable
Leafy spurge*	<i>Euphorbia esula</i>	97.2	142	22	9	57.1
Wild parsnip*	<i>Pastinaca sativa</i>	97.0	365	26	15	62.1
Hedge-parsley spp.*	<i>Torilis</i> spp.	96.9	228	16	14	33.7
Spotted knapweed*	<i>Centaurea stoebe</i>	95.9	98	22	22	72.8
Bush honeysuckle spp.*	<i>Lonicera</i> spp.	92.3	1291	19	30	54.9
Crownvetch*	<i>Securigera varia</i>	91.4	151	15	12	36.0
Japanese barberry*	<i>Berberis thunbergii</i>	86.4	674	17	20	33.7
Garlic mustard*	<i>Alliaria petiolata</i>	83.7	604	33	32	32.7
Canada thistle	<i>Cirsium arvense</i>	73.9	329	15	21	57.7
Purple loosestrife	<i>Lythrum salicaria</i>	72.9	48	20	17	46.0
Knotweed spp.	<i>Reynoutria</i> spp.	68.0	50	27	22	38.4
Common buckthorn	<i>Rhamnus cathartica</i>	62.7	465	20	28	58.0
Autumn olive	<i>Elaeagnus umbellata</i>	56.4	454	13	19	46.1
Common tansy	<i>Tanacetum vulgare</i>	52.5	59	12	16	30.6
Teasel spp.	<i>Dipsacus</i> spp.	42.0	50	15	13	10.9

^aAn asterisk (*) indicates species performing above an 80% correct classification threshold per a χ^2 -test (P < 0.05).

and Japanese barberry (*Berberis thunbergii* DC.) predicted $\leq 36\%$ of Wisconsin suitable. While results suggest some level of generalization, with large areas of the state deemed suitable habitat for a given species, this reduction would still result in substantial decreases in monitoring efforts for land managers. If a higher level of reduction is desired, increasing the number of models that agree (we selected one) could be an approach to further reduce area. However, this may increase the potential for false negatives and should be used with caution.

Although referenced as a key component of the modeling process, field validation is not widely observed in the literature (Barbet-Massin et al. 2018). Our data corroborate this notion that useful models must be tested with independently acquired data, as relying on the evaluation metrics (training and cross-validation test) from the model development phase does not correlate with accuracy from newly acquired occurrence data. With only 8 of the 15 models performing at or above 80% correct classification rates, we would recommend further development of the remaining 7 models using the newly collected data as a means to improve model performance (Crall et al. 2013). Below this threshold, these models would likely have limited

applicability for land managers employing EDRR management efforts. End users, however, should set their own expectations for accuracy requirements. We hope that our results provide additional momentum in field validating HSMs, as other have also labeled this as a high priority issue (Sofaer et al. 2019).

Finally, to investigate whether any single metric or set of metrics was important in describing correct classification of community scientist observations, an RF analysis was performed to determine the most important factors driving correct classification (n = 15). The predictors included each of the five model evaluation metrics (AUC, TSS) as well as information about the community scientist data to explore the importance these variables had on predicting suitable habitat. These variables only explained 14.9% of variability in the data. The number of community scientist presence data and the number of total reporters of these new records were the two most important variables (1.0 and 0.81 scaled importance, respectively), all other variables had scaled importance < 0.62. Partial dependence plots (Figure 2) depict the directionality of these predictor variables on the response when all other variables are held constant. For example, new occurrences above 250 and number of reporters above 19 had a dramatic improvement on correct

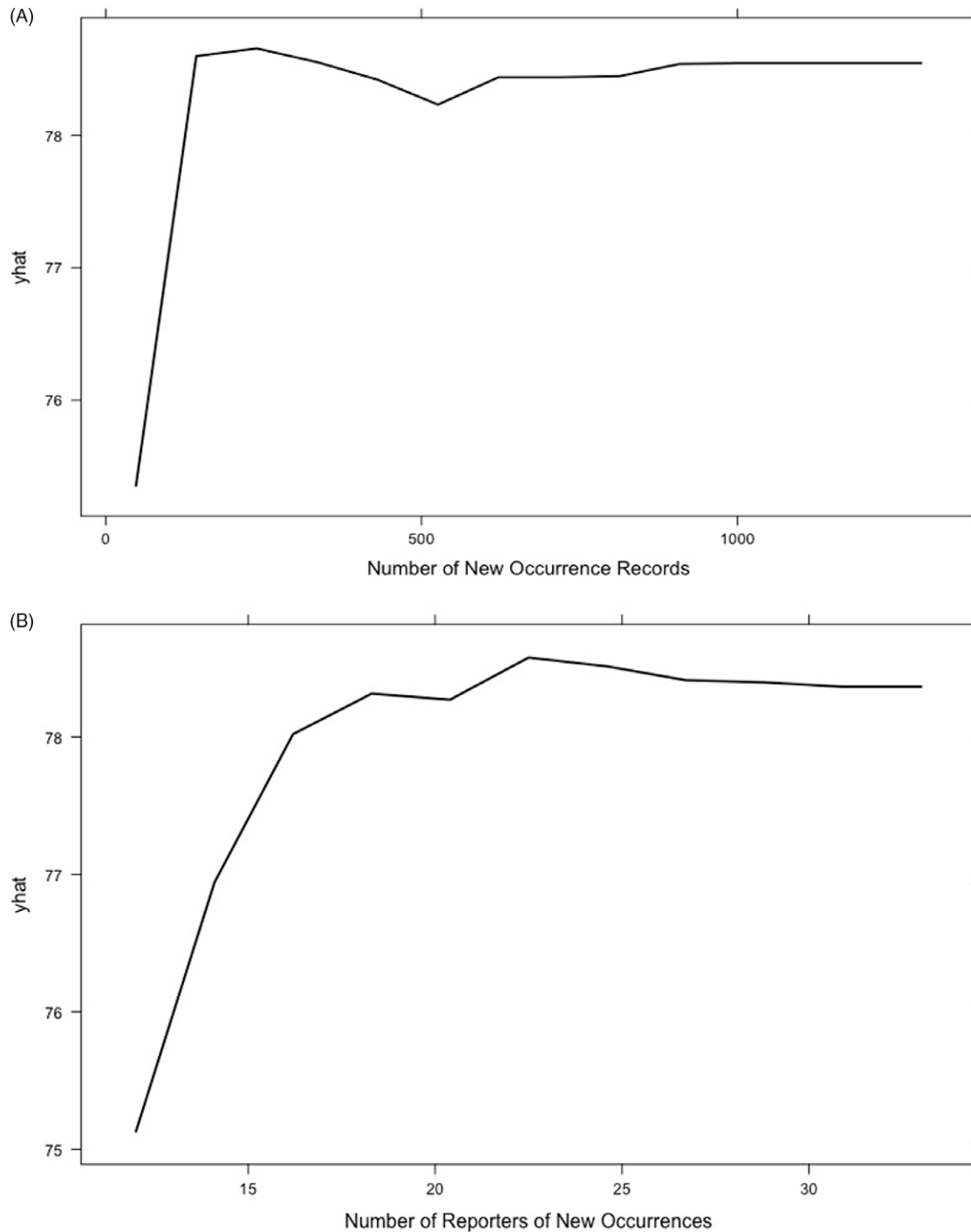


Figure 2. Partial dependence plots (PDPs) for the top two most important variables in a random forest model to determine drivers of accurate classification from community scientist observations of 15 habitat suitability models for individual invasive plant species in Wisconsin. A PDP depicts the marginal impact a single variable has on the result of a machine learning algorithm when all other variables are held at their mean values. The number of independently collected occurrence records (A) not used to train the original habitat models was the most important variable, followed closely by the number of reporters (B) of these new occurrence records.

classification. As these new data came from across the state of Wisconsin, we consider the number of reporters of these data as a proxy for spatial extent but found it surprising that another proxy for spatial extent (number of counties with reports) was not important. Evaluation metrics for models did not have much influence on the overall result. This agrees with our variability in correct classification performance at or above our 80% threshold, despite

relatively similar model evaluation metrics, and suggests that evaluation metrics alone should not be used to assess model performance. Others have suggested that specific model evaluation metrics are a poor indicator of model performance (AUC [Lobo et al. 2008]; TSS [Wunderlich et al. 2019]). Additional research is needed to better understand which metrics are best at predicting model performance in the field.

While robust data sets can help with performance, other factors are likely influencing success/failure of these model performances. More explicit descriptions of the distribution of these new occurrence data (distance to roads or trails, elevation, etc.; Mair and Ruete 2016) could offer better insight into the drivers of successful classification, as spatial bias may occur during the data-collection phase. Regardless, Crall et al. (2013) and Jarnevich et al. (2018) recommend overcoming poor field validation performance by using an iterative approach, whereby models are created and vetted in a development phase, then passed through a field evaluation test, and then updated. Our results support the need for this process when developing HSMs for invasive plants for early detection of new populations, until acceptable results are achieved. Life history, habitats invaded, and other species-specific attributes may also influence results and should be evaluated in the future.

Here we presented 15 ensemble HSMs for a range of invasive plants in the state of Wisconsin. While climate predictor variables consistently were most often the most important variable, the specific variable, directionality, and mixture of other variables differed among models, suggesting modeling should be done on individual species, if possible.

Working with community scientists allowed for the development of a separate data set to confirm performance of models. While all but one species performed at or above acceptable ranges for model evaluation metrics, when community scientist data were tested, just 53% of models could correctly classify these new data at >80%. Known issues of transferability to new environments or environmental conditions (e.g., Latif et al. 2016) or to new locations entirely (e.g., Lauria et al. 2015) that were not used to train the models can produce these low-performing models. Therefore, while not all of our models are ready for deployment in land monitoring/management systems, at least eight exceeded expectations for detection of new populations. Model evaluation metrics (e.g., AUC, TSS) were not a good predictor of accuracy in community scientist observations. This effort highlights the benefits of developing a network of community scientists and stakeholders for HSM efforts. While validation of data submitted and education are required, the additional information made available across wide spatial extents broadened the ability for the scientific community to create useful models that perform well in the field and improve interaction between the scientists and the land managers.

Supplementary Material. To view supplementary material for this article, please visit <https://doi.org/10.1017/inp.2021.27>

Acknowledgments. This material is based upon work support by the National Institute of Food and Agriculture, U.D. Department of Agriculture, Hatch project no. 1006586. We would also like to thank all of the community scientists, invasive species enthusiasts, and professionals who contributed invasive plant occurrences to inform this research, as such a large-scale undertaking would be nearly impossible without such a widespread, educated network. No conflicts of interest have been declared.

References

Allen JM, Bradley BA (2016) Out of the weeds? Reduced plant invasion risk with climate change in the continental United States. *Biol Conserv* 203:306–312

Araújo MB, Anderson RP, Barbosa AM, Beale CM, Dormann CF, Early R, García RA, Guisan A, Maiorano L, Naimi B, O'Hara RB (2019) Standards for distribution models in biodiversity assessments. *Sci Adv* 5:4858

Barbet-Massin M, Rome Q, Villemant C, Courchamp F (2018) Can species distribution models really predict the expansion of invasive species? *PLoS ONE* 13:e0193085

Cook G, Jarnevich C, Warden M, Downing M, Withrow J, Leinwand I (2019) Iterative models for early detection of invasive species across spread pathways. *Forests* 10(2):108

Crall AW, Jarnevich CS, Panke B, Young N, Renz M, Morissette J (2013) Using habitat suitability models to target invasive plant species surveys. *Ecol Appl* 23:60–72

Crall AW, Renz M, Panke BJ, Newman GJ, Chapin C, Graham J, Bargeron C (2012) Developing cost-effective early detection networks for regional invasions. *Biol Invasions* 14:2461–2469

Cutler DR, Edwards TC Jr, Beard KH, Cutler A, Hess KT, Gibson J, Lawler JJ (2007) Random forests for classification in ecology. *Ecology* 88:2783–2792

Didan K (2015) MOD13Q1 MODIS/Terra Vegetation Indices 16-Day L3 Global 250m SIN Grid V006 [Data set]. NASA EOSDIS Land Processes DAAC. <https://doi.org/10.5067/MODIS/MOD13Q1.006>. Accessed: April 15, 2016

DiTomaso JM (2000) Invasive weeds in rangelands: species, impacts, and management. *Weed Sci* 48:255–265

Dukes JS, Mooney HA (2004) Disruption of ecosystem processes in western North America by invasive species. *Rev Chil Hist Nat* 77:411–437

Engler R, Waser LT, Zimmermann NE, Schaub M, Berdos S, Ginzler C, Psomas A (2013) Combining ensemble modeling and remote sensing for mapping individual tree species at high spatial resolution. *For Ecol Manag* 310:64–73

Epanchin-Niell RS, Hufford MB, Aslan CE, Sexton JP, Port JD, Waring TM (2010) Controlling invasive species in complex social landscapes. *Front Ecol Environ* 8:210–216

Evangelista PH, Kumar S, Stohlgren TJ, Jarnevich CS, Crall AW, Norman JB III, Barnett DT (2008) Modelling invasion for a habitat generalist and a specialist plant species. *Divers Distrib* 14:808–817

Freire J, Silva CT (2012) Making computations and publications reproducible with VisTrails. *Comput Sci Eng* 14(4):18–25

Funk JL, Parker IM, Matzek V, Flory SL, Aschehoug ET, D'Antonio CM, Dawson W, Thomson DM, Valliere J (2020) Keys to enhancing the value of invasion ecology research for management. *Biol Invasions* 22:2431–2445

Hosmer DW, Lemeshow S (1989) *Applied Logistic Regression*. New York: Wiley. 307 p

Iannone BV III, Oswalt CM, Liebhold AM, Guo Q, Potter KM, Nunez-Mir GC, Oswalt SN, Pijanowski BC, Fei S (2015) Region-specific patterns and drivers of macroscale forest plant invasions. *Divers Distrib* 21:1181–1192

Jarnevich CS, Hayes MA, Fitzgerald LA, Yackel Adams AA, Falk BG, Collier MAM, Bonewell LR, Klug PE, Naretto S, Reed RN (2018) Modeling the distributions of tegu lizards in native and potential invasive ranges. *Sci Rep* 8:10193

Jorgensen N, Leary J, Renz M, Mahnken B (2021) Characterizing the suitable habitat of *Miconia calvescens* in the East Maui Watershed. *Manage Biol Inv* 12:313–330

Kumar S, Stohlgren TJ, Chong GW (2006) Spatial heterogeneity influences native and nonnative plant species richness. *Ecology* 87:3186–3199

Latif QS, Saab VA, Hollenbeck JP, Dudley JG (2016) Transferability of habitat suitability models for nesting woodpeckers associated with wildfire. *The Condor: Ornithol Appl* 118:766–790

Lauria V, Power AM, Lordan C, Weetman A, Johnson MP (2015) Spatial transferability of habitat suitability models of *Nephrops norvegicus* among fished areas in the Northeast Atlantic: sufficiently stable for marine resource conservation? *PLoS One* 10:e0117006

Levine JM, Vila M, Antonio CM, Dukes JS, Grigulis K, Lavorel S (2003) Mechanisms underlying the impacts of exotic plant invasions *Proc R Soc London B* 270:775–781

Lobo JM, Jiménez-Valverde A, Real R (2008) AUC: a misleading measure of the performance of predictive distribution models. *Global Ecol Biogeogr* 17:145–151

Lötter D, Maitre D le (2014) Modelling the distribution of *Aspalathus linearis* (Rooibos tea): implications of climate change for livelihoods dependent on both cultivation and harvesting from the wild. *Ecol Evol* 4:1209–1221

Magarey R, Newton L, Hong SC, Takeuchi Y, Christie D, Jarnevich CS, Kohl L, Damus M, Higgins SI, Millar L, Castro K (2018) Comparison of four modeling tools for the prediction of potential distribution for non-indigenous weeds in the United States. *Biol Invasions* 20:679–694

Mair L, Ruete A (2016) Explaining spatial variation in the recording effort of citizen science data across multiple taxa. *PLoS One* 11:e0147796

- Maistrello L, Dioli P, Bariselli M, Mazzoli GL, Giacalone-Forini I (2016) Citizen science and early detection of invasive species: phenology of first occurrences of *Halyomorpha halys* in Southern Europe. *Biol Invasions* 18:3109–3116
- Maxwell BD, Backus V, Hohmann MG, Irvine KM, Lawrence P, Lehnhoff EA, Rew LJ (2012) Comparison of transect-based standard and adaptive sampling methods for invasive plant species. *Invasive Plant Sci Manag* 5:178–193
- Mazza G, Tricarico E, Genovesi P, Gherardi F (2014) Biological invaders are threats to human health: an overview. *Ethol Ecol Evol* 26:112–129
- Morisette JT, Jarnevich CS, Holcombe TR, Talbert CB, Ignizio D, Talbert MK, Silva C, Koop D, Swanson A, Young NE (2013) VisTrails SAHM: visualization and workflow management for species habitat modeling. *Ecography* 36:129–135
- [NISC] National Invasive Species Council (2018) National Invasive Species Council Crosscut Budget. doi: <https://www.doi.gov/invasivespecies/nisc-resources>
- Pearson RG (2007) Species' distribution modeling for conservation educators and practitioners. *Synthesis. Lessons in Conservation* 3:54–89
- Phillips SJ, Dudík M, Elith J, Graham CH, Lehmann A, Leathwick J, Ferrier S (2009) Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. *Ecol Appl* 19:181–197
- Probst P, Wright MN, Boulesteix AL (2019) Hyperparameters and tuning strategies for random forest. *Data Min Knowl Disc* 9:e1301
- R Core Team (2019) R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. <https://www.R-project.org>
- Reaser JK, Burgiel SW, Kirkey J, Brantley KA, Veatch SD, Burgos-Rodríguez J (2019) The early detection of and rapid response (EDRR) to invasive species: a conceptual framework and federal capacities assessment. *Biol Invasions* 22:1–9
- Renz M, Gibson KD, Hillmer J, Howe KM, Waller DM, Cardina J (2009) Land manager and researcher perspectives on invasive plant research needs in the midwestern United States. *Invasive Plant Sci Manag* 2:83–91
- Rothlisberger JD, Chadderton WL, McNulty J, Lodge DM (2010) Aquatic invasive species transport via trailered boats: what is being moved, who is moving it, and what can be done. *Fisheries* 35:121–132
- Sexton JO, Song XP, Feng M, Noojipady P, Anand A, Huang C, Kim DH, Collins KM, Channan S, DiMiceli C, Townshend JRG (2013) Global, 30-m resolution continuous fields of tree cover: Landsat-based rescaling of MODIS Vegetation Continuous Fields with lidar-based estimates of error. *Int J Digit Earth* 6:427–448
- Sofaer HR, Jarnevich CS, Pearse IS, Smyth RL, Auer S, Cook GL, Edwards TC Jr, Guala GF, Howard TG, Morisette JT, Hamilton H (2019) Development and delivery of species distribution models to inform decision-making. *BioScience* 69:544–557
- Stohlgren TJ, Barnett DT, Kartesz JT (2003) The rich get richer: patterns of plant invasions in the United States. *Front Ecol Environ* 1:11–14
- Wang O, Zachmann LJ, Sesnie SE, Olsson AD, Dickson BG (2014) An iterative and targeted sampling design informed by habitat suitability models for detecting focal plant species over extensive areas. *PLoS One* 9:e101196
- Wang T, Hamann A, Spittlehouse D, Carroll C (2016) Locally downscaled and spatially customizable climate data for historical and future periods for North America. *PLoS One* 11:e0156720
- West AM, Kumar S, Brown CS, Stohlgren TJ, Bromberg J (2016) Field validation of an invasive species Maxent model. *Ecol Inf* 36:126–134
- Westbrooks RG (2004) New approaches for early detection and rapid response to invasive plants in the United States. *Weed Technol* 18:1468–1471
- [WIFDN] Wisconsin First Detector Network (2020) Home page. <https://fyi.extension.wisc.edu/wifdn>. Accessed: June 20, 2020
- [WISC] Wisconsin Invasive Species Council (2013) Statewide Plan for Invasive Species (2013–2016). Madison, WI: Wisconsin Department of Natural Resources Bureau of Science Services, PUB-SS-1107 2013. 18 p
- Wunderlich RF, Lin YP, Anthony J, Petway JR (2019) Two alternative evaluation metrics to replace the true skill statistic in the assessment of species distribution models. *J Nat Conserv* 35:97
- Young NE, Jarnevich CS, Sofaer HR, Pearse I, Sullivan J, Engelstad P, Stohlgren TJ (2020) A modeling workflow that balances automation and human intervention to inform invasive plant management decisions at multiple spatial scales. *PLoS One* 15:e0229253