Aquatic Botany 151 (2018) 43–50

Contents lists available at ScienceDirect

Aquatic Botany

journal homepage: www.elsevier.com/locate/aquabot

Prediction of starry stonewort (*Nitellopsis obtusa*) invasion risk in upper Midwest (USA) lakes using ecological niche models

Ranjan Muthukrishnan[a,*], Robin S. Sleith[b], Kenneth G. Karol[b], Daniel J. Larkin[a]

[a] Department of Fisheries, Wildlife and Conservation Biology and Minnesota Aquatic Invasive Species Research Center, University of Minnesota-Twin Cities, St. Paul, MN, 55108, USA

[b] Lewis B. and Dorothy Cullman Program for Molecular Systematics, The New York Botanical Garden, Bronx, NY, 10458, USA

ARTICLE INFO

Keywords:
Ecological niche model
Invasive species
Lake
Random forests
Boosted regression trees
Ecological niche factor analysis

ABSTRACT

Starry stonewort (*Nitellopsis obtusa*; Characeae) is a freshwater macroalga that is considered an invasive species in North America and has only recently been identified in the upper Midwest states of Minnesota and Wisconsin (USA). While the current known extent of *N. obtusa* invasion in the Midwestern U.S. is limited, there is significant potential habitat for continued expansion and thus a pressing need to target surveillance and response efforts to limit further spread. Here we use data on *N. obtusa* presence and lake-level environmental conditions from locations in New York state to train a set of ecological niche models using three separate algorithms: random forests, boosted regression trees, and ecological niche factor analysis. These models were then used to predict habitat suitability and potential invasion risk for a set of ~900 lakes in the upper Midwest using publicly available lake-level water quality data. Based on a cross-validation study we found that the random forest method provided the most accurate predictions, though only marginally better than boosted regression trees. Ecological niche factor analysis, while offering better than random predictions, had the highest cross-validation error rates. Though there were some inconsistencies between modeling approaches, all three tended to agree on regions of relatively high risk in central Minnesota and eastern Wisconsin and relatively low risk in north-central Wisconsin. However, there are inherent limitations to developing ecological niche models with data from one geographical region and predicting into a different region, thus there is a need for additional efforts to validate model predictions.

1. Introduction

The spread of invasive species is a major driver of global ecological change with significant impacts on biodiversity, ecosystem services, and human use of natural systems (Brook et al., 2008; Pyšek et al., 2012; Vilà et al., 2011). Priorities for invasive species management include preventing new invasions into susceptible habitats and focusing control efforts on recently established populations, i.e., early detection and rapid response (Mack et al., 2000; Simberloff, 2003). Unfortunately, early detection of invasive species is difficult and new invasions are often only identified after populations are well-established (Mehta et al., 2007). Management efforts also require significant effort and resources, making broad, non-targeted efforts difficult to justify. Thus there is a need to predict invasion risk to help prioritize efforts to identify and control new populations (Lodge et al., 2006).

Starry stonewort (*Nitellopsis obtusa*; Characeae) is a freshwater green macroalga considered invasive in North America, where it has spread throughout the Great Lakes basin and beyond since its discovery in the St. Lawrence River in the mid-1970s (Geis et al., 1981; Karol and Sleith, 2017; Larkin et al., 2018; Midwood et al., 2016; Sleith et al., 2015). It was recently recorded for the first time in Wisconsin (2014) and Minnesota (2015). But some infestations in those states are extensive enough to suggest that *N. obtusa* established several years before it was identified. *Nitellopsis obtusa* can grow in dense beds up to 2-m tall and in depths down to 10 m (Simons and Nat, 1996; Sleith et al., 2015), producing nuisance aggregations that impair recreational activities such as boating, swimming, and fishing. Additionally, *N. obtusa* can decrease native plant diversity, which in turn could have further ecological consequences (Brainard and Schulz, 2016).

Lakes are both a major locus of recreational activity and an important cultural touchstone for the upper Midwest; thus the spread of aquatic invasive species can be rapid and consequential. In response to the recent spread of *N. obtusa*, natural resource agencies have focused significant attention and resources on surveillance and control efforts.

* Corresponding author.
E-mail address: mrunj@umn.edu (R. Muthukrishnan).

https://doi.org/10.1016/j.aquabot.2018.08.001
Received 11 January 2018; Received in revised form 25 July 2018; Accepted 1 August 2018
Available online 17 August 2018

The number of verified invaded lakes has grown quickly since *N. obtusa* became a regional priority, likely due in substantial part to increased awareness and detection efforts and not simply new infestations. Currently, 11 lakes in Minnesota and 14 waterbodies in Wisconsin (including multiple nearshore locations in Lake Michigan) have known populations of *N. obtusa*. There are also undoubtedly additional lakes in this region where *N. obtusa* has already established but not yet been identified, precluding opportunities to limit transmission through public awareness campaigns, increased inspection resources, or other interventions. Only male individuals of *N. obtusa* have been identified in North America, implicating human movement of fragments and bulbils (asexual reproductive structures) in its spread rather than waterbird movement of zygotes (or oospores) or other natural pathways (Sleith et al., 2015).

Ecological niche modeling (ENM) is well-established and commonly used for predicting changing species distributions under different future scenarios, such as climate change (Pearson and Dawson, 2003; Thuiller et al., 2005). ENM is also increasingly used to predict potential habitat for invasive species (Jiménez-Valverde et al., 2011; Peterson, 2003), though there are notable challenges to this approach (Hirzel et al., 2002). In most applications of ENMs to invasive species, broad-scale bioclimatic variables have been used to estimate ecological niches, but for aquatic invasive species, these bioclimatic variables may not be the most important parameters. For example, lakes buffer aquatic invasive species from much of the temperature and precipitation variability described by bioclimatic variables (Kriticos et al., 2012), limiting exposure to drought, freezing, and other stressors that constrain distributions of terrestrial species. Furthermore, models based on bioclimatic variables provide very broad spatial predictions (e.g., Escobar et al., 2016) that may have limited value for geographically targeting interventions. Therefore, environmental variables at the scale of individual water bodies (e.g., water chemistry variables) may be more informative, but such data are often difficult to acquire and are rarely used in niche modeling (Kilroy et al., 2008; Kumar et al., 2009). However, if sufficient and appropriate data can be acquired, several ENM algorithms can be targeted to discrete geographic features, such as lakes.

A variety of niche modeling approaches have been developed, with tradeoffs in terms of model complexity, accuracy, interpretability, and computational efficiency. A major distinction lies between approaches that use empirical data of both presences and absences and those that only incorporate presence data, which are generally more reliable than absence data. For active invasions in particular, species distributions are fundamentally non-equilibrial, violating a key assumption of many ENM methods (Hirzel et al., 2002). If an invasive species has not yet had the opportunity to fully explore a region, treating locations without the species as true absences—and thus as poor-quality habitat—may skew predictions. Presence-only ENM approaches, such as ecological niche factor analysis (ENFA; Hirzel et al., 2002), have been developed to address this issue by focusing only on presences that positively indicate suitable habitat and remaining agnostic about the suitability of locations where a species has not been observed.

Alternative approaches that incorporate both presence and absence data are often built using decision trees, which can estimate complex, non-linear effects of multiple predictor variables on suitability through a series of nested bifurcations (Breiman et al., 1984; De'ath and Fabricius, 2000). Using a dataset with known outcomes, samples are split based on single predictors into the most homogenous groups possible. By recursively continuing this splitting procedure, an optimal set of rules can be determined to predict outcomes. The simplest form of this approach is classification and regression trees (CART; Breiman et al., 1984), which use an entire dataset to create a single optimal tree for prediction. While this approach allows for non-linear relationships and is highly interpretable, CART models tend to have low predictive accuracy relative to other methods. Newer approaches, such as random forests (RF; Breiman, 2001) and boosted regression trees (BRT; Elith et al., 2008) build upon the basic framework of decision trees, but differ from CART in that they generate many decision trees using subsets of the data and then employ different strategies for integrating the results of individual trees to provide more refined predictions.

All ENM approaches require location data for known presences (and potentially absences) and corresponding environmental data to train the model, but acquiring such data can be challenging for incipient invasions. For example, predicting habitat suitability for *N. obtusa* in Minnesota and Wisconsin based on occurrence records from those states is problematic because there are relatively few known invaded lakes in those states with which to train an ENM. The longer history of *N. obtusa* invasion and extensive search efforts in New York (Sleith et al., 2018, 2015) provide an alternative data source with which to develop an ENM. This has the added benefit of allowing for stronger independence of training data (New York) and test data (Minnesota and Wisconsin), which can strengthen model robustness (Houlahan et al., 2017).

We used ENMs to predict vulnerability of individual lakes to *N. obtusa* invasion on the western edge of known *N. obtusa* distribution (Minnesota and Wisconsin). We compiled environmental data from a variety of sources, including direct measurements taken during surveys as well as publicly available data collected by state agencies and citizen scientists in New York, Minnesota, and Wisconsin. We then trained models using *N. obtusa* presence and absence data from systematic surveys conducted throughout New York using RF, BRT, and ENFA algorithms. Lastly, we used these models to predict lake-level invasion risk in Minnesota and Wisconsin and evaluated models' predictive ability based on known occurrences in these states.

2. Methods

2.1. Species distribution data

In a previous study, Sleith et al. (2018) searched for *N. obtusa* at 390 locations throughout New York (as part of a larger study throughout New York and New England) during field surveys in 2014. Water bodies to be surveyed were randomly chosen by identifying the closest public boat launch to a set of randomly stratified points throughout the region. At each access point a combination of wading and tossing a dredge were used to sample Characeae populations. Most *N. obtusa* presences were aggregated in north central New York. To limit the potential for including "false absences" from lakes where *N. obtusa* has not yet had a chance to colonize, we only included as absences lakes within a geographic region where *N. obtusa* was recorded and in which other Characeae species were found, i.e., characean-supporting lakes in an *N. obtusa*-invaded region (providing a dataset of 25 lakes with *N. obtusa* present and 29 lakes where it was absent).

2.2. Environmental data

We collected data on lake environmental conditions from a variety of sources. For New York sampling locations, we collected water samples during field surveys (Sleith et al., 2018) and analyzed samples for ammonium and nitrate (which we summed as total inorganic N) and total dissolved phosphorus. We also directly measured conductivity and pH at each site using an In Situ SmarTroll MP (Ft. Collins, CO, U.S.A). In addition, we compiled records from the Citizens Statewide Lake Assessment Program (CSLAP) for water clarity (Secchi depth) and chlorophyll *a* concentrations (chlA) from sampling locations. Not all survey sites had corresponding water clarity and chlorophyll *a* measurements. For locations missing these data (approximately half of sites), we imputed Secchi depth and chlorophyll *a* levels based on four measured water chemistry variables (pH, conductivity, N, and P) using generalized linear models (GLM). We estimated statistical models using the glm and predict.glm functions from the stats package and using the quasi-poisson family to constrain distributions to only positive values using R version 3.1.2 (which was used for all statistical and modeling analyses;

R Core Team, 2014).

To estimate environmental parameters for lakes in the upper Midwest (i.e., Minnesota and Wisconsin), we aggregated data from three publicly available sources. We collected measurements of long-term average Secchi depth for ~11,000 lakes in Minnesota derived from remote sensing data (Olmanson et al., 2014). Secondly, we accessed a large dataset of direct lake measurements (~6 million records) collected by a variety of State, local, and citizen-based organizations, and on a wide variety of environmental parameters, managed by the Minnesota Pollution Control Agency. Lastly, we downloaded data for Wisconsin lakes from the Surface Water Integrated Water Monitoring System of the Wisconsin Department of Natural Resources. From these datasets we aggregated data for six environmental parameters that are likely to influence macrophyte distributions and that were sampled in a large number of lakes: pH, conductivity, inorganic nitrogen, total dissolved phosphorus, chlorophyll *a* concentration, and Secchi depth.

Data were heterogeneous in space and time and collected by a wide variety of groups, thus we took several steps to assure data quality. We limited environmental measures to only those collected since the year 2000 and during the growing season (June through September). We also excluded hypolimnetic samples. To identify outlier data, we calculated the mean and standard deviations for each environmental variable across all lakes and excluded any samples that had values > 5 standard deviations from the mean—values extreme enough to suggest that they were erroneous. Because standard deviations were also sometimes strongly influenced by extreme outliers, we recalculated standard deviations and repeated the process a second time. We then averaged all measurements for a given environmental parameter from a lake and identified lakes with data for all six environmental variables. This yielded a dataset of 920 lakes: 692 in Minnesota and 228 in Wisconsin.

2.3. Niche model parameterization

Using the data from New York, we developed models to estimate niche preferences of *N. obtusa* (all data and code used for ENMs are included in supplementary material). For two methods, RF and BRT, we used presence and absence data to parameterize suitability models using all six environmental parameters. These models were then used to predict the probability that additional locations would be classified as presence locations. Rather than attempting to identify a single optimal decision tree, RF produces a large number of decision trees where, at each branching point, potential variables that could be used to split the data are a randomly selected subset of all potential variables (Breiman, 2001; Cutler et al., 2007). As a result, any given tree is unlikely to provide a best fit for the entire dataset, but by classifying data based on the summed "votes" across all trees, better predictions are generally produced. The RF algorithm was implemented with the randomForest function (from the randomForest package; Liaw and Wiener, 2002) using default parameters, except that we estimated 1500 trees (tree number was selected using a cross-validation procedure similar to the one described below).

Use of BRT allows for an iterative machine-learning process where an initial decision tree is first constructed, and then additional trees are built to augment areas where the original tree predicted poorly (Elith et al., 2008, 2006). This process is repeated and a composite model is constructed from a large number of trees. This method allows for non-linear relationships, variable weighting of different parameters, and identification of complex interactions. We implemented BRT using the gbm.step function (from the gbm package; Ridgeway, 2015) following the approach of Elith et al. (2008). This required tuning of three algorithm parameters: learning rate, tree complexity, and bagging fraction. We used an *ad-hoc* approach to tuning parameters to identify a parameter set that consistently produced low cross-validation deviance. This led us to select a learning rate of 0.001, tree complexity of 25, and a bagging fraction of 0.67 and produced a composite model with 1530 trees.

The third method we used was ENFA (Hirzel et al., 2002). This method uses only data from known presence locations to estimate the niche by comparing, for each environmental parameter, the distribution of conditions in locations where the species is present against the conditions across the entire potential range. For each parameter, the range of utilized conditions relative to the complete range (specialization) and the difference between the average condition of utilized locations and all locations (marginality) are calculated (Basille et al., 2008). With these data, a centroid delimiting the optimal conditions for a species in multivariate space is determined and the multivariate (Mahalanobis) distance from that centroid can be calculated for any given location as a measure of relative suitability (Calenge et al., 2008). We conducted ENFA for *N. obtusa* using the enfa function (from the adehabitatHS package; Calenge, 2006), with presence data from New York, but environmental data from both New York and the entirety of the Midwest dataset in order to predict into the entire range. Because this process generates distances from optimal conditions for locations rather than explicit probabilities of presence, numeric values of outputs from ENFA cannot be directly compared to the other methods. However, relative risks among locations are internally consistent and were evaluated analogously across all methods.

2.4. Cross-validation for model comparison

We conducted a cross-validation analysis to evaluate the relative accuracy of the different approaches. For this evaluation, we constrained both our training and prediction datasets to the New York data. We randomly selected 75% of New York lakes (of the 54 within the invasion region, including both presence and absence lakes) to train all three models, with the other 25% withheld to test model predictions. To calculate a metric that would apply consistently across methods, we used each model to predict presence/absence in each of the withheld lakes and then calculated error rate. For RF and BRT, we treated location probabilities > 0.5 as classifications of presence. Because ENFA does not provide explicit probabilities, we incorporated an additional classification step: using estimated Mahalanobis distances for each of the training data points, we calculated error rates for a range of distances that could be used as cut-off values to identify presence classifications. We then identified the smallest cut-off distance with the lowest error rate as the optimal cut-off. We applied that threshold to the distance estimates for the withheld data to determine if they would be classified as presence or absence locations.

This process was repeated for 100 folds of the data, each with a random selection of 75% training and 25% testing data. For each fold, the error rate was calculated for each model, as well as the individual error rates for presence and absence locations. Then average error rates and class errors were calculated across all folds.

2.5. Evaluation of relative importance of model parameters

We next calculated the relative influence of each environmental parameter on ENM models, using approaches adapted to each algorithm. For RF we evaluated relative importance using the mean decrease in accuracy when values for a given parameter were randomly permuted. This functionally transforms the variable into noise, and values were extracted with the importance function from the randomForest package. Values were then scaled so they summed to 100 for easier comparison. For BRT, relative influences were calculated using a similar permutation method, the summary.gbm function using the "permutation.test.gbm" method. Values were again scaled so they summed to 100. For ENFA there was no equivalent method for calculating relative influence. However, a graphical representation of relative importance was created with a biplot of the marginality and first specialization axis of the ENFA model with vectors depicting the loadings of each of the environmental parameters (Basille et al., 2008).

Such biplots show both the entire range of potential habitat and the subset used by the species; the shape of the used habitat relative to loadings of the different environmental parameters can be used as an indicator of their relative importance.

2.6. Invasion risk prediction

Once models were parameterized, we estimated risks for the 920 upper Midwest lakes for which we had measurements of all six environmental parameters. For RF and BRT, we applied models to the Midwest environmental dataset using the predict.randomForest and predict.gbm functions, respectively. To acquire probabilities of classification as a presence location, the "type" argument for the predict calls was "prob" for RF and "response" for BRT. These probabilities were then directly interpreted as relative suitability. Habitat suitability based on ENFA was estimated using the predict.enfa function, which provides the Mahalanobis distance from the multivariate centroid of the conditions observed at the presence locations. This distance can be directly interpreted as a measure of suitability, but we further transformed these values to aid interpretation. Lower distances indicate better habitat, but quality generally decays rapidly and there is no upper limit to distance values. Thus we calculated log of the distance (+1 to account for zero distances) and subtracted distances from the maximum distance. This resulted in more-suitable habitat having higher values, allowing for more intuitive comparisons with the other methods.

3. Results

Lakes in Minnesota and Wisconsin had much broader ranges of environmental conditions compared to New York lakes (Fig. 1). This is partly due to the much larger number of Midwestern lakes included in the analysis. Conditions observed in New York lakes were generally still encompassed within the ranges of the Midwestern lakes. Though in some cases, particularly Secchi depth and chlorophyll *a* concentration, the New York values tended to be near the edge of the Midwestern distribution.

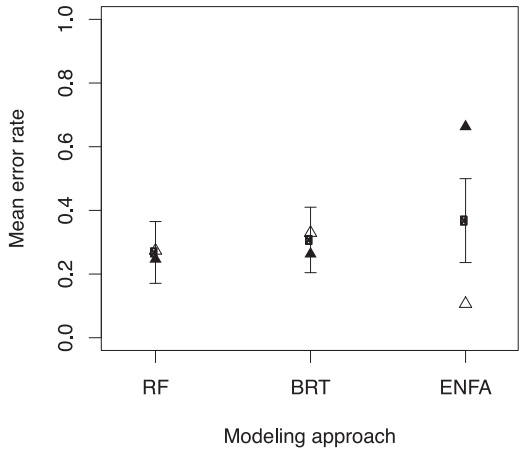

Fig. 2. Comparison of accuracy of different algorithms based on cross-validation analyses of New York data. Circles with error bars show the mean ± standard deviation of the overall error rate while the closed and open triangles show mean error for presence locations (false positives) and absence locations (false negatives), respectively.

Comparison of model predictions for New York showed that all approaches were informative in that they outperformed random predictions. However, there remained significant unexplained variance, with error rates averaging ~30% (Fig. 2). RF produced the most accurate predictions, though these were only marginally better than BRT. Overall error rate for ENFA was relatively close to RF and BRT, but there was much greater disparity between error rates for false positive

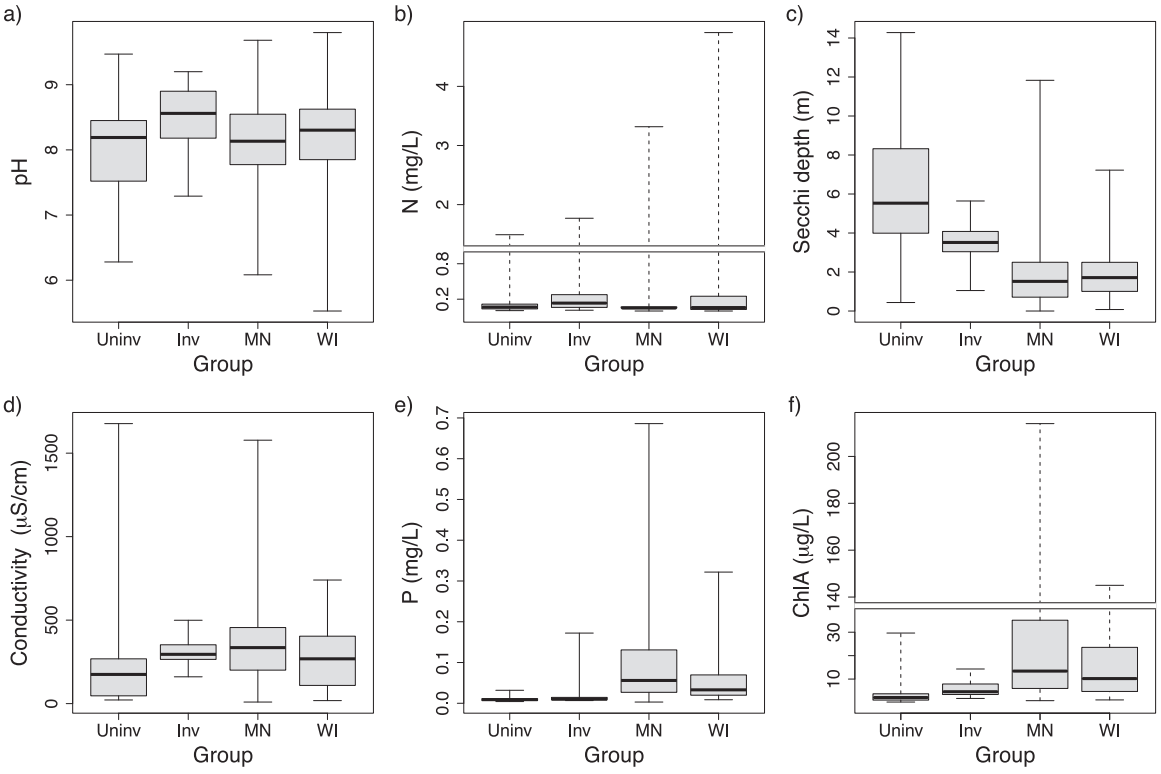

Fig. 1. Observed ranges of environmental parameters in both invaded ("Inv") and uninvaded ("Uninv") lakes in New York and in lakes from Minnesota and Wisconsin where predictions were made. Different panels display ranges for each of the parameters considered: a) pH, b) inorganic nitrogen, c) Secchi depth, d) conductivity, e) total dissolved phosphorus, and f) chlorophyll *a* concentration. Note the broken axes and different scales in the upper and lower portions of panels b and f used to improve visual clarity.

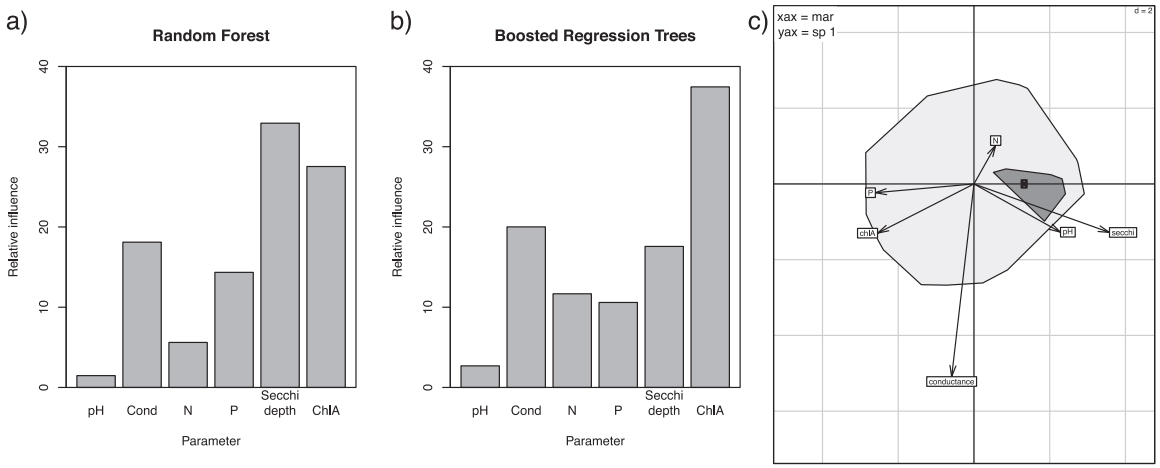

Fig. 3. Analysis of the relative influence of each environmental parameter in each modeling algorithm. The relative influence of each parameter (scaled to a maximum of 100) were explicitly calculated for random forest (a) and boosted regression trees (b). For ENFA (c) a biplot is shown based on the marginality (x-axis) and first axis of specialization (y-axis) that describes the environmental conditions seen across all lakes (light gray polygon) and the invaded New York lakes (dark grey polygon), with a white circle indicating the centroid of the conditions of the invaded lakes. The loadings of each environmental parameter on those two axes are plotted with arrows.

and false negative predictions, indicating that ENFA over-predicted presences.

Broadly speaking, there was consilience among models in the relative importance of parameters that determined the environmental niche for *N. obtusa*. Calculation of the relative influence of different parameters for RF and BRT both identified chlorophyll *a* concentration, conductivity, and Secchi depth as the strongest predictors of invader presence (Fig. 3a,b). Those parameters also showed strong loading on ENFA primary axes (Fig. 3c; Secchi depth and chlorophyll *a* on the marginality axis and conductivity on the first axis of specialization). The models showed intermediate, somewhat inconsistent importance for nitrogen and phosphorus levels, with BRT and ENFA identifying greater importance of nitrogen and RF placing more emphasis on phosphorus. All models were consistent in identifying pH as a relatively uninformative parameter.

Predictions of invasion risk using environmental variables in Minnesota and Wisconsin showed wide ranges in estimated lake-level suitability for *N. obtusa* across all models, indicating substantial model discrimination in habitat quality (Fig. 4). Risk predictions across models were aligned for some regions, but inconsistent in others (Fig. 5). Predictions tended to converge for higher-risk lakes and be more disparate at lower risk levels. The least consistent predictions were observed in southwestern Minnesota, which BRT indicated as high risk, ENFA indicated as low risk, and RF indicated as intermediate risk. A smaller region in south-central Wisconsin also showed similar inconsistency in predictions. Across all models, relatively high risk was predicted for areas in central Minnesota and eastern Wisconsin and relatively low risk was predicted for north-central Wisconsin. This pattern generally aligns with where *N. obtusa* has been found to date, though we caution that current knowledge of *N. obtusa* distribution is limited and likely incomplete. Locations where *N. obtusa* has been identified in Minnesota and Wisconsin are also generally in the upper portions of the distributions of our risk predictions for BRT and RF (Fig. 6a–b), providing support for model validity, however ENFA had less accurate predictions (Fig. 6c).

4. Discussion

Ecological niche models are typically based on broad-scale climatic or topographic factors and are rarely used for predictions within smaller landscape features such as lakes (though see for example Kulhanek et al., 2011). By using publicly available datasets for multiple environmental parameters, we were able to adapt standard ENM methods to predict lake-level habitat suitability and, by extension, invasion risk for an incipient invasion of the characean macroalga *N. obtusa*. We used multiple ENM approaches and, despite some

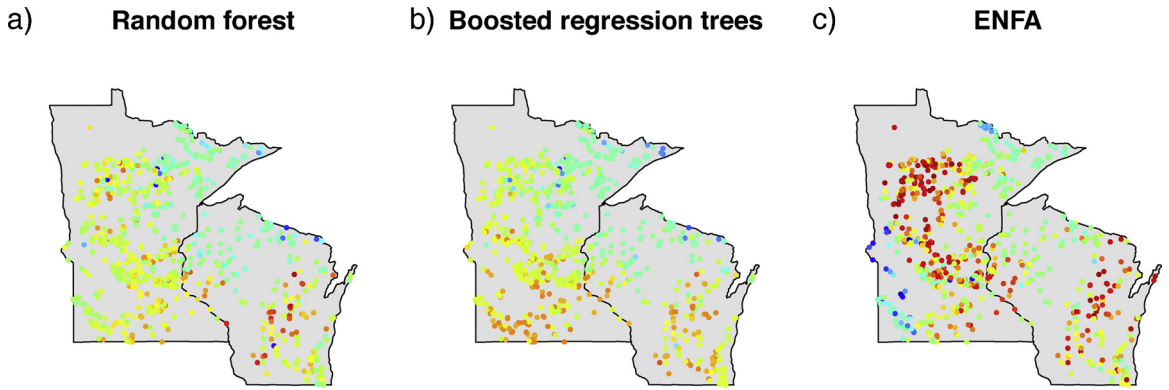

Fig. 4. Predicted suitability for *N. obtusa* of individual lakes across Minnesota and Wisconsin. Colors indicate relative risk from low (blue) to high (red). Risks for random forest (a) and boosted regression trees (b) are on the scale of 0-1. For ENFA (c) risks are based on Mahalanobis distance from the optimal condition, and the color gradient was determined by taking the log of distances and scaling it from the largest value (blue) to the optimal condition, a distance of 0 (red). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article).

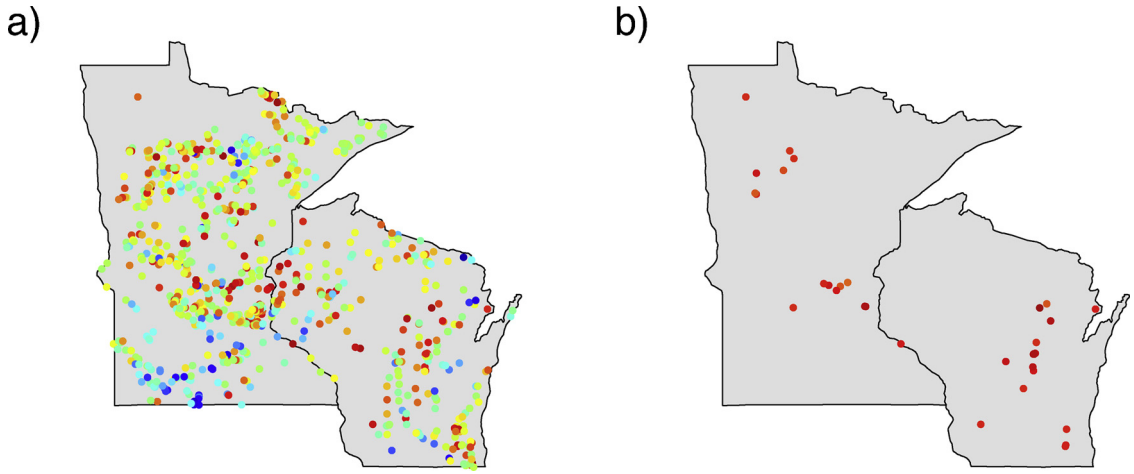

Fig. 5. Comparison of the consistency of predictions across all three models. Consistency was quantified by calculating the ranked percentile suitability for each lake based on each model and then taking the standard deviation of suitability percentiles. In panel a) colors indicate relative consistency of predictions going from blue (least consistent) to red (most consistent) across the entire range of consistency values. Additionally, the lakes that ranked in the top 25% of predicted suitability across all 3 models are shown in b) with the color indicating mean percentile (using the same blue to red color scale). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article).

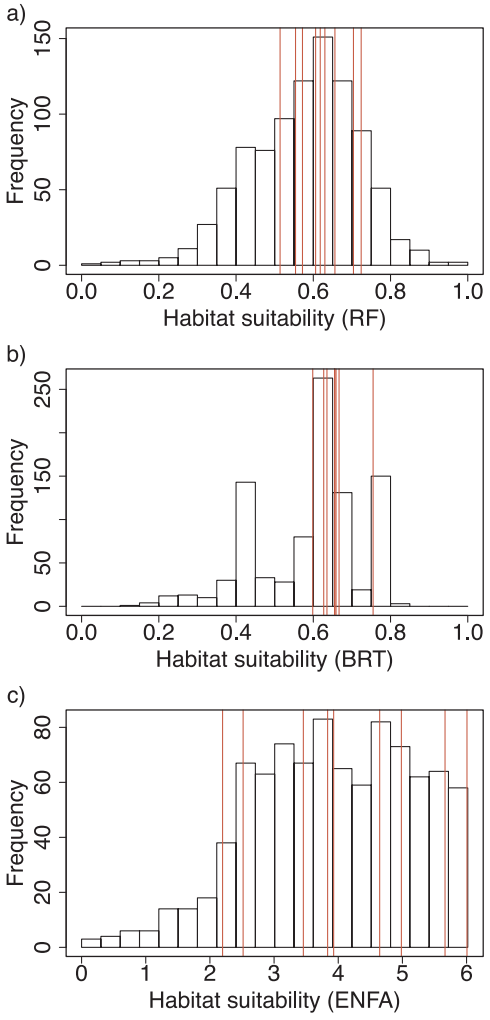

Fig. 6. Distributions of risk predictions for all lakes in the dataset for a) random forest and b) boosted regression trees and c) ENFA approaches. Vertical lines indicate the specific predicted risk values for each known starry stonewort lake in Minnesota and Wisconsin for which we also had sufficient environmental data to estimate risk.

disagreement between methods, all indicated substantial vulnerability to further spread within the upper Midwest. By developing lake-level predictions based on water quality metrics rather than climatological data, we were able to significantly refine the spatial scale offered by previous risk predictions for *N. obtusa* (Escobar et al., 2016; Romero-Alvarez et al., 2017). This finer granularity provides a practicable scale for targeting interventions aimed at reducing *N. obtusa* spread.

Our analyses did not identify a single environmental parameter as a clear driver of the niche for *N. obtusa*, but we were able to identify multiple environmental parameters likely to relate to habitat suitability. Secchi depths and chlorophyll *a* concentrations (which are often correlated; Tilzer, 1988) tended to have the strongest influence on models, with lower clarity and higher chlorophyll *a* indicating higher risk. This suggests intermediate to higher productivity lakes are more vulnerable. Alternatively, lower clarity and higher algal production are also indicators of human activity and anthropogenic influence (Rabalais, 2002; Smith et al., 1999). Thus, Secchi depths and chlorophyll *a* may instead or additionally be indirect indicators of increased opportunities for transmission. Conductivity was also a strong environmental predictor and may be a direct, positive determinant of the invasion niche for *N. obtusa* (Midwood et al., 2016; Sleith et al., 2015). In the native range of *N. obtusa*, it is typically found in mesotrophic to eutrophic, calcareous waters (Larkin et al., 2018), which is consistent with our model predictions of suitable habitat having relatively low clarity and high conductivity.

While ENMs generally focus only on environmental conditions, the influence of the extant community on the ability of *N. obtusa* to establish in a particular location should not be discounted. Many locations that are highly suitable for *N. obtusa* are likely to have high suitability for other characeans (Sleith et al., 2018) and vascular macrophytes, and establishment is likely to be mediated by interspecific interactions as well as habitat suitability—though recent work in Minnesota lakes has suggested that environmental constraints may be more important determinants of macrophyte invasions than species interactions (Muthukrishnan et al., 2018)

4.1. Ecological niche modeling of invasive species

While there is a strong rationale for interpreting our model results as indicating the preferences or requirements of *N. obtusa* (Basille et al., 2008; Rushton et al., 2004), there is also a risk of overinterpreting what are fundamentally correlative data. It can be difficult to determine if

any given environmental parameter directly influences the niche of *N. obtusa* or if it simply covaries with some other true driver. In the case of invasive species, this problem is confounded by environmental parameters potentially covarying with human activity that directly increases dispersal probability (e.g., Westphal et al., 2008). As such, the purpose for which models are being used—for example whether the goal is to simply predict occurrence (Williams et al., 2009) or to determine fundamental environmental constraints (Hirzel and Le Lay, 2008)—need to be carefully considered when choosing which variables to include and how to interpret model outputs. These issues also argue for conducting more direct experimental evaluations of how species respond to specific environmental parameters, as well as spending more effort on model validation, such as searching lakes predicted to be high vs. low risk to test model predictions.

Applying ENMs to predict habitat suitability for invasive species also has particular challenges that are not present for long-established species (Araújo and Peterson, 2012; Václavík and Meentemeyer, 2012). Most ENM approaches, and particularly those that use both presence and absence information, work from the assumption that species have existed in a region long enough for their distribution to have equilibrated with respect to the environment (Guisan and Thuiller, 2005). This assumption is often violated in the case of invasive species (Václavík and Meentemeyer, 2012, 2009). These challenges do not invalidate model predictions, but they make clear the importance of recognizing the limits of the approach (Araújo and Peterson, 2012). The fact that we trained our model based on data from New York but then used it to predict risk in the upper Midwest creates a scenario where the models are inherently confounded and broader regional differences (e.g., in climatic, geochemical, or biological factors) which may limit the applicability of the model in the new region. This can be exacerbated when the new region has conditions at the edge of the range observed in the training region, as was the case for phosphorus and chlorophyll *a* and potentially climatic conditions (Escobar et al., 2016). These uncertainties have the potential to lead to underestimates of suitability or invasion risk as models are less likely to effectively predict risk in novel habitats (Atwater et al., 2018; Fitzpatrick and Hargrove, 2009). Thus ENM-based invasion predictions need to be interpreted cautiously and should be part of an iterative process of validation and refinement. However, our set of known infested lakes in the upper Midwest does provide a limited test of model validity. The fact that those lakes tended to have higher risk predictions with models trained in a completely different region supports the accuracy and robustness of the models. These results suggest that the models are indeed identifying important environmental correlates and could be useful for targeting interventions across broad geographic ranges.

4.2. Model evaluation

Broadly speaking, there is a strong argument for making explicit predictions that can be evaluated and falsified (Houlahan et al., 2017) and it is important that those additional steps be undertaken in order to more efficiently refine models. ENM and machine-learning approaches to evaluate species' niches have the benefit that they can readily incorporate new data. Thus, sampling of locations predicted to have high suitability is a powerful approach to both test models and provide new information to update model assumptions (Hirzel et al., 2006). Inconsistencies in predicted risk between the different methods make it clear that there is uncertainty in our understanding of the key drivers of habitat suitability for *N. obtusa*, and a need for further refinement.

Without appropriate verification data, it is difficult to evaluate the relative accuracy or value of the different modeling approaches (Araújo et al., 2005). As such, without additional data we are hesitant to clearly prefer any of our modeling approaches and looking for consensus is a conservative initial approach—though ENFA did appear to predict less well, which may argue for methods that incorporate absence data. However, even if imperfect, model predictions can be useful for designing surveillance strategies targeting not only areas of high risk for invasion, but also areas that are most useful for model validation and refinement. Areas of high inconsistency between models should be prioritized for additional empirical evaluation because alignment with a particular modeling approach may indicate deficiencies in others. Areas of inconsistency, as well as areas of intermediate risk, likely reflect environmental conditions that are poorly covered in the training data and could have the strongest influence on improving models. Most importantly, model uncertainties underscore the importance of an iterative process of prediction, validation, and refinement to improve models (Luck, 2002). Additionally, this modeling effort purposefully focused on the environmental preferences of *N. obtusa*, while a true risk analysis would also incorporate information on the potential for dispersal to particular lakes. Integrating estimates of habitat suitability with information on current invasion locations and habitat connectivity will increase the quality of risk predictions (Gallardo et al., 2012).

5. Conclusions

The potential spread of invasive species is clearly a key challenge for ecosystems across the globe (Pyšek et al., 2012; Vilà et al., 2011). Management for invasive species is most effective when it can prevent dispersal or if established populations are relatively small, thus spread prevention and early detection are particularly critical (Mack et al., 2000; Simberloff, 2003). However, conservation and management resources are limited, so means to refine expectations of invasion risk are valuable for prioritization and resource allocation. There are numerous challenges for such efforts, but by using rigorously developed and data-driven models it may be possible to improve both our understanding of the ecology of specific invaders and our ability to manage invasions.

Acknowledgements

Funding for this project was provided through the Minnesota Aquatic Invasive Species Research Center from the Minnesota Environment and Natural Resources Trust Fund. We would like to thank John D. Wehr for assistance with New York water chemistry analysis. This work was also supported by the Sarah K. de Coizart Article TENTH Perpetual Charitable Trust, a Northeast Algal Society Student Grant to Support Research, an International Phycological Society Paul C. Silva Student Grant, a Phycological Society of America Grant-in-Aid of Research, a Northeast Aquatic Plant Management Society Graduate Scholarship, and a City University of New York Doctoral Student Research Grant. This material is also based upon work supported by the National Science Foundation under Grant Numbers DEB-1020660, DEB-1036466 and DBI-1348920 as well as the Wisconsin Department of Natural Resources under grant number AIRD10716.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.aquabot.2018.08.001.

References

Araújo, M.B., Peterson, A.T., 2012. Uses and misuses of bioclimatic envelope modeling. Ecology 93, 1527–1539. https://doi.org/10.1890/11-1930.1.

Araújo, M.B., Pearson, R.G., Thuiller, W., Erhard, M., 2005. Validation of species–climate impact models under climate change. Glob. Chang. Biol. 11, 1504–1513. https://doi.org/10.1111/j.1365-2486.2005.01000.x.

Atwater, D.Z., Ervine, C., Barney, J.N., 2018. Climatic niche shifts are common in introduced plants. Nat. Ecol. Evol. 2, 34–43. https://doi.org/10.1038/s41559-017-0396-z.

Basille, M., Calenge, C., Marboutin, É., Andersen, R., Gaillard, J.-M., 2008. Assessing habitat selection using multivariate statistics: some refinements of the ecological-niche factor analysis. Ecol. Modell. 211, 233–240. https://doi.org/10.1016/j.ecolmodel.2007.09.006.

Brainard, A.S., Schulz, K.L., 2016. Impacts of the cryptic macroalgal invader, *Nitellopsis obtusa*, on macrophyte communities. Freshw. Sci. 36, 55–62. https://doi.org/10.

1086/689676.
Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32. https://doi.org/10.1023/A:1010933404324.
Breiman, L., Friedman, J., Stone, C.J., Olshen, R.A., 1984. Classification and Regression Trees, the Wadsworth and Brooks-cole Statistics-probability Series. Taylor & Francis.
Brook, B.W., Sodhi, N.S., Bradshaw, C.J.A., 2008. Synergies among extinction drivers under global change. Trends Ecol. Evol. (Amst.) 23, 453–460. https://doi.org/10.1016/j.tree.2008.03.011.
Calenge, C., 2006. The package adehabitat for the R software: tool for the analysis of space and habitat use by animals. Ecol. Modell. 197, 1035.
Calenge, C., Darmon, G., Basille, M., Loison, A., Jullien, J.-M., 2008. The factorial decomposition of the mahalanobis distances in habitat selection studies. Ecology 89, 555–566. https://doi.org/10.1890/06-1750.1.
Cutler, D.R., Edwards, T.C., Beard, K.H., Cutler, A., Hess, K.T., Gibson, J., Lawler, J.J., 2007. Random forests for classification in ecology. Ecology 88, 2783–2792. https://doi.org/10.1890/07-0539.1.
De'ath, G., Fabricius, K.E., 2000. Classification and regression trees: a powerful yet simple technique for ecological data analysis. Ecology 81, 3178–3192. https://doi.org/10.1890/0012-9658(2000)081[3178:CARTAP]2.0.CO;2.
Elith, J., Graham, C.H., Anderson, R.P., Dudík, M., Ferrier, S., Guisan, A., Hijmans, R.J., Huettmann, F., Leathwick, J.R., Lehmann, A., Li, J., Lohmann, L.G., Loiselle, A.B., Manion, G., Moritz, C., Nakamura, M., Nakazawa, Y., Overton, J.M.c C.M., Townsend Peterson, A., Phillips, J.S., Richardson, K., Scachetti-Pereira, R., Schapire, R.E., Soberón, J., Williams, S., Wisz, M.S., Zimmermann, N.E., 2006. Novel methods improve prediction of species' distributions from occurrence data. Ecography (Cop.) 29, 129–151. https://doi.org/10.1111/j.2006.0906-7590.04596.x.
Elith, J., Leathwick, J.R., Hastie, T., 2008. A working guide to boosted regression trees. J. Anim. Ecol. 77, 802–813. https://doi.org/10.1111/j.1365-2656.2008.01390.x.
Escobar, L.E., Qiao, H., Phelps, N.B.D., Wagner, C.K., Larkin, D.J., 2016. Realized niche shift associated with the Eurasian charophyte *Nitellopsis obtusa* becoming invasive in North America. Sci. Rep. 6, 29037.
Fitzpatrick, M.C., Hargrove, W.W., 2009. The projection of species distribution models and the problem of non-analog climate. Biodivers. Conserv. 18, 2255. https://doi.org/10.1007/s10531-009-9584-8.
Gallardo, B., Errea, M.P., Aldridge, D.C., 2012. Application of bioclimatic models coupled with network analysis for risk assessment of the killer shrimp, *Dikerogammarus villosus*, Great Britain. Biol. Invasions 14, 1265–1278. https://doi.org/10.1007/s10530-011-0154-0.
Geis, J.W., Schumacher, G.J., Raynal, D.J., Hyduke, N.P., 1981. Distribution of *Nitellopsis obtusa*(Charophyceae, Characeae) in the St Lawrence River: a new record for North America. Phycologia 20, 211–214. https://doi.org/10.2216/i0031-8884-20-2-211.1.
Guisan, A., Thuiller, W., 2005. Predicting species distribution: offering more than simple habitat models. Ecol. Lett. 8, 993–1009.
Hirzel, A.H., Le Lay, G., 2008. Habitat suitability modelling and niche theory. J. Appl. Ecol. 45, 1372–1381. https://doi.org/10.1111/j.1365-2664.2008.01524.x.
Hirzel, A.H., Hausser, J., Chessel, D., Perrin, N., 2002. Ecological-niche factor analysis: how to compute habitat-suitability maps without absence data? Ecology 83, 2027–2036. https://doi.org/10.1890/0012-9658(2002)083[2027:ENFAHT]2.0.CO;2.
Hirzel, A.H., Le Lay, G., Helfer, V., Randin, C., Guisan, A., 2006. Evaluating the ability of habitat suitability models to predict species presences. Ecol. Modell. 199, 142–152. https://doi.org/10.1016/j.ecolmodel.2006.05.017.
Houlahan, J.E., McKinney, S.T., Anderson, T.M., McGill, B.J., 2017. The priority of prediction in ecological understanding. Oikos 126, 1–7. https://doi.org/10.1111/oik.03726.
Jiménez-Valverde, A., Peterson, A.T., Soberón, J., Overton, J.M., Aragón, P., Lobo, J.M., 2011. Use of niche models in invasive species risk assessments. Biol. Invasions 13, 2785–2797. https://doi.org/10.1007/s10530-011-9963-4.
Karol, K.G., Sleith, R.S., 2017. Discovery of the oldest record of *Nitellopsis obtusa* (Charophyceae, Charophyta) North America. J. Phycol. 53, 1106–1108. https://doi.org/10.1111/jpy.12557.
Kilroy, C., Snelder, T.H., Floerl, O., Vieglais, C.C., Dey, K.L., 2008. A rapid technique for assessing the suitability of areas for invasive species applied to New Zealand's rivers. Divers. Distrib. 14, 262–272. https://doi.org/10.1111/j.1472-4642.2007.00406.x.
Kriticos, D.J., Webber, B.L., Leriche, A., Ota, N., Macadam, I., Bathols, J., Scott, J.K., 2012. CliMond: global high-resolution historical and future scenario climate surfaces for bioclimatic modelling. Methods Ecol. Evol. 3, 53–64. https://doi.org/10.1111/j.2041-210X.2011.00134.x.
Kulhanek, S.A., Leung, B., Ricciardi, A., 2011. Using ecological niche models to predict the abundance and impact of invasive species: application to the common carp. Ecol. Appl. 21, 203–213. https://doi.org/10.1890/09-1639.1.
Kumar, S., Spaulding, S.A., Stohlgren, T.J., Hermann, K.A., Schmidt, T.S., Bahls, L.L., 2009. Potential habitat distribution for the freshwater diatom *Didymosphenia geminata* in the continental US. Front. Ecol. Environ. 7, 415–420. https://doi.org/10.1890/080054.
Larkin, D.J., Monfils, A.K., Boissezon, A., Sleith, R.S., Skawinski, P.M., Welling, C.H., Cahill, B.C., Karol, K.G., 2018. Biology, ecology, and management of starry stonewort (*Nitellopsis obtusa*; Characeae): A Red-listed Eurasian green alga invasive in North America. Aquat. Bot. 148, 15–24. https://doi.org/10.1016/j.aquabot.2018.04.003.
Liaw, A., Wiener, M., 2002. Classification and Regression by randomForest. R News 2, 18–22.
Lodge, D.M., Williams, S., MacIsaac, H.J., Hayes, K.R., Leung, B., Reichard, S., Mack, R.N., Moyle, P.B., Smith, M., Andow, D.A., Carlton, J.T., McMichael, A., 2006. Biological invasions: recommendations for US policy and management. Ecol. Appl. 16, 2035–2054. https://doi.org/10.1890/1051-0761(2006)016[2035:BIRFUP]2.0.CO;2.
Luck, G.W., 2002. The habitat requirements of the rufous treecreeper (*Climacteris rufa*). 2. Validating predictive habitat models. Biol. Conserv. 105, 395–403. https://doi.org/10.1016/S0006-3207(01)00223-3.
Mack, R.N., Simberloff, D., Lonsdale, W.M., Evans, H., Clout, M., Bazzaz, F.A., 2000. Biotic invasions: causes, epidemiology, global consequences, and control. Ecol. Appl. 10, 689–710.
Mehta, S.V., Haight, R.G., Homans, F.R., Polasky, S., Venette, R.C., 2007. Optimal detection and control strategies for invasive species management. Ecol. Econ. 61, 237–245. https://doi.org/10.1016/j.ecolecon.2006.10.024.
Midwood, J.D., Darwin, A., Ho, Z.-Y., Rokitnicki-Wojcik, D., Grabas, G., 2016. Environmental factors associated with the distribution of non-native starry stonewort (*Nitellopsis obtusa*) in a Lake Ontario coastal wetland. J. Great Lakes Res. 42, 348–355. https://doi.org/10.1016/j.jglr.2016.01.005.
Muthukrishnan, R., Hansel-Welch, N., Larkin, D.J., 2018. Environmental filtering and competitive exclusion drive biodiversity-invasibility relationships in shallow lake plant communities. J. Ecol. 0. https://doi.org/10.1111/1365-2745.12963.
Olmanson, L.G., Brezonik, P.L., Bauer, M.E., 2014. Geospatial and temporal analysis of a 20-Year record of Landsat-based water clarity in Minnesota's 10,000 Lakes. JAWRA J. Am. Water Resour. Assoc. 50, 748–761. https://doi.org/10.1111/jawr.12138.
Pearson, R.G., Dawson, T.P., 2003. Predicting the impacts of climate change on the distribution of species: are bioclimate envelope models useful? Glob. Ecol. Biogeogr. 12, 361–371. https://doi.org/10.1046/j.1466-822X.2003.00042.x.
Peterson, A.T., 2003. Predicting the geography of species' invasions via ecological niche modeling. Q. Rev. Biol. 78, 419–433. https://doi.org/10.1086/378926.
Pyšek, P., Jarošík, V., Hulme, P.E., Pergl, J., Hejda, M., Schaffner, U., Vilà, M., 2012. A global assessment of invasive plant impacts on resident species, communities and ecosystems: the interaction of impact measures, invading species' traits and environment. Glob. Chang. Biol. 18, 1725–1737. https://doi.org/10.1111/j.1365-2486.2011.02636.x.
R Core Team, 2014. R: A Language and Environment for Statistical Computing.
Rabalais, N.N., 2002. Nitrogen in aquatic ecosystems. AMBIO A J. Hum. Environ. 31, 102–112. https://doi.org/10.1579/0044-7447-31.2.102.
Ridgeway, G., 2015. gbm: Generalized Boosted Regression Models.
Romero-Alvarez, D., Escobar, L.E., Varela, S., Larkin, D.J., Phelps, N.B.D., 2017. Forecasting distributions of an aquatic invasive species (*Nitellopsis obtusa*) under future climate scenarios. PLoS One 12, e0180930.
Rushton, S.P., Ormerod, S.J., Kerby, G., 2004. New paradigms for modelling species distributions? J. Appl. Ecol. 41, 193–200. https://doi.org/10.1111/j.0021-8901.2004.00903.x.
Simberloff, D., 2003. How much information on population biology is needed to manage introduced species? Conserv. Biol. 17, 83–92. https://doi.org/10.1046/j.1523-1739.2003.02028.x.
Simons, J., Nat, E., 1996. Past and present distribution of stoneworts (Characeae) in The Netherlands. In: Caffrey, J.M., Barrett, P.R.F., Murphy, K.J., Wade, P.M. (Eds.), Management and Ecology of Freshwater Plants: Proceedings of the 9th International Symposium on Aquatic Weeds. European Weed Research Society. Springer, Netherlands, Dordrecht, pp. 127–135. https://doi.org/10.1007/978-94-011-5782-7_20.
Sleith, R.S., Havens, A.J., Stewart, R.A., Karol, K.G., 2015. Distribution of *Nitellopsis obtusa*(Characeae) in New York, U.S.A. Brittonia 67, 166–172. https://doi.org/10.1007/s12228-015-9372-6.
Sleith, R.S., Wehr, J.D., Karol, K.G., 2018. Untangling climate and water chemistry to predict changes in freshwater macrophyte distributions. Ecol. Evol. 8, 2802–2811. https://doi.org/10.1002/ece3.3847.
Smith, V.H., Tilman, G.D., Nekola, J.C., 1999. Eutrophication: impacts of excess nutrient inputs on freshwater, marine, and terrestrial ecosystems. Environ. Pollut. 100, 179–196. https://doi.org/10.1016/S0269-7491(99)00091-3.
Thuiller, W., Richardson, D.M., Pyšek, P., Midgley, G.F., Hughes, G.O., Rouget, M., 2005. Niche-based modelling as a tool for predicting the risk of alien plant invasions at a global scale. Glob. Chang. Biol. 11, 2234–2250. https://doi.org/10.1111/j.1365-2486.2005.001018.x.
Tilzer, M.M., 1988. Secchi disk — chlorophyll relationships in a lake with highly variable phytoplankton biomass. Hydrobiologia 162, 163–171. https://doi.org/10.1007/BF00014539.
Václavík, T., Meentemeyer, R.K., 2009. Invasive species distribution modeling (iSDM): Are absence data and dispersal constraints needed to predict actual distributions? Ecol. Modell. 220, 3248–3258. https://doi.org/10.1016/j.ecolmodel.2009.08.013.
Václavík, T., Meentemeyer, R.K., 2012. Equilibrium or not? Modelling potential distribution of invasive species in different stages of invasion. Divers. Distrib. 18, 73–83. https://doi.org/10.1111/j.1472-4642.2011.00854.x.
Vilà, M., Espinar, J.L., Hejda, M., Hulme, P.E., Jarošík, V., Maron, J.L., Pergl, J., Schaffner, U., Sun, Y., Pyšek, P., 2011. Ecological impacts of invasive alien plants: a meta-analysis of their effects on species, communities and ecosystems. Ecol. Lett. 14, 702–708. https://doi.org/10.1111/j.1461-0248.2011.01628.x.
Westphal, M.I., Browne, M., MacKinnon, K., Noble, I., 2008. The link between international trade and the global distribution of invasive alien species. Biol. Invasions 10, 391–398. https://doi.org/10.1007/s10530-007-9138-5.
Williams, J.N., Seo, C., Thorne, J., Nelson, J.K., Erwin, S., O'Brien, J.M., Schwartz, M.W., 2009. Using species distribution models to predict new occurrences for rare plants. Divers. Distrib. 15, 565–576. https://doi.org/10.1111/j.1472-4642.2009.00567.x.