Using Species Distribution Models to Predict Suitable Habitat for Threatened Plant Species of Southern Ontario

by

Hanna Rosner-Katz

A thesis submitted to the Faculty of Graduate and Postdoctoral Affairs in partial fulfillment of the requirements for the degree of

Master of Science

in

Biology

Carleton University

Ottawa, Ontario

Overview

One of the greatest challenges to rare species management is our lack of complete knowledge concerning their geographic distributions. Part of the Wallacean shortfall (Lomolino, 2004), this problem makes knowing a species' true conservation status difficult on global, regional, and local levels. This situation is exemplified by the many at-risk species of plants native to Southern Ontario for which there is incomplete data on the locations of their populations. Therefore, it is important to uncover effective methodologies that can be used to help fill this information gap. Towards this end, one conservation tool that has proliferated in scope and use in recent years is species distribution models (SDMs). By using species occurrence data (most commonly just species presences) along with relevant environmental predictors, the habitat suitability or probability of presence for a species can be predicted across a region (Guisan & Zimmerman 2000). The output of these models can thus be used to direct survey efforts and discover previously unknown populations of threatened species (McCune 2016, Williams et al. 2009). However, with the limited time and resources available for such conservation activities, models must a) be used in as efficient a manner as possible while also b) make use of reliable data so that outputs are accurate depictions of a species' realized niche.

The work presented here is composed of two chapters. The first chapter, "Using stacked SDMs with accuracy and threat weighting to optimize surveys for threatened plant species" examines how best to target search efforts for woodland plant species at-risk in Southern Ontario by testing the success of using multiple SDMs simultaneously (known as stacked SDMs) with and without weighting by species rarity and model

output accuracy. Specifically, this chapter asks 1) What effect does weighting by model accuracy and/or species rarity have on survey efficiency?; 2) do areas with high predicted probability of presence for multiple species have a higher probability of supporting target species in comparison with areas that have high predicted probability of presence for only one species? and 3) is total vascular plant species richness significantly higher in areas with high predicted probability of presence for multiple threatened species?

In the second chapter, “Incorporating older presence data in species distribution models does not negatively impact predictive performance when records are spatially accurate,” I use seven rare plant species of Southern Ontario to explore the effect of using different proportions of older presence records alongside newer records in distribution models on the predictive performance of the models. Previous research has shown that incorporating older, coarse-resolution data with low spatial accuracy can decrease model predictive performance (Reside et al. 2011), but it is unknown whether including spatially accurate older records would have the same effect. Here, I address this question in terms of independent AUC values and sensitivity values as measures of model accuracy.

The results of this research show that by stacking the probability of occurrence model outputs for threatened plant species, discovery of new occurrences of both species included in the stacked SDM and threatened species not included in the stacked SDM could be made. These discoveries were approximately twice as likely to occur in multi-species cells (MSC), where multiple species were predicted to have suitable habitat, than in single-species cells (SSC), where only one species was

predicted to have suitable habitat. There was no difference in species richness between the two types of sites, so the methodology used here provides a unique, more efficient way to target survey sites than building models to estimate total species richness. Weighting by species rarity and/or model accuracy did not affect survey success. Lastly, incorporating older records in the distribution models generally did not significantly lower model predictive performance in terms of AUC and sensitivity. Therefore, using these older, spatially accurate presence records can be an effective way to increase sample size, especially for rare species which are often limited in their data availability.

References

Guisan, A., & Zimmermann, N. E. (2000). Predictive habitat distribution models in ecology. *Ecological Modelling*, *135*(2-3), 147-186.

Lomolino, M.V. (2004) Conservation biogeography. *Frontiers of Biogeography: new directions in the geography of nature* (ed. by M.V. Lomolino and L.R. Heaney), Sinauer Associates, Sunderland, Massachusetts.

McCune, J. L. (2016). Species distribution models predict rare species occurrences despite significant effects of landscape context. *Journal of Applied Ecology*, 53(6), 1871-1879.

Reside, A. E., Watson, I., VanDerWal, J., & Kutt, A. S. (2011). Incorporating low-resolution historic species location data decreases performance of distribution models. Ecological Modelling, 222(18), 3444-3448.

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. *Diversity and Distributions*, 15(4), 565-576.

Acknowledgements

This thesis greatly benefitted from the help of many people in ineffable ways. Firstly, I need to thank my supervisor, Joseph Bennett, who from our very first conversation three years ago to the present has been nothing but encouraging, patient, understanding, and inspiring. I am so grateful that he has afforded me the opportunity to work on a research project concerning a conservation issue that is so important and close to us both. Throughout this process he has always provided helpful, thoughtful feedback on matters such as field work logistics, data analysis, writing, and all steps in-between while showing the utmost concern for my mental and physical well-being. His kindness and generosity of time is unmatched.

Except perhaps by Jenny McCune. Her incredible vision and motivation to begin this larger research project while a postdoc at the University of Guelph is what made my thesis research possible. From training me on the ins and outs of Maxent, to helping me with identification of pesky (yet loveable) asters and grasses, to providing me with carefully constructed comments on my thesis drafts (emphasis on the plural), she has been there every step of the way, always with a smile on her face. Jenny both encouraged and inspired me to consider different potential explanations and possibilities throughout my research and for all of this I cannot thank her enough.

I would also like to sincerely thank my other two committee members, Andrew Simons and David Currie, both of whom provided me with invaluable comments and suggestions during my committee meetings. This thesis greatly benefitted from those discussions.

I would be remiss to not thank my field assistant during the summer, Jed I. Lloren, who not only braved the ever-present poison ivy, fields of stinging nettle, allergy-inducing ragweed, swarms of mosquitos, hot weather, and rainy days, but managed to stay positive and pleasant to be around throughout. Perhaps greatest of all was his ability to withstand my quirks and eccentricities good-naturedly. I'm glad I didn't turn him off from this line of work completely.

I also would like to thank Nature Conservancy of Canada (NCC) staff for allowing me to survey on NCC land and providing me with the necessary maps and accessibility information to do so. I especially want to thank Mhairi McFarlane, the Conservation Science Manager for Ontario for the NCC who helped to coordinate said surveys. Additionally, I thank all the private landowners who allowed me to survey on their property and whose participation was integral to the project. I would also like to thank the Natural History Information Centre for providing me access to rare plant records which were essential to the completion of this research.

Finally, I would like to thank the family and friends whose support allowed me to complete the work necessary for this thesis. My mom, dad, and aunt all provided me with much-appreciated encouragement, not just during my thesis but at all times in my life leading up to it. Emily, Kaitlyn, Sunu, Therese: thank you all for giving me the space to talk it out, for giving me confidence, and most of all for giving me laughs, fun times, and cherished memories throughout our friendship. And of course, where would I be without Pouncer. For the past two years you have been my Edelweiss and have given me a much-needed daily dose of jellicle felinity.

Table of Contents

List of Figures and Tables

Chapter 1

Chapter 2

CHAPTER 1: Using stacked SDMs with accuracy and threat weighting to optimize surveys for threatened plant species

ABSTRACT

Surveys are required to locate previously unknown occurrences of threatened species. Here, I test a methodology to determine how surveys for threatened plant species can be conducted most efficiently to discover the greatest number of new occurrences possible. To optimize efficiency of species-at-risk surveys, I used species distribution models for 22 rare plant species throughout southern Ontario to predict the best possible survey locations among individual 1-ha pixels. For each pixel, I weighted model outputs by accuracy and species rarity to create an efficiency value. Based on efficiency values, I conducted field surveys in multi-species cells, “MSC” (areas with high predicted efficiency for multiple species) and in single species cells, “SSC” (areas with high probability for only one species) to determine the utility of multi-species survey optimization. MSC were over two times more likely to have at least one threatened plant species discovered than SSC. Efficiency ranks were also useful in directing surveyors toward incidental discoveries of other rare species that were not modeled. This technique can help direct surveys to more efficiently find threatened plant species occurrences.

Key-words: rare plants, species distribution model, prioritization, S-SDM, Maxent, forest, species at-risk, conservation

Introduction

To effectively monitor and protect threatened plant species, we first must know the geographic locations of their populations. As important as this knowledge is, many plant species at risk of extinction are lacking precise distribution information. While field surveys can help to fill these gaps (Peterson et al. 2011), they must be performed in an efficient manner so as not to waste either time, which is critical for accurate surveys (Zhang et al. 2014), or resources that could be put toward other conservation activities (Lindenmayer et al. 2013). Reliable protocols are thus needed to help efficiently direct survey efforts.

Species distribution models have recently proliferated as conservation tools (Guisan et al. 2013). By using species occurrence data (presence-absence or presence only) along with relevant environmental or spatial predictors, the habitat suitability or probability of presence for a species can be predicted across an area of interest (Guisan & Zimmerman 2000). Depending on the type of data available and the end goal, many techniques are available (Elith et al. 2006). In addition to their most basic purpose of modeling a species' distribution, SDMs have also been used to predict current species richness (e.g. Guisan & Theurillat 2000, Algar et al. 2009, Newbold et al. 2009, Raes et al. 2009, Parvianen et al. 2009, Pineda & Lobo 2009, Benito 2013, Amaral et al. 2017, Scherrer et al. 2017), predict future species richness with changing climate (Thuiller et al. 2005), predict shifts in species' distributions (e.g. Thuiller 2004, Elith et al. 2010, McKenney et al. 2014), model endemicity patterns (Raes et al. 2009), designate floristic regions (Zhang et al. 2012), assess the efficacy of current protected

areas and propose locations for new ones (Loiselle et al. 2003, Koch et al. 2017, Amaral et al. 2017), select species for ecological restoration (Gastón & García-Viñas 2013), aid in bioassessment (Labay et al. 2015), prescribe management for invasive species (Bennett 2014), and find new occurrences of threatened species (Williams et al. 2009, Rebelo & Jones 2010, Peterson et al. 2011, Peterman et al. 2013, Searcy & Shaffer 2014, Fois et al. 2015, McCune 2016). Increasingly, outputs of SDMs for individual species are stacked together to create one composite map (Ferrier & Guisan 2006); this technique is often referred to as S-SDM (where the "S" stands for stacking) (Dubois et al. 2011).

Stacked distribution models offer an effective and relatively simple method for highlighting areas of special conservation importance based on predictions of species richness of a taxonomic group of interest (Newbold et al. 2009, Yu et al. 2017), endemic species richness (Raes et al. 2009), or threatened species richness (Parvianen et al. 2009; Koch et al. 2017). This last factor has significant implications for threatened species management and recovery. Knowledge of which areas have the highest potential concentration of threatened species can point to distinctly promising survey sites, thus leading to the discovery of previously unknown threatened species locations (Williams et al. 2009) which can then allow for management of these new populations and a better assessment of the species' conservation status.

Simple stacking of individual SDMs to derive threatened species richness maps, however, may not be the most efficient way to choose survey locations with the purpose of discovering new occurrences. Two additional factors that have received little attention in the S-SDM literature but nonetheless should be considered are 1) the accuracy of the

SDM for each species and 2) the conservation status of each of the species being modelled.

The accuracy and predictive performance of distribution models for individual species has been discussed extensively (e.g. Segurado & Araaujo 2004, Elith & Leathwick 2009), and with good reason when considering the conservation implications. A model that tends to overpredict suitable habitat and make commission errors will waste survey resources by sending surveyors to supposedly suitable areas where the species is actually absent, while a model that tends to underpredict suitable habitat and make omission errors will cause surveyors to overlook supposedly unsuitable areas where the species is present (Loiselle 2003). Few studies have accounted for individual model accuracy in S-SDMs. An exception to this is Dunn et al. (2016) who weighted SDM outputs by both AUC value and expert opinion to analyze the effectiveness of protected areas for Himalayan Galliformes. When working within an S-SDM framework, there will be variation in the accuracy of the individual models being stacked due to the differing characteristics of the species being modelled (Guisan et al. 2007, Zimmerman et al. 2007, Syphard & Franklin 2010), the quality of the data for each species (e.g. Graham et al. 2004, Moudrý & Šímová 2012, Soultan & Safi 2017), and the relative prevalence of the species (Hernandez et al. 2006, Le Lay et al. 2010, Soultan & Safi 2017). Thus, it is important to take this variation into account in the stacking process if the goal is to focus on survey locations with the highest probabilities of containing at least one of the species of interest.

Secondly, many agencies prioritize management of species based on threat level (e.g. ESA 1973; SARA 2002), so knowing where the most threatened species are may

be more important than knowing where other, less threatened species are. Some modelers have noted this fact and have weighted individual models in stacked SDMs by species' relative rarity (Albuquerque & Beier 2016, Tukainen et al. 2017, Yu et al. 2017), or have used the habitat specificity of the species (Miličić et al. 2017) to highlight potential areas important for threatened species conservation. However, to my knowledge, weighting individual model outputs by rarity and accuracy within an S-SDM framework has not yet been shown to lead to an increased detection rate of priority threatened species.

Here, I test a methodology to determine how surveys for threatened plant species can be conducted most efficiently to discover the greatest number of new occurrences possible. I built SDMs for 22 threatened vascular plant species in southern Ontario and stacked them - weighting each individual model by model accuracy, species conservation status, both, or neither. I conducted field surveys in areas predicted to be suitable for one or more species to test the effectiveness of these S-SDM maps in leading to discovery of new threatened plant occurrences, which I define here as those species which are considered vulnerable, imperiled, or critically imperiled at the provincial level (i.e. S-rank S3, S2, or S1; Faber-Langendoen et al. 2012). My main research questions were i) What effect does weighting by model accuracy and/or species rarity have on survey efficiency?; ii) do areas with high predicted probability of presence for multiple rare species have a higher probability of supporting target species in comparison with areas that have high predicted probability of presence for only one species? and iii) is total vascular plant species richness significantly higher in areas with high predicted probability of presence for multiple threatened species (in which case

richness models may be useful as an easier alternative to S-SDM that rely on models for individual rare species)?

Methods

Study Region and Species

I focused this study on the forests of Southern Ontario (Figure 1), predominantly in the Carolinian zone (Lake Erie-Lake Ontario Ecoregion; Crins et al. 2009), which is characterized by deciduous canopy cover. This region is ecologically important for vascular plants at both the provincial and national level, as approximately 72% of Ontario's and more than 40% of Canada's plant species occur in this relatively small area (Oldham 2017), despite the dominance of urban development and agricultural land use in the region (Crins et al. 2009). Approximately 1,520 vascular plants are known to occur in the region, with over half considered rare (Oldham 2017). I also located some study sites in the Lake Simcoe-Rideau Ecoregion forests, characterized by mixed deciduous-evergreen cover (Crins et al. 2009). The total extent of the study region encompasses approximately seven million hectares.

I initially selected 27 vascular plant species native to Ontario and growing in woodland habitats, based on their conservation status, their particular interest to the Ontario Ministry of Natural Resources and Forestry (MNRF), and their relative ease of identification in the field. Here, threatened species refer to those ranked S1, S2, or S3 at the provincial level, corresponding to critically imperiled, imperiled, and vulnerable species, respectively (Table 1). Most of the species included here are at the northern edge of their range within southern Ontario, with large portions of their ranges extending

south into the United States. This list contains a variety of life forms including trees, shrubs, herbs and fern species. Nomenclature follows the Ontario Natural Heritage Information Centre (NHIC, https://www.ontario.ca/page/get-natural-heritage-information).

Building and Evaluating Individual Models

I obtained presence-only occurrence records for each species from the NHIC. These records include herbarium records, field surveys by MNRF biologists, and other confirmed sightings. I used only records that had <100 m accuracy to build the models to correspond with the resolution for the environmental variables (see below). The number of occurrence records I used to build the models ranged from four to 1594 with a mean of 83.7 records per species and a median of 17.5 (Table 1). Species with occurrence records on the lower end of this range are potentially subject to overfitting (Merow et al. 2013); however, SDMs can perform well even with sample sizes as low as five (Hernandez et al. 2006; Pearson et al. 2007, van Proosdij et al. 2015) and I am specifically studying threatened species with limited available records. Records date from 1897 to 2012. I considered high spatial accuracy to be the most important criterion for including occurrence records. However, I later explored the effect of record age on the performance of SDMs which suggests that this has minimal effect on model accuracy given that records are spatially accurate (see chapter 2).

I collected data on climatic, topographic, soil, and surficial geology environmental variables to use as predictors in the models (Table 2). Previous research (McCune 2016) suggests that forest type and the amount of forest on the landscape surrounding

a site can influence whether a threatened plant species is present. Therefore, I also collected data on forest contiguity (number of 1-hectare cells out of the 80 cells immediately surrounding the focal cell that are forested) and land cover type (deciduous forest, mixed forest, swamp, etc.) across the study region. Multicollinearity was tested between environmental predictors and only those that were correlated at $r < 0.7$ were used, following Gogol-Prokurat (2011). Where two variables were correlated at $r > 0.7$, the more generalized variable was used (e.g., mean annual temperature and mean temperature of the warmest quarter were correlated so the former was kept). Multicollinearity is less of an issue for MaxEnt models than for regression (Elith et al. 2011), nonetheless, Merow et al. (2013) recommend removing highly correlated predictors. I resampled each variable to a 100m x 100m resolution.

Because the available data were limited to presence-only records and records were often limited in number, I chose to use Maxent (Phillips et al. 2006) to build SDMs. In comparison with other modelling techniques, Maxent has been shown to perform very well using a range of modelling performance measures (Elith & Graham 2009), even when few species occurrence records are available (Hernandez et al. 2006, Pearson et al. 2007, van Proosdij et al. 2015). I built eight different models for each species. All models for each species contained the climatic, topographic, soil, and surficial geology predictor variables (14 predictors). Additionally, I built models that included either the forest contiguity or land cover variables or both. I ran each of these model types twice: once with a regularization multiplier set to 1 and once with regularization set to 0.5. Varying the regularization multiplier is recommended in order to maximize predictive performance of the models (Merow et al. 2013). I set aside a random subsample of 25%

of available records to test each model and fit the model 10 times with different model fitting and test subsamples, making the final result an average of the outputs from the subsamples. As a measure of habitat suitability, I used Maxent's cumulative output in which each grid cell receives a score from 0 to 100 and can be interpreted as the percent of cells in the study area that have a cumulative value equal to or less than that cell's value (Merow et al. 2013).

I evaluated the models with independent presence and absence data originating from two sources. First, I obtained independent presences from the NHIC's central holdings database, which consists of records that have not yet been approved for addition to the main database of occurrences. Second, I obtained independent absences (and occasionally presences) from 2014 and 2015 field data in which botanists surveyed 156 100m x 100m cells throughout Southern Ontario predicted to be suitable for one or more threatened plant species (see McCune 2016; McCune et al. 2017). I excluded any independent presence located within the same grid cell as a record used to build the SDM, using the dismo package in R (R Core Team 2016, Hijmans et al. 2017). I also excluded absences if the grid cell was surveyed outside the time of year during which the plant is present and identifiable.

It is difficult to choose one "best" model among several choices because of the different performance measures available. One of the most commonly used methods is AUC, a threshold independent measure that can have values from 0 to 1 where 1 indicates a model that perfectly discriminates between presences and absences and 0.5 indicates a model that does no better than random at discriminating between the two (Fielding & Bell 1997). Sensitivity measures the percent of actual presences correctly

classified as presences by the model. Sensitivity is dependent on the chosen threshold for classifying a cell as 'suitable'. Minimal predicted area selects the model that predicts the lowest total number of cells to be suitable, while correctly predicting a certain percentage of the presence records used to build the SDM (Engler et al. 2004). Often, the model with the highest AUC may not be the one with the highest sensitivity or lowest minimal predicted area (MPA) (Gogol-Prokurat 2011). I chose one model for each species based on three criteria: 1) first, I chose the model with the highest sensitivity based on calculations with the independent data as the best model for that species (i.e. the model that predicted the highest number of independent presence records as suitable) using a threshold that allowed for 90% of presence records used to build the SDM to be correctly predicted as present; 2) then if more than one model was tied for the highest sensitivity, I chose the model with the highest AUC; 3) if there was still a tie between models, I chose the model with the lowest area predicted suitable. I considered sensitivity to be the most important measure of predictive performance for this study because I wanted to minimize the number of false negatives produced by each model. If an area is falsely characterized as unsuitable, then it will not be surveyed and rare occurrences will be missed as a result of omission error (Liu et al. 2016).

It is important to recognize that the MaxEnt predicted habitat suitability does not always have a linear relationship with species probability of occurrence (Gogul-Prokurat 2011). Therefore, I converted the output of the best SDM for each species to probability of presence based on the independent presence and absence data described above, using GLMs with a binomial link function.

S-SDM Efficiency Maps

I used only the species that had acceptable AUC values and GLM probability of presence models that were significantly better than a null model in the stacking process, to avoid using models that were not useful for predicting species distributions. Note that there may be several traits associated with poor-performing models for plant species including dispersal method, longevity, growth rate, and disturbance association (Guisan et al. 2007, Hanspach et al. 2010, Syphard & Franklin 2010), but determining the specific reasons for the species included here is beyond the scope of this work. I defined acceptable AUC values as those ≥0.6 since an AUC value of 0.5 is no better than random. Although a common practice is to consider AUC values ≥0.7 as adequate (Swets 1988), I adopted 0.6 as the cutoff because the species modeled here are rare, and any information that is better than random is potentially important. After eliminating five species using the criteria above, I was left with 22 species with which to create the efficiency maps which highlight the areas with the greatest potential for rare plant discovery.

To create the efficiency maps, I stacked the probability of occurrence model outputs for these 22 species in four ways: 1) individual probability of occurrence maps added together with no weighting; 2) probabilities of occurrence added together and weighted by accuracy of the SDM (as measured by AUC) and by the S-rank (threat level) of the species, using the following equation:

(1) $E_i = \sum_{j=1}^{n} (Pij \cdot Rj \cdot Aj)$

Where P_{ij} is the probability of occurrence of any species *j* occurring in a cell *i*, R_j is the rarity weight of species *j*, A_j is the accuracy of the best model associated with species *j*

as measured by AUC, and E_j is the resulting Efficiency Index of cell *i*. 3) models weighted by accuracy but not by rarity; and 4) models weighted by rarity but not by accuracy. Because of these different weighting systems, the range of values of the final maps are different from one another (Figure 2).

Field Surveys

I selected candidate cells for surveys that were either highly suitable for multiple species or suitable for just one species according to the efficiency map using AUC and threat level weighting. I term the former multi-species cells (MSC) and the latter single-species cells (SSC). There has been some debate about the appropriateness of applying a suitability threshold and obtaining binary predicted presence/absence maps for individual species (bS-SDM) compared to using raw probability of presence values (pS-SDM) (Guisan & Rahbek 2011, Calabrese et al. 2014, D'Amen et al. 2015); I used both methods in identifying MSC and SSC. Specifically, I defined MSC as being suitable for two or more species using a threshold that allows for a 10% omission rate for species with greater than 15 records used to build their model and a 0% omission rate for species with fewer than 15 records (bS-SDM) while also having an efficiency value within the top 5% of all grid cell values within the study region (pS-SDM) (Parviainen et al. 2009). I defined SSC as being suitable for only one species (again using the 10% omission rate threshold) independent of the cell's efficiency value. I chose not to randomly select cells with low and high efficiency indices because this could lead to surveying areas not predicted suitable for any species. Given limited time and the urgency of finding new threatened plant populations, such sampling would have been wasteful. Additionally, choosing sites in such a manner would not allow me to assess

the efficacy of surveying MSC vs. SSC. I attempted to survey the same number of MSC and SSC within each tertiary level watershed (subdivisions of secondary watersheds which are mostly made up of large river systems) to reduce spatial bias. Visual examination of the efficiency map revealed that high efficiency areas are concentrated in the southwest portion of the study area, as are known threatened species occurrences. Therefore, pairing MSC and SSC locations by watershed ensures that MSC and SSC are not strongly geographically separated.

I surveyed 70 cells (including 33 MSC and 37 SSC) on privately owned sites as well as protected areas (e.g. Nature Conservancy of Canada, Nature Trust, or Provincial Park land), between May and August 2017. I obtained written or verbal permission from the landowners for all privately owned sites, and research permits for protected areas. At each site, I navigated to the center of the 100m by 100m grid cell using a handheld GPS unit and used flagging tape and a compass to delineate four quadrants. The field team included one to four people with at least one trained botanist present walking the entire square grid cell in straight lines, noting each of the vascular plant species observed until the entire area was surveyed, usually lasting approximately 2.5-5 hours. I was present for all surveys except for two led by Dr. J.L. McCune. If we observed any threatened plants, we marked the point location of the species with a GPS unit. We took a photo or a small sample of any plant species we could not identify in the field, for later identification. We identified a small number of taxa only to the genus level (e.g. *Rubus sp., Viola sp*.).

Comparison of S-SDM Methods

To test for differences between MSC and SSC in the total, native, and exotic species richness recorded, I performed t-tests assuming unequal variances. I then built logistic regression models to test the significance of the relationship between the efficiency values (four versions for each surveyed cell) and the probability of finding any threatened plant species at that location. I performed all data analysis in R 3.3.1 (R Foundation for Statistical Computing, Vienna, Austria 2016).

Results

For eight species, the best MaxEnt model included both forest type and amount, for six species it included only forest amount, for five species it only forest type, and for eight species only the original climatic, edaphic, and topographic variables were included in the best model (Table 3). The independent AUC value for the best MaxEnt models for each species ranged from 0.442 to 0.999 with a mean of 0.83. Three species*, Juglans cinerea, Erigenia bulbosa*, and *Arisaema dracontium*, had independent AUC values below the pre-determined acceptable level of 0.6 and I excluded them from further analysis. Two additional species, *Morus rubra* and *Phegopteris hexagonoptera*, had independent AUC values above 0.6, but their models (GLMs) relating probability of presence to Maxent output did not perform better than a null model using the criterion of $\Delta AIC < 2$ (Burnham & Anderson 2002) and therefore I also excluded them from further analysis and from the efficiency maps. This left 22 species that had acceptable AUC values as well as GLMs that predicted their probability of presence throughout the study region based on the SDM habitat suitability output.

Of the seventy 1 hectare cells surveyed, 22 had at least one threatened plant species (ranked S1, S2, or S3). Sixteen plots had one species, four plots had two, and two plots had three. I found a total of 30 occurrences of 16 threatened plant species. Interestingly, only four out of the sixteen species were those I modelled and incorporated into the efficiency maps (*Castanea dentata, Celtis tenuifolia, Cornus florida, and Lithospermum latifolium*), the rest being incidental discoveries of species not modelled by the SDMs (Table 4).

The probability of finding at least one threatened plant species was approximately double in MSC compared to SSC. Including incidental threatened plant discoveries, MSC had a 42.4% success rate while the SSC had an 21.6% success rate. Not including the incidental species, the MSC had an 18.2% success rate while the SSC had an 8.1% success rate (Figure 3). Excluding those SSC locations in watersheds where there were no paired MSC, results were very similar (42.4% MSC success rate vs 18.9% SSC success rate including incidentals, 18.2% MSC success rate vs 8.1% SSC success rate not including incidentals). The mean efficiency index score of MSC locations was 4.95 while the mean efficiency index score of SSC locations was 2.68 which represented a significant difference between the two (t-test assuming unequal variance, $P<0.001$). None of the measures of species richness (total, native, or exotic) were significantly different between the two site types. Mean total species richness at SSC sites plus/minus standard error was 96.4 ± 3.67 species compared to 91.9 ± 4.34 species at MSC sites while mean native plant species richness was 78.8 ± 3.37 at SSC and 85.8 ± 3.33 at MSC. Exotic species richness was 12.8 ± 1.44 at SSC compared to 10.4 ± 0.87 at MSC (Figure 4).

The GLMs relating the probability of presence of at least one threatened plant species to the efficiency index showed relatively little differences among the weighting procedures used to create the S-SDMs. All ΔAIC values were <5 and percent deviance explained by the full models were similar (Table 5). There was a significant positive relationship in the logistic regression between the efficiency index values of all the S-SDMs, including the fully weighted S-SDM, and the probability of at least one threatened species (target and incidental species included) being present. Results were similar when only target species were included. There was a slight negative relationship in the linear regression between the efficiency index values and species richness and a slight positive relationship in the logistic regression for species richness and probability of presence of a rare species, but these were not significant (Figure 5).

Discussion

The efficiency maps I created by stacking the probability of occurrence model outputs for 22 threatened plant species allowed for the discovery of new occurrences of fifteen at-risk species. These discoveries were approximately twice as likely to occur in multi-species (MSC) sites, where multiple species were predicted to have suitable habitat, than in single-species (SSC) sites, where only one species was predicted to have suitable habitat. Additionally, species richness was not significantly different between the two site types according to any measure (total, native, or exotic). While I found new occurrences of only four out of the 22 species modelled, because I was working with rare species this discovery rate is not unusual (MacDougall & Loo 2002, Williams et al. 2009, McCune 2016).

I found eleven other plant species considered threatened in Ontario which were not modelled and not incorporated into the efficiency maps. This preponderance of incidental finds is interesting, because it means that the efficiency index not only helps to find species explicitly modelled, but also those which are not modelled but also threatened. Thus, my efficiency index is capturing something shared in the ecological niches of the species modelled as well as in some threatened species that are not modelled. The efficiency index also performed better at predicting rare species presence than a simple measure of species richness (because species richness and the probability of presence of a threatened species were not related, see Figure 5). This is important because instead of trying to model as many species as possible to get a full picture of the potential threatened species richness across an area of interest and make conservation inferences based on the resulting S-SDM which would take considerable time and financial resources (Tulloch et al. 2016), my results show that, at least for threatened plants in the region, creating SDMs for a subset of rare species may be sufficient if the main goal is to locate new rare species sharing a given habitat type. It may be possible to further improve results of future studies by purposefully choosing a relatively small subset of species to model, instead of the selection of species used in this analysis.

It must be noted that this efficiency index will not be useful for finding new rare plant species occurrences in every situation. Firstly, the efficiency index did not result in the discovery of any species ranked S1, those which are the most imperiled in the province. This is likely because these species are the least prevalent and thus any new population locations (if they exist) will be the most difficult to find. Le Lay et al. (2010)

had similar results when attempting to discover new occurrences of two extremely rare plant species based on ensembles of SDMs and were not successful in their searches.

Secondly, the efficiency index will not aid in the discovery of rare species that have very distinct habitat requirements and thus are not likely to occur in the MSC of my efficiency maps. The MSC represent areas that have especially suitable habitat based on the ecological requirements of a subset of the species modelled. Species that do not share these ecological requirements will generally not overlap in distribution. Consequently, new locations of this type of species will be more likely to occur in the SSC areas of the efficiency map. For example*, Asplenium scolependrium,* a fern species that can only grow on limestone substrate (Oldham & Brinker 2009), was not found at any MSC locations I searched in 2017. With this species' distinct habitat requirements, its areas of suitable habitat do not occur in MSC sites. There is an assortment of results in the literature concerning whether or not species with distinct habitat requirements are more easily modelled with SDMs than more generalist ones (Elith & Burgman 2002, Hernandez et al. 2006, Le Lay et al. 2010, Grenouillet et al. 2011, McCune 2016, Soultan & Safi 2017, Rhoden et al. 2017). If species with distinct habitats are in fact easily modelled, then their lack of discovery in my efficiency maps is not a major point of concern because any specialized species of particular interest can be separately modelled. The methodology presented here focuses on most efficiently finding the most occurrences of threatened plant species and not necessarily on finding individual species. If the latter is the desired conservation outcome, an individual SDM should be used.

My results did not show a difference between the different weighting systems (species S-rank, model accuracy, both, neither) in predicting the presence of at least one threatened plant species using my ΔAIC criterion. However, there was a strong relationship between the efficiency index and both the rare species probability of presence and total species richness. Thus, although the stacking of individual model outputs was useful for discovering new threatened plant occurrences, the weighting of model outputs by threat level and model accuracy was unnecessary in this study for the goal of field site survey prioritization. Dunn et al. (2016) had similar results when weighting by Maxent model accuracy to find important conservation areas for Galliformes using Zonation analysis, finding that weighting the models did not change the areas highlighted for conservation very much when compared to unweighted model stacking. This does not mean that model weighting by either threat level or accuracy should be completely discounted for future S-SDM analysis. It is possible that a similar weighting system used in a different study area with different species or more surveyed cells would have different results.

Conclusions

Optimizing survey efficiency is important for locating previously unknown occurrences of threatened species. My results show that stacking individual species distribution model outputs can not only lead to the discovery of new populations of threatened plant species that are modelled, but also to the discovery of populations of unmodeled threatened species sharing similar habitats. Using an efficiency index such as that used in this study, whether weighted or unweighted, can help practitioners focus field surveys on areas most likely to contain threatened species occurrences.

Tables

Table 1.1 Number of records used to build each model, independent AUC of the best MaxEnt model and S-Rank of the 27 species tested, including the 22 species included in the S-SDM. Total weight is the resulting weight by which the probability of occurrence of each species in each cell is multiplied before stacking across species to get the efficiency map in which both accuracy are rarity are weighted. *Species not included in the S-SDM due to low independent AUC and/or GLM that does not predict probability of presence better than a null model.

Scientific Name	Number of Records	AUC	S-rank	Total Weight
Aplectrum hyemale	5	0.734	S2	1.468
*Arisaema dracontium**	73	0.595	S3	-
Asclepias quadrifolia	7	0.999	S1	2.997
Asimina triloba	43	0.779	S3	0.779
Asplenium scolependrium	144	0.981	S3	0.981
Castanea dentata	153	0.813	S2	1.626
Celtis tenuifolia	73	0.968	S2	1.936
Chimaphila maculata	15	0.983	S1	2.949
Corallorhiza odontorhiza	4	0.913	S2	1.826
Cornus florida	295	0.775	S2	1.55
Cypripedium arietinum	65	0.860	S3	0.860
Enemion biternatum	14	0.862	S2	1.724
*Erigenia bulbosa**	8	0.453	S2S3	-
Frasera caroliniensis	30	0.882	S2	1.764
Fraxinus quadrangulata	46	0.865	S3	0.865
Heuchera americana	19	0.907	S2	1.814
Hydrastis canadensis	44	0.826	S2	1.652
*Juglans cinerea**	1594	0.467	S2	-
Liparis liliifolia	16	0.971	S2	1.942
Lithospermum latifolium	11	0.664	S3	0.664
Magnolia acuminata	58	0.847	S2	1.694
Mertensia virginica	6	0.828	S2	1.656
*Morus rubra**	42	0.780	S2	-
*Phegopteris hexagonoptera**	38	0.632	S3	-
Stylophorum diphyllum	6	0.959	S1	2.877
Trillium flexipes	4	0.967	S1	2.901
Uvularia perfoliata	9	0.938	S1	2.814

.

Table 2.1: Environmental variables used in the models for all species.

Variable	Code	Unit	Source	Reference/ Access
Elevation	cdem	meters	Canadian Digital Elevation Model	http://geogratis.gc.ca/
Slope	slope	degrees	Canadian Digital Elevation Model	http://geogratis.gc.ca/
Aspect	north	unitless	Canadian Digital Elevation Model	http://geogratis.gc.ca/
Soil Texture	soil1	Categorical, 24 categories	Soil Survey Complex, Ontario Ministry of Agriculture	https://www.ontario.ca/page/land-information-ontario
Soil Drainage	soil2	Categorical, 9 categories	Soil Survey Complex, Ontario Ministry of Agriculture	https://www.ontario.ca/page/land-information-ontario
Surficial Geology	sgu	Categorical, 40 categories	Canada Forest Service	https://www.ontario.ca/page/land-information-ontario
Isothermality	bio03	Percentage	Canada Forest Service	McKenney et al. 2011*
Mean temperature of wettest quarter	bio08	°C	Canada Forest Service	McKenney et al. 2011
Annual precipitation	bio12	mm	Canada Forest Service	McKenney et al. 2011
Precipitation seasonality	bio15	percentage	Canada Forest Service	McKenney et al. 2011
Precipitation of warmest quarter	bio18	mm	Canada Forest Service	McKenney et al. 2011
Total precipitation for	sg06	mm	Canada Forest Service	McKenney et al. 2011

growing season				
Annual mean temperature	sg12	°C	Canada Forest Service	McKenney et al. 2011
Mean temperature of growing season	sg15	°C	Canada Forest Service	McKenney et al. 2011
Land cover	SOLRIS	Categorical, 25 land use categories	Southern Ontario Land Resource Information System (MNRF)	https://www.ontario.ca/data/southern-ontario-land-resource-information-system-solris-20
Forest extent among 81 contiguous grid cells	cont	Number of 1 hectare grid cells	Southern Ontario Land Resource Information System (MNRF)	https://www.ontario.ca/data/southern-ontario-land-resource-information-system-solris-20**

* Also available at https://cfs.nrcan.gc.ca/projects/3/2

** Created from this data using ArcMap 10.4.1

Table 3.1: Results from evaluating the models using independent data collected from field surveys and the Ontario NHIC. Bolded text indicates the best model (see main text for criteria). Red text indicates models that have independent AUC values below the predefined acceptable level (<0.6).

species	environmental variables included and regularization multiplier	number of environmental variables	# records used to build model	# independent absences	# independent presences	independent AUC	TPR[1]	percent area predicted suitable with 100% presences correctly predicted
Aplectrum hyemale	original[2](1)	14	5	48	10	0.6313	0.4000	62.54
Aplectrum hyemale	original(0.5)	14	5	48	10	0.6313	0.4000	61.67
Aplectrum hyemale	original+SOLRIS(1)	15	5	38	10	0.5868	0.4000	59.96
Aplectrum hyemale	original+SOLRIS(0.5)	15	5	38	10	0.6368	0.4000	53.70
Aplectrum hyemale	original+cont(1)	15	5	38	10	0.6563	0.4000	74.63
Aplectrum hyemale	original+cont(0.5)	15	5	38	10	0.7208	0.4000	54.70
Aplectrum hyemale	original+SOLRIS+cont (1)	16	5	38	10	0.6053	0.4000	56.35

species	environmental variables included and regularization multiplier	number of environmental variables	# records used to build model	# independent absences	# independent presences	independent AUC	TPR[1]	percent area predicted suitable with 100% presences correctly predicted
Aplectrum hyemale	**original+SOLRIS+cont (0.5)**	**16**	**5**	**38**	**10**	**0.7342**	**0.4000**	**55.19**
Arisaema dracontium	original(1)	14	73	141	99	0.6720	0.6465	23.44
Arisaema dracontium	original(0.5)	14	73	141	99	0.6860	0.6162	21.80
Arisaema dracontium	**original+SOLRIS(1)**	**15**	**73**	**141**	**99**	**0.5947**	**0.7778**	**26.74**
Arisaema dracontium	original+SOLRIS(0.5)	15	73	141	99	0.6084	0.6465	18.36
Arisaema dracontium	original+cont(1)	15	73	151	99	0.7057	0.697	11.48
Arisaema dracontium	original+cont(0.5)	15	73	151	99	0.7243	0.7071	11.93
Arisaema dracontium	original+SOLRIS+cont (1)	16	73	141	99	0.6783	0.7576	13.62
Arisaema dracontium	original+SOLRIS+cont (0.5)	16	73	141	99	0.7021	0.6768	10.82
Asclepias quadrifolia	original-soil1[3] (1)	13	11	133	6	0.9912	1	0.21
Asclepias quadrifolia	original-soil1(0.5)	13	11	133	6	0.9787	1	0.23
Asclepias quadrifolia	original-soil1+SOLRIS	14	11	133	6	0.9946	1	0.04

species	environmental variables included and regularization multiplier	number of environmental variables	# records used to build model	# independent absences	# independent presences	independent AUC	TPR[1]	percent area predicted suitable with 100% presences correctly predicted
Asclepias quadrifolia	original-soil1+SOLRIS	14	11	133	6	0.9973	1	0.04
Asclepias quadrifolia	original-soil1+cont(1)	14	11	133	6	0.9937	1	0.05
Asclepias quadrifolia	original-soil1+cont(0.5)	14	11	133	6	0.995	1	0.04
Asclepias quadrifolia	original-soil1+SOLRIS+cont (1)	15	11	123	6	0.993225	1	0.09
Asclepias quadrifolia	**original-soil1+SOLRIS+cont (0.5)**	**15**	**11**	**123**	**6**	**0.9987**	**1**	**0.02**
Asimina triloba	original(1)	14	43	156	16	0.7548	0.6450	22.94
Asimina triloba	original(0.5)	14	43	156	16	0.7508	0.3750	24.22
Asimina triloba	original+SOLRIS(1)	15	43	146	16	0.7573	0.5000	22.08
Asimina triloba	original+SOLRIS(0.5)	15	43	146	16	0.7427	0.5000	25.18
Asimina triloba	**original+cont(1)**	**15**	**43**	**156**	**16**	**0.7788**	**0.6875**	**22.12**
Asimina triloba	original+cont(0.5)	15	43	156	16	0.7941	0.6250	18.66
Asimina triloba	original+SOLRIS+cont (1)	16	43	146	16	0.8052	0.5000	20.92
Asimina triloba	original+SOLRIS+cont (0.5)	16	43	146	16	0.7817	0.4375	19.04
Asplenium scolependrium var. americanum	original(1)	14	144	153	21	0.9409	0.7619	24.94

species	environmental variables included and regularization multiplier	number of environmental variables	# records used to build model	# independent absences	# independent presences	independent AUC	TPR[1]	percent area predicted suitable with 100% presences correctly predicted
Asplenium scolependrium var. americanum	original(0.5)	14	144	153	21	0.9490	0.7619	30.57
Asplenium scolependrium var. americanum	original+SOLRIS(1)	15	144	143	21	0.9457	0.7619	11.12
Asplenium scolependrium var. americanum	original+SOLRIS(0.5)	15	144	143	21	0.9547	0.7619	15.76

species	environmental variables included and regularization multiplier	number of environmental variables	# records used to build model	# independent absences	# independent presences	independent AUC	TPR[1]	percent area predicted suitable with 100% presences correctly predicted
Asplenium scolependrium var. americanum	**original+cont(1)**	**15**	**144**	**153**	**21**	**0.9807**	**0.7619**	**5.48**
Asplenium scolependrium var. americanum	original+cont(0.5)	15	144	153	21	0.9642	0.7619	8.34
Asplenium scolependrium var. americanum	original+SOLRIS+cont (1)	16	144	143	21	0.9800	0.7619	4.99

species	environmental variables included and regularization multiplier	number of environmental variables	# records used to build model	# independent absences	# independent presences	independent AUC	TPR[1]	percent area predicted suitable with 100% presences correctly predicted
Asplenium scolependrium var. americanum	original+SOLRIS+cont (0.5)	16	144	143	21	0.9570	0.7143	5.85
Castanea dentata	original(1)	14	153	134	275	0.8129	0.8909	10.63
Castanea dentata	original(0.5)	14	153	134	275	0.8374	0.7673	9.82
Castanea dentata	**original+SOLRIS(1)**	**15**	**153**	**124**	**275**	**0.8127**	**0.9055**	**12.53**
Castanea dentata	original+SOLRIS(0.5)	15	153	124	275	0.8287	0.8073	12.11
Castanea dentata	original+cont(1)	15	153	134	275	0.8426	0.8918	15.17
Castanea dentata	original+cont(0.5)	15	153	134	275	0.8537	0.8582	11.46
Castanea dentata	original+SOLRIS+cont (1)	16	153	124	275	0.8399	0.9018	15.06
Castanea dentata	original+SOLRIS+cont (0.5)	16	153	124	275	0.8376	0.8036	17.84
Celtis tenuifolia	original(1)	14	73	141	17	0.9591	0.5294	3.39

species	environmental variables included and regularization multiplier	number of environmental variables	# records used to build model	# independent absences	# independent presences	independent AUC	TPR[1]	percent area predicted suitable with 100% presences correctly predicted
Celtis tenuifolia	**original(0.5)**	**14**	**73**	**141**	**17**	**0.9675**	**0.7059**	**1.95**
Celtis tenuifolia	original+SOLRIS(1)	15	73	131	17	0.956	0.7059	3.18
Celtis tenuifolia	original+SOLRIS(0.5)	15	73	131	17	0.9672	0.6471	1.82
Celtis tenuifolia	original+cont(1)	15	73	141	17	0.9395	0.5882	4.05
Celtis tenuifolia	original+cont(0.5)	15	73	141	17	0.9604	0.5294	2.63
Celtis tenuifolia	original+SOLRIS+cont (1)	16	73	131	17	0.9349	0.5294	4.07
Celtis tenuifolia	original+SOLRIS+cont (0.5)	16	73	131	17	0.9627	0.5294	2.06
Chimaphila maculata	original(1)	14	15	154	9	0.9524	0.6667	0.2
Chimaphila maculata	original(0.5)	14	15	154	9	0.9416	0.5556	0.32
Chimaphila maculata	original+SOLRIS(1)	15	15	144	9	0.9429	0.3333	0.4
Chimaphila maculata	original+SOLRIS(0.5)	15	15	144	9	0.929	0.7778	0.74
Chimaphila maculata	original+cont(1)	15	15	154	9	0.9755	0.8889	0.29
Chimaphila maculata	original+cont(0.5)	15	15	154	9	0.9711	0.8889	0.26
Chimaphila maculata	**original+SOLRIS+cont (1)**	**16**	**15**	**144**	**9**	**0.9830**	**0.8889**	**0.2**

species	environmental variables included and regularization multiplier	number of environmental variables	# records used to build model	# independent absences	# independent presences	independent AUC	TPR[1]	percent area predicted suitable with 100% presences correctly predicted
Chimaphila maculata	original+SOLRIS+cont (0.5)	16	15	144	9	0.9699	0.5556	0.4005
Corallorhiza odontorhiza	original(1)	14	4	42	6	0.8889	1	0.63
Corallorhiza odontorhiza	original(0.5)	14	4	42	6	0.7421	0.3333	5.35
Corallorhiza odontorhiza	original+SOLRIS(1)	15	4	42	6	0.6349	0	3.74
Corallorhiza odontorhiza	original+SOLRIS(0.5)	15	4	42	6	0.4444	0	30.23
Corallorhiza odontorhiza	**original+cont(1)**	**15**	**4**	**42**	**6**	**0.9127**	**1**	**0.32**
Corallorhiza odontorhiza	original+cont(0.5)	15	4	42	6	0.8730	0.6667	0.85
Corallorhiza odontorhiza	original+SOLRIS+cont (1)	16	4	42	6	0.8929	0.5000	1.34
Corallorhiza odontorhiza	original+SOLRIS+cont (0.5)	16	4	42	6	0.6786	0.5000	5.47
Cornus florida	original(1)	14	295	147	89	0.7602	0.6629	10.66
Cornus florida	original(0.5)	14	295	147	89	0.7859	0.6854	9.07
Cornus florida	original+SOLRIS(1)	15	295	137	89	0.7281	0.6629	12.63
Cornus florida	**original+SOLRIS(0.5)**	**15**	**295**	**137**	**89**	**0.775**	**0.7191**	**8.33**
Cornus florida	original+cont(1)	15	295	147	89	0.7289	0.3933	18.99
Cornus florida	original+cont(0.5)	15	295	147	89	0.7787	0.5506	17.31

species	environmental variables included and regularization multiplier	number of environmental variables	# records used to build model	# independent absences	# independent presences	independent AUC	TPR[1]	percent area predicted suitable with 100% presences correctly predicted
Cornus florida	original+SOLRIS+cont (1)	16	295	137	89	0.6937	0.4045	20.1
Cornus florida	original+SOLRIS+cont (0.5)	16	295	137	89	0.7312	0.4719	20.84
Cypripedium arietinum	**original(1)**	**14**	**77**	**41**	**27**	**0.8600**	**0.9259**	**9.82**
Cypripedium arietinum	original(0.5)	14	77	41	27	0.8229	0.8519	17.68
Cypripedium arietinum	original+SOLRIS(1)	15	65	31	27	0.8411	0.8148	22.8
Cypripedium arietinum	original+SOLRIS(0.5)	15	65	31	27	0.8608	0.8148	12.99
Cypripedium arietinum	original+cont(1)	15	77	41	27	0.8157	0.8519	23.37
Cypripedium arietinum	original+cont(0.5)	15	77	41	27	0.8076	0.8889	29.36
Cypripedium arietinum	original+SOLRIS+cont (1)	16	65	31	27	0.8686	0.7407	13.88
Cypripedium arietinum	original+SOLRIS+cont (0.5)	16	65	31	27	0.8692	0.8148	14.81
Enemion biternatum	original(1)	14	14	57	26	0.8576	0.8077	4.98
Enemion biternatum	original(0.5)	14	14	57	26	0.8617	0.7308	7.38

species	environmental variables included and regularization multiplier	number of environmental variables	# records used to build model	# independent absences	# independent presences	independent AUC	TPR[1]	percent area predicted suitable with 100% presences correctly predicted
Enemion biternatum	original+SOLRIS(1)	15	14	47	26	0.8208	0.7308	3.88
Enemion biternatum	original+SOLRIS(0.5)	15	14	47	26	0.8028	0.6923	4.78
Enemion biternatum	**original+cont(1)**	**15**	**14**	**57**	**26**	**0.8623**	**0.9231**	**4.50**
Enemion biternatum	original+cont(0.5)	15	14	57	26	0.8873	0.8077	3.49
Enemion biternatum	original+SOLRIS+cont (1)	16	14	47	26	0.8347	0.7308	6.72
Enemion biternatum	original+SOLRIS+cont (0.5)	16	14	47	26	0.8224	0.6923	3.09
Erigenia bulbosa	original(1)	14	16	23	12	0.4529	0.8333	10.34
Erigenia bulbosa	original(0.5)	14	16	23	12	0.4638	0.5000	12.54
Erigenia bulbosa	original+SOLRIS(1)	15	16	23	12	0.4529	0.4167	9.41
Erigenia bulbosa	original+SOLRIS(0.5)	15	16	23	12	0.4457	0.2500	10.00
Erigenia bulbosa	original+cont(1)	15	16	23	12	0.5217	0.5000	3.37
Erigenia bulbosa	original+cont(0.5)	15	16	23	12	0.5399	0.5833	5.34
Erigenia bulbosa	original+SOLRIS+cont (1)	16	16	23	12	0.3877	0.5000	6.22
Erigenia bulbosa	original+SOLRIS+cont (0.5)	16	16	23	12	0.4384	0.3333	6.93
Frasera caroliniana	original(1)	14	30	155	19	0.8995	0.8421	2.56

species	environmental variables included and regularization multiplier	number of environmental variables	# records used to build model	# independent absences	# independent presences	independent AUC	TPR[1]	percent area predicted suitable with 100% presences correctly predicted
Frasera caroliniana	original(0.5)	14	30	155	19	0.9182	0.7368	3.68
Frasera caroliniana	original+SOLRIS(1)	15	30	145	19	0.8715	0.8421	4.28
Frasera caroliniana	original+SOLRIS(0.5)	15	30	145	19	0.8762	0.7895	3.39
Frasera caroliniana	original+cont(1)	15	30	155	19	0.8927	0.7368	8.86
Frasera caroliniana	original+cont(0.5)	15	30	155	19	0.8815	0.7895	10.32
Frasera caroliniana	**original+SOLRIS+cont (1)**	**16**	**30**	**145**	**19**	**0.8817**	**0.8947**	**10.46**
Frasera caroliniana	original+SOLRIS+cont (0.5)	16	30	145	19	0.8584	0.5789	10.70
Fraxinus quadrifolia	**original(1)**	**14**	**46**	**131**	**65**	**0.8652**	**0.7385**	**4.35**
Fraxinus quadrifolia	original(1)	14	46	131	65	0.8979	0.6769	6.02
Fraxinus quadrifolia	original+SOLRIS(1)	15	46	121	65	0.8432	0.6923	4.69
Fraxinus quadrifolia	original+SOLRIS(0.5)	15	46	121	65	0.8479	0.6154	5.99
Fraxinus quadrifolia	original+cont(1)	15	46	131	65	0.8162	0.6615	4.30

species	environmental variables included and regularization multiplier	number of environmental variables	# records used to build model	# independent absences	# independent presences	independent AUC	TPR[1]	percent area predicted suitable with 100% presences correctly predicted
Fraxinus quadrifolia	original+cont(0.5)	15	46	131	65	0.855	0.6154	12.66
Fraxinus quadrifolia	original+SOLRIS+cont (1)	16	46	121	65	0.8257	0.7077	5.65
Fraxinus quadrifolia	original+SOLRIS+cont (0.5)	16	46	121	65	0.8563	0.5692	8.18
Heuchera americana	original(1)	6	19	156	12	0.9097	0.5000	29.38
Heuchera americana	original(1)	6	19	156	12	0.9225	0.5000	16.97
Heuchera americana	**original+SOLRIS(1)**	**7**	**19**	**146**	**9**	**0.9072**	**0.6667**	**14.90**
Heuchera americana	original+SOLRIS(0.5)	7	19	146	9	0.9323	0.3333	13.83
Heuchera americana	original-climate+cont(1)	7	19	156	12	0.9236	0.4167	14.43
Heuchera americana	original-climate+cont(0.5)	7	19	156	12	0.9343	0.5000	9.31
Heuchera americana	original+SOLRIS+cont (1)	8	19	146	9	0.9224	0.5556	13.66
Heuchera americana	original+SOLRIS+cont (0.5)	8	19	146	9	0.9338	0.4444	11.35
Hydrastis canadensis	original(1)	14	44	156	35	0.8057	0.6571	48.07

species	environmental variables included and regularization multiplier	number of environmental variables	# records used to build model	# independent absences	# independent presences	independent AUC	TPR[1]	percent area predicted suitable with 100% presences correctly predicted
Hydrastis canadensis	**original(1)**	**14**	**44**	**156**	**35**	**0.8258**	**0.6571**	**50.52**
Hydrastis canadensis	original+SOLRIS(1)	15	44	146	35	0.7955	0.5429	19.35
Hydrastis canadensis	original+SOLRIS(0.5)	15	44	146	35	0.8485	0.6000	24.04
Hydrastis canadensis	original+cont(1)	15	44	156	35	0.8332	0.6000	31.74
Hydrastis canadensis	original+cont(0.5)	15	44	156	35	0.8535	0.5429	37.35
Hydrastis canadensis	original+SOLRIS+cont (1)	16	44	146	35	0.8153	0.4571	21.91
Hydrastis canadensis	original+SOLRIS+cont (0.5)	16	44	146	35	0.8188	0.5429	25.07
Juglans cinerea	original(1)	14	1888	138	522	0.4663	0.4023	75.12
Juglans cinerea	original(1)	14	1888	138	522	0.4553	0.3257	74.36
Juglans cinerea	original+SOLRIS(1)	15	1594	129	513	0.4453	0.4366	60.26
Juglans cinerea	original+SOLRIS(0.5)	15	1594	129	513	0.4305	0.3665	59.87
Juglans cinerea	original+cont(1)	15	1888	138	522	0.4824	0.4636	77.24
Juglans cinerea	original+cont(0.5)	15	1888	138	522	0.4458	0.4061	74.12
Juglans cinerea	original+SOLRIS+cont (1)	16	1594	129	513	0.467	0.4854	58.49
Juglans cinerea	**orig+SOLRIS+cont (0.5)**	**16**	**1594**	**129**	**513**	**0.4417**	**0.3918**	**55.61**

species	environmental variables included and regularization multiplier	number of environmental variables	# records used to build model	# independent absences	# independent presences	independent AUC	TPR[1]	percent area predicted suitable with 100% presences correctly predicted
Liparis liliifolia	original(1)	14	16	133	53	0.9628	0.9434	27.19
Liparis liliifolia	original(1)	14	16	133	53	0.9625	0.9434	24.19
Liparis liliifolia	original+SOLRIS(1)	15	16	123	53	0.9646	0.9434	13.22
Liparis liliifolia	original+SOLRIS(0.5)	15	16	123	53	0.9612	0.9245	18.20
Liparis liliifolia	original+cont(1)	15	16	133	53	0.9699	0.9623	31.26
Liparis liliifolia	original+cont(0.5)	15	16	133	53	0.9664	0.9434	17.54
Liparis liliifolia	**original+SOLRIS+cont (1)**	**16**	**16**	**123**	**53**	**0.9712**	**0.9623**	**16.36**
Liparis liliifolia	original+SOLRIS+cont (0.5)	16	16	123	53	0.9693	0.8868	13.65
Lithospermum latifolium	original(1)	14	11	136	23	0.6563	0.7391	21.23
Lithospermum latifolium	**original(1)**	**14**	**11**	**136**	**23**	**0.6643**	**0.7391**	**31.95**
Lithospermum latifolium	original+SOLRIS(1)	15	11	126	23	0.6370	0.4783	26.11
Lithospermum latifolium	original+SOLRIS(0.5)	15	11	126	23	0.6070	0.2609	25.80
Lithospermum latifolium	original+cont(1)	15	11	136	23	0.7008	0.6960	11.76
Lithospermum latifolium	original+cont(0.5)	15	11	136	23	0.7017	0.6957	18.9
Lithospermum latifolium	original+SOLRIS+cont (1)	16	11	126	23	0.6563	0.6087	20.75

species	environmental variables included and regularization multiplier	number of environmental variables	# records used to build model	# independent absences	# independent presences	independent AUC	TPR[1]	percent area predicted suitable with 100% presences correctly predicted
Lithospermum latifolium	original+SOLRIS+cont (0.5)	16	11	126	23	0.5769	0.2609	32.57
Magnolia acuminata	original(1)	14	58	155	12	0.8457	0.2500	1.90
Magnolia acuminata	original(1)	14	58	155	12	0.8887	0.4167	1.41
Magnolia acuminata	**original+SOLRIS(1)**	**15**	**58**	**145**	**12**	**0.8471**	**0.5833**	**1.54**
Magnolia acuminata	original+SOLRIS(0.5)	15	58	145	12	0.8552	0.1667	1.19
Magnolia acuminata	original+cont(1)	15	58	155	12	0.8500	0.3333	2.45
Magnolia acuminata	original+cont(0.5)	15	58	155	12	0.8720	0.2500	1.56
Magnolia acuminata	original+SOLRIS+cont (1)	16	58	145	12	0.8115	0.1667	2.27
Magnolia acuminata	original+SOLRIS+cont (0.5)	16	58	145	12	0.8626	0.1667	1.45
Mertensia virginica	**original(1)**	**14**	**6**	**51**	**28**	**0.8284**	**0.3571**	**17.81**
Mertensia virginica	original(1)	14	6	51	28	0.7983	0.2500	17.81
Mertensia virginica	original+SOLRIS(1)	15	6	41	28	0.7143	0.3571	16.55

species	environmental variables included and regularization multiplier	number of environmental variables	# records used to build model	# independent absences	# independent presences	independent AUC	TPR[1]	percent area predicted suitable with 100% presences correctly predicted
Mertensia virginica	original+SOLRIS(0.5)	15	6	41	28	0.8066	0.2857	11.61
Mertensia virginica	original+cont(1)	15	6	51	28	0.8578	0.3571	15.36
Mertensia virginica	original+cont(0.5)	15	6	51	28	0.8326	0.2500	17.36
Mertensia virginica	original+SOLRIS+cont (1)	16	6	41	28	0.7065	0.3214	17.75
Mertensia virginica	original+SOLRIS+cont (0.5)	16	6	41	28	0.7274	0.2857	14.42
Morus rubra	**original(1)**	**14**	**42**	**156**	**4**	**0.7804**	**0**	**10.15**
Morus rubra	original(1)	14	42	156	4	0.7580	0	12.88
Morus rubra	original+SOLRIS(1)	15	42	146	4	0.6524	0	6.26
Morus rubra	original+SOLRIS(0.5)	15	42	146	4	0.6558	0	5.58
Morus rubra	original+cont(1)	15	42	156	4	0.7484	0	22.00
Morus rubra	original+cont(0.5)	15	42	156	4	0.7243	0	12.81
Morus rubra	original+SOLRIS+cont (1)	16	42	146	4	0.7140	0	8.48
Morus rubra	original+SOLRIS+cont (0.5)	16	42	146	4	0.7123	0	4.68
Phegopteris hexagonoptera	original(1)	14	38	132	10	0.4598	0.6000	30.35
Pheogopteris hexagonoptera	original(1)	14	38	132	10	0.4742	0.4000	31.75

species	environmental variables included and regularization multiplier	number of environmental variables	# records used to build model	# independent absences	# independent presences	independent AUC	TPR[1]	percent area predicted suitable with 100% presences correctly predicted
Pheogopteris hexagonoptera	original+SOLRIS(1)	15	38	122	10	0.3984	0.4000	31.46
Pheogopteris hexagonoptera	original+SOLRIS(0.5)	15	38	122	10	0.3566	0.3000	32.49
Phegopteris hexagonoptera	**original+cont(1)**	**15**	**38**	**132**	**10**	**0.6318**	**0.6000**	**20.17**
Pheogopteris hexagonoptera	original+coont(0.5)	15	38	132	10	0.5977	0.6000	25.37
Pheogopteris hexagonoptera	original+SOLRIS+cont (1)	16	38	122	10	0.4402	0.5000	25.20
Pheogopteris hexagonoptera	original+SOLRIS+cont (0.5)	16	38	122	10	0.4410	0.1000	30.89
Stylophorum diphyllum	original(1)	14	4	154	3	0.7143	0	10.84
Stylophorum diphyllum	original(1)	14	4	154	3	0.7056	0	16.78
Stylophorum diphyllum	original+SOLRIS(1)	15	4	144	3	0.6620	0	9.61
Stylophorum diphyllum	original+SOLRIS(0.5)	15	4	144	3	0.6690	0	10.55
Stylophorum diphyllum	original+cont(1)	15	4	154	3	0.7013	0.3333	11.77
Stylophorum diphyllum	original+cont(0.5)	15	4	154	3	0.7662	0.3333	3.29

species	environmental variables included and regularization multiplier	number of environmental variables	# records used to build model	# independent absences	# independent presences	independent AUC	TPR[1]	percent area predicted suitable with 100% presences correctly predicted
Stylophorum diphyllum	original+SOLRIS+cont (1)	16	4	144	3	0.7106	0.3333	0.75
Stylophorum diphyllum	original+SOLRIS+cont (0.5)	16	4	144	3	0.7407	0.3333	5.68
Trillium flexipes	original-north-soil1[4](1)	12	4	70	2	0.9357	1	0.30
Trillium flexipes	original-north-soil1 (0.5)	12	4	70	2	0.9143	0.5000	2.12
Trillium flexipes	original-north-soil1+SOLRIS(1)	13	4	60	2	0.9667	1	0.05
Trillium flexipes	original-north-soil1+SOLRIS(0.5)	13	4	60	2	0.9500	0	0.21
Trillium flexipes	original-north-soil1+cont(1)	13	4	70	2	0.9571	1	0.07
Trillium flexipes	original-north-soil1+cont(0.5)	13	4	70	2	0.9500	0	0.16
Trillium flexipes	**original+SOLRIS+cont (1)**	**14**	**4**	**60**	**2**	**0.9667**	**1**	**0.02**
Trillium flexipes	original+SOLRIS+cont (0.5)	14	4	60	2	0.9500	0.5000	0.04
Uvularia perfoliata	original(1)	14	9	69	3	0.8068	0.3333	5.03
Uvularia perfoliata	original(1)	14	9	69	3	0.7536	0.3333	7.93

species	environmental variables included and regularization multiplier	number of environmental variables	# records used to build model	# independent absences	# independent presences	independent AUC	TPR[1]	percent area predicted suitable with 100% presences correctly predicted
Uvularia perfoliata	original+SOLRIS(1)	15	9	59	3	0.8305	1	1.72
Uvularia perfoliata	original+SOLRIS(0.5)	15	9	59	3	0.8079	0.3333	2.08
Uvularia perfoliata	original+cont(1)	15	9	69	3	0.913	0.6667	0.32
Uvularia perfoliata	original+cont(0.5)	15	9	69	3	0.9034	0.6667	0.94
Uvularia perfoliata	**original+SOLRIS+cont (1)**	**16**	**9**	**59**	**3**	**0.9379**	**1**	**0.16**
Uvularia perfoliata	original+SOLRIS+cont (0.5)	16	9	59	3	0.9322	0.6667	0.26

[1]The true positive rate (TPR) represents the percent of independent presences of a species correctly predicted as presences by the model when using either a 10% (number of records >10) or a 0% (number of records < 10) omission threshold.

[2]"Original" refers to the 14 climatic, topographic, and soil characteristic environmental variables used to train the models, not including the land cover (SOLRIS) and forest amount (cont.) variables.

[3]Soil texture predictor not used for this species because too many records lie within areas with no data for this variable.

[4]Both soil texture and aspect predictors not used for this species because too many records lie within areas with no data for these variables.

Table 4.1: List of the plant species of conservation concern found in at least one plot during field surveys and whether their models were included in the final efficiency map.

Species Found	Included in Efficiency Map?
Arisaema dracontium	No
Carex muskingumensis	No
Carex virescens	No
Carya glabra	No
Castanea dentata	Yes
Celtis tenuifolia	Yes
. Cornus florida	Yes
Erigenia bulbosa	No
Hybanthus concolor	No
. Lithospermum canescens	No
. Lithospermum latifolium	Yes
Phegopteris hexagonoptera	No
Poa languida	No
Saururus cernuus	No
Verbesina alternifolia	No
Vernonia missurica	No

Table 5.1: Results of the generalized linear models predicting probability of at least one species of conservation concern based on S-SDM score for each of the different weighting procedures.

Weighting	AIC	Null Deviance	Residual Deviance	% Deviance Explained
None	78.130	81.686	73.836	9.6
Model Accuracy	77.592	81.686	73.592	9.9
Species S-rank	78.546	81.686	74.546	8.7
Both	78.375	81.686	74.375	9.0

Figures

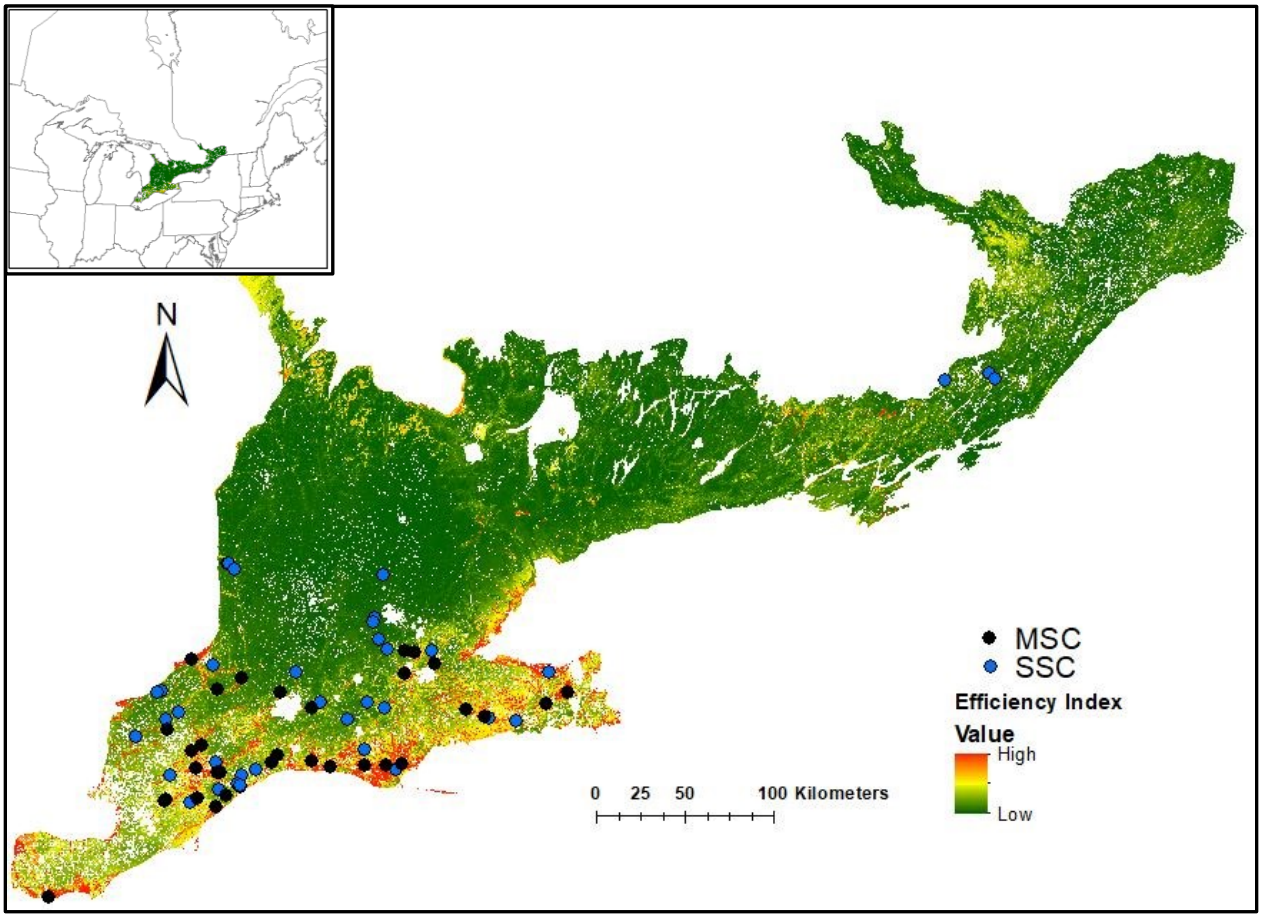
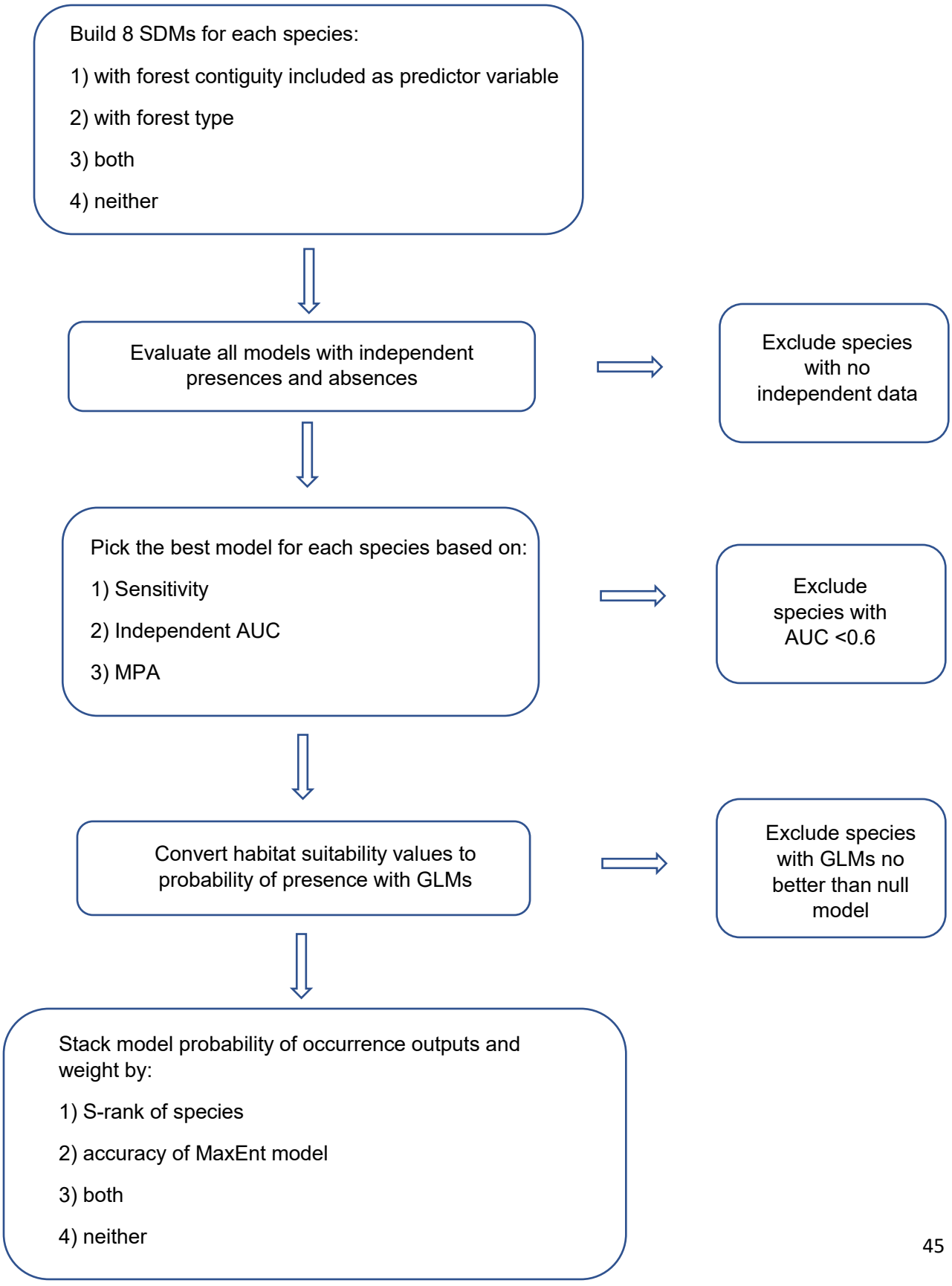


Figure 1.1: Study area showing the range of efficiency index values based on weighting individual model outputs by species rarity and model accuracy with the sites surveyed in 2017 overlaid.

Figure 2.1: Flowchart showing progression of steps taken to attain efficiency maps from Maxent SDMs.



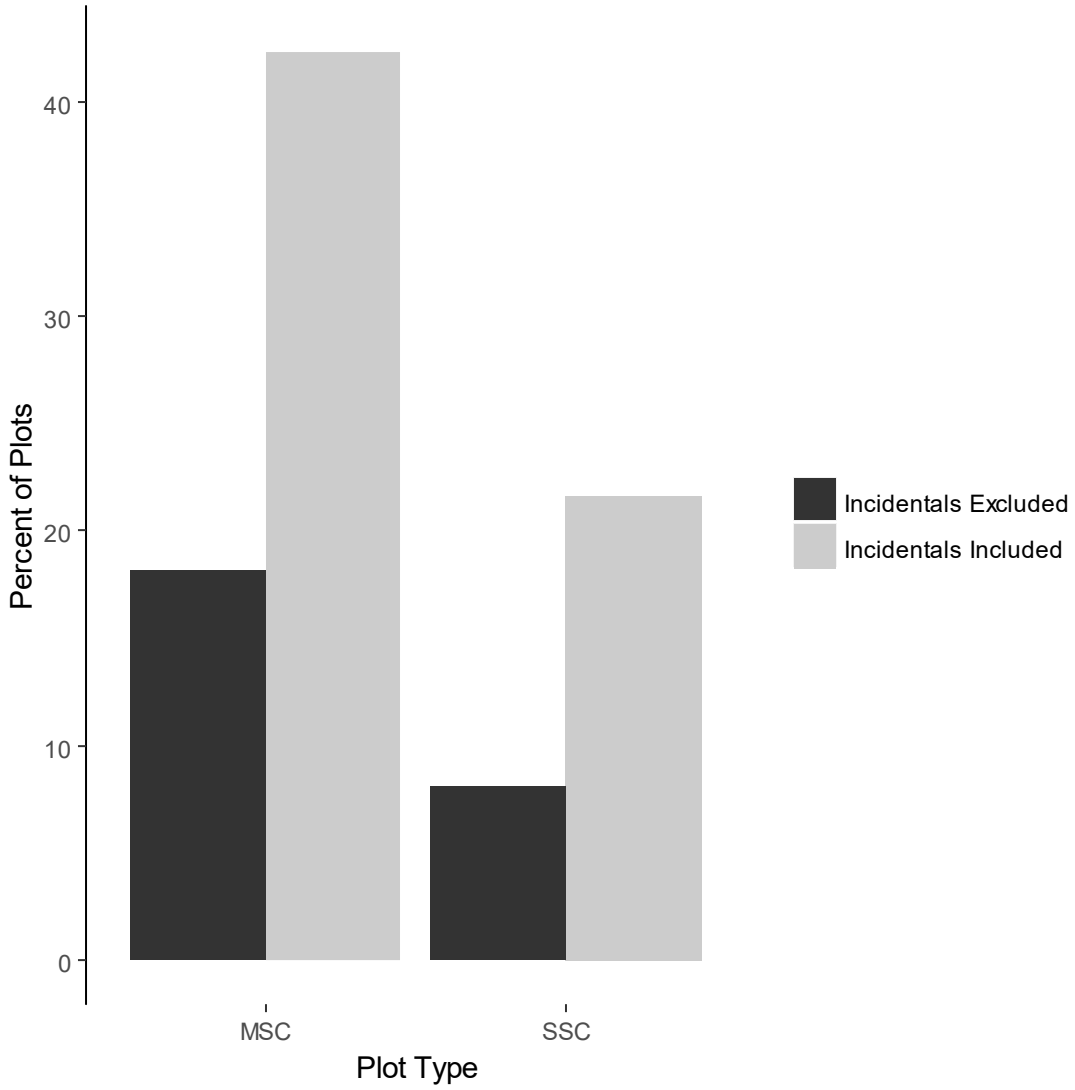


Figure 3.1: Comparison of the percent of multi-species cell (MSC) and single species cell (SSC) plots that had at least one plant species of conservation concern discovered, either including or excluding incidental species discoveries (species which were not modelled).

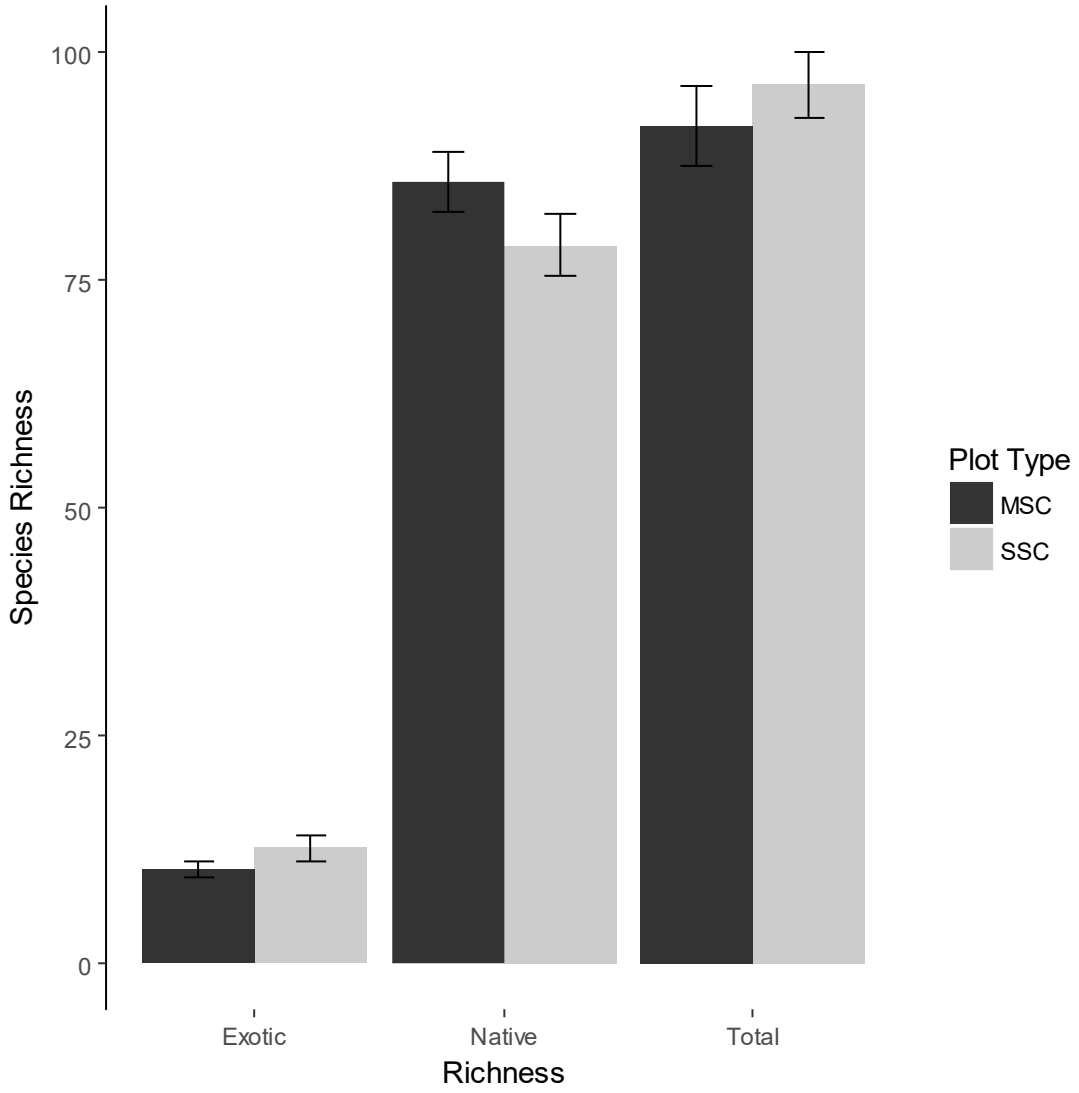


Figure 4.1: Comparison of the total, native, and exotic species richness in multi-species cells and single species cells with standard error bars.

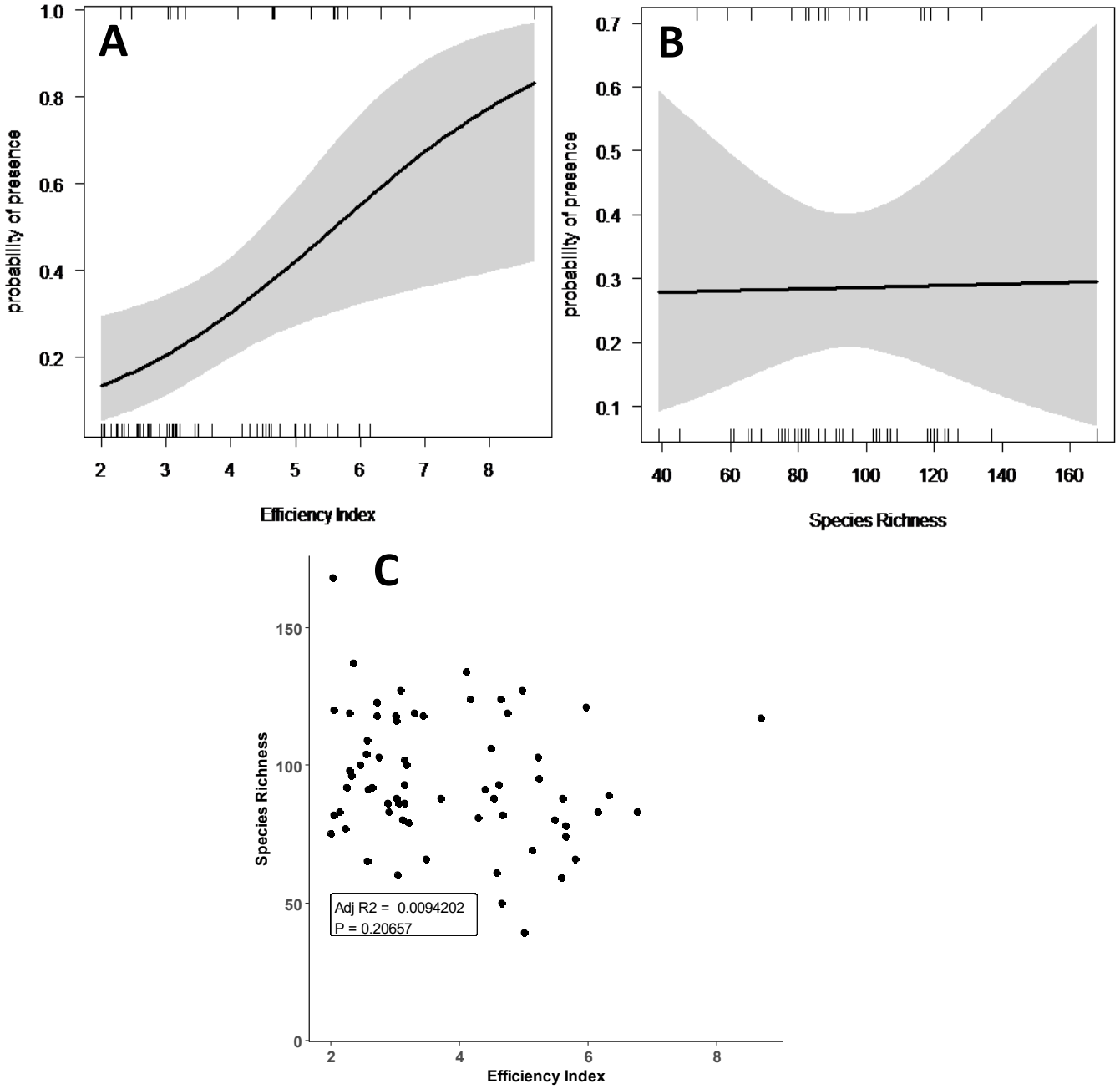


Figure 5.1: Visualization of regression models with 95% confidence intervals for: A) logistic regression showing probability of presence of at least one tracked plant species across efficiency index values of the S-SDM weighted by both rarity and accuracy, B) logistic regression showing probability of presence of at least one threatened plant by species richness, C) linear regression showing species richness of surveyed plots across efficiency index values of the S-SDM weighted by both rarity and accuracy.

References

Albuquerque, F., & Beier, P. (2016). Predicted rarity-weighted richness, a new tool to prioritize sites for species representation. *Ecology and Evolution*, 6(22), 8107-8114.

Algar, A. C., Kharouba, H. M., Young, E. R., & Kerr, J. T. (2009). Predicting the future of species diversity: macroecological theory, climate change, and direct tests of alternative forecasting methods. *Ecography*, 32(1), 22-33.

Amaral, A. G., Munhoz, C. B., Walter, B. M., Aguirre-Gutiérrez, J., & Raes, N. (2017). Richness pattern and phytogeography of the Cerrado's herb-shrub flora and implications for conservation. *Journal of Vegetation Science*. 28, 1-15.

Benito, B. M., Cayuela, L., & Albuquerque, F. S. (2013). The impact of modelling choices in the predictive performance of richness maps derived from species-distribution models: guidelines to build better diversity models. *Methods in Ecology and Evolution*, 4(4), 327-335.

Bennett, J. R. (2014). Comparison of native and exotic distribution and richness models across scales reveals essential conservation lessons. *Ecography*, 37(2), 120-129.

Boakes, E. H., McGowan, P. J., Fuller, R. A., Chang-qing, D., Clark, N. E., O'Connor, K., & Mace, G. M. (2010). Distorted views of biodiversity: spatial and temporal bias in species occurrence data. *PLoS Biology*, 8(6), e1000385.

Burnham, K. P. and Anderson, D. R. (2002). Model selection and inference: a practical information - theoretic approach, 2nd ed. - Springer.

Calabrese, J. M., Certain, G., Kraan, C., & Dormann, C. F. (2014). Stacking species distribution models and adjusting bias by linking them to macroecological models. *Global Ecology and Biogeography*, 23(1), 99-112.

Crins, W.J., Gray, P.A., Uhlig, P.W.C. & Wester, M.C. (2009) The ecosystems of Ontario, part I: ecozones and ecoregions. Ontario Ministry of Natural Resources, Peterborough, ON

D'Amen, M., Dubuis, A., Fernandes, R. F., Pottier, J., Pellissier, L., & Guisan, A. (2015). Using species richness and functional traits predictions to constrain assemblage predictions from stacked species distribution models. *Journal of Biogeography*, 42(7), 1255-1266.

Dunn, J. C., Buchanan, G. M., Stein, R. W., Whittingham, M. J., & McGowan, P. J. (2016). Optimising different types of biodiversity coverage of protected areas with a case study using Himalayan Galliformes. *Biological Conservation*, 196, 22-30.

Dubuis, A., Pottier, J., Rion, V., Pellissier, L., Theurillat, J. P., & Guisan, A. (2011). Predicting spatial patterns of plant species richness: a comparison of direct

macroecological and species stacking modelling approaches. *Diversity and Distributions*, 17(6), 1122-1131.

Elith, J., & Burgman, M. A. (2002). Predictions and their validation: rare plants in the Central Highlands, Victoria, Australia. Predicting species occurrences: issues of accuracy and scale, (eds J.M. Scott, P.J. Heglund, M.L. Morrison, J.B. Haufler, M.G. Raphael, W.A. Wall & F.B. Samson), pp. 303–313. Island Press, Washington, DC.

Elith, J., & Graham, C. H. (2009). Do they? How do they? WHY do they differ? On finding reasons for differing performances of species distribution models. *Ecography*, 32(1), 66-77.

Elith, J., Graham, C. H., Anderson, R. P., Dudík, M., Ferrier, S., Guisan, A., *et al*. (2006). Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, 129-151.

Elith, J., Kearney, M., & Phillips, S. (2010). The art of modelling range-shifting species. *Methods in Ecology and Evolution*, 1(4), 330-342.

Elith, J., & Leathwick, J. R. (2009). Species distribution models: ecological explanation and prediction across space and time. *Annual Review of Ecology, Evolution, and Systematics*, 40, 677-697.

Engler, R., Guisan, A., & Rechsteiner, L. (2004). An improved approach for predicting the distribution of rare and endangered species from occurrence and pseudo-absence data. *Journal of Applied Ecology*, 41(2), 263-274.

Faber-Langendoen, D., J. Nichols, L. Master, K. Snow, A. Tomaino, R. Bittman, G. Hammerson, B. Heidel, L. Ramsay, A. Teucher, and B. Young. (2012). NatureServe Conservation Status Assessments: Methodology for Assigning Ranks. NatureServe, Arlington, VA.

Ferrier, S., & Guisan, A. (2006). Spatial modelling of biodiversity at the community level. *Journal of Applied Ecology*, 43(3), 393-404.

Fielding, A.H. & Bell, J.F. (1997) A review of methods for the assessment of prediction errors in conservation presence/absence models*. Environmental Conservation*, 24, 38–49.

Fois, M., Fenu, G., Lombrana, A. C., Cogoni, D., & Bacchetta, G. (2015). A practical method to speed up the discovery of unknown populations using Species Distribution Models. *Journal for Nature Conservation*, 24, 42-48.

Gastón, A., & García-Viñas, J. I. (2013). Evaluating the predictive performance of stacked species distribution models applied to plant species selection in ecological restoration. *Ecological Modelling*, 263, 103-108.

Gogol-Prokurat, M. (2011). Predicting habitat suitability for rare plants at local spatial scales using a species distribution model. *Ecological Applications*, 21(1), 33-47.

Graham, C. H., Ferrier, S., Huettman, F., Moritz, C., & Peterson, A. T. (2004). New developments in museum-based informatics and applications in biodiversity analysis. *Trends in Ecology & Evolution*, 19(9), 497-503.

Grenouillet, G., Buisson, L., Casajus, N., & Lek, S. (2011). Ensemble modelling of species distribution: the effects of geographical and environmental ranges. *Ecography*, 34(1), 9-17.

Guisan, A., & Rahbek, C. (2011). SESAM—a new framework integrating macroecological and species distribution models for predicting spatio-temporal patterns of species assemblages. *Journal of Biogeography*, 38, 1433-1444.

Guisan, A., Tingley, R., Baumgartner, J. B., Naujokaitis-Lewis, I., Sutcliffe, P. R., *et al*. (2013). Predicting species distributions for conservation decisions. *Ecology Letters*, 16(12), 1424-1435.

Guisan, A., & Theurillat, J. P. (2000). Equilibrium modeling of alpine plant distribution: how far can we go? *Phytocoenologia*, 30(3/4), 353-384.

Guisan, A., & Zimmermann, N. E. (2000). Predictive habitat distribution models in ecology. *Ecological Modelling*, 135(2-3), 147-186.

Guisan, A. , Zimmermann, N. , Elith, J. , Graham, C. , Phillips, S. , and Peterson, A. (2007). What matters for predicting spatial distributions of tree occurrences: techniques, data, or species' characteristics. *Ecological Monographs*, 77, 615–630.

Hanspach, J., Kühn, I., Pompe, S., & Klotz, S. (2010). Predictive performance of plant species distribution models depends on species traits. *Perspectives in Plant Ecology, Evolution and Systematics*, 12(3), 219-225.

Hernandez P.A., Graham C.H., Master L.L., Albert D.L. (2006) The effect of sample size and species characteristics on performance of different species distribution modeling methods. *Ecography*, 29, 773–785

R Core Team (2016) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna

Hijmans, R.J., Phillips S., Leathwick J., and Elith J. (2017). dismo: Species Distribution Modeling. R package version 1.1-4. https://CRAN.R-project.org/package=dismo

Koch, R., Almeida-Cortez, J. S., & Kleinschmit, B. (2017). Revealing areas of high nature conservation importance in a seasonally dry tropical forest in Brazil: Combination of modelled plant diversity hot spots and threat patterns. *Journal for Nature Conservation*, 35, 24-39.

Labay, B. J., Hendrickson, D. A., Cohen, A. E., Bonner, T. H., King, R. S., *et al*. (2015). Can species distribution models aid bioassessment when reference sites are lacking? Tests based on freshwater fishes. *Environmental Management*, 56(4), 835-846.

Le Lay, G., Engler, R., Franc, E., & Guisan, A. (2010). Prospective sampling based on model ensembles improves the detection of rare species. *Ecography*, 33(6), 1015-1027.

Lindenmayer, D. B., Piggott, M. P., & Wintle, B. A. (2013). Counting the books while the library burns: why conservation monitoring programs need a plan for action. *Frontiers in Ecology and the Environment*, 11(10), 549-555.

Liu, C., Newell, G., & White, M. (2016). On the selection of thresholds for predicting species occurrence with presence-only data. *Ecology and Evolution*, 6(1), 337-348.

Loiselle, B. A., Howell, C. A., Graham, C. H., Goerck, J. M., Brooks, T., Smith, K. G., & Williams, P. H. (2003). Avoiding pitfalls of using species distribution models in conservation planning. *Conservation Biology*, 17(6), 1591-1600.

MacDougall, A., & Loo, J. (2002). Land use history, plant rarity, and protected area adequacy in an intensively managed forest landscape. *Journal for Nature Conservation*, 10(3), 171-183.

McCune, J. L. (2016). Species distribution models predict rare species occurrences despite significant effects of landscape context. *Journal of Applied Ecology*, 53(6), 1871-1879.

McCune, J. L., Van Natto, A., & MacDougall, A. S. (2017). The efficacy of protected areas and private land for plant conservation in a fragmented landscape. *Landscape Ecology*, 32(4), 871-882.

McKenney, D. W., Pedlar, J. H., Lawrence, K., Papadopol, P., & Campbell, K. (2014). Hardiness Zones and Bioclimatic Modelling of Plant Species Distributions in North America. In Proceedings of the 2014 Annual Meeting of the International Plant Propagators Society 1085 (pp. 139-148).

Merow, C., Smith, M. J., & Silander, J. A. (2013). A practical guide to MaxEnt for modeling species' distributions: what it does, and why inputs and settings matter. *Ecography*, 36(10), 1058-1069.

Meyer, C., Weigelt, P., & Kreft, H. (2016). Multidimensional biases, gaps and uncertainties in global plant occurrence information. *Ecology Letters*, 19(8), 992-1006.

Miličić, M., Vujić, A., Jurca, T., & Cardoso, P. (2017). Designating conservation priorities for Southeast European hoverflies (Diptera: Syrphidae) based on species distribution models and species vulnerability. *Insect Conservation and Diversity*, 10(4), 354-366.

Moudrý, V., & Šímová, P. (2012). Influence of positional accuracy, sample size and scale on modelling species distributions: a review. *International Journal of Geographical Information Science*, 26(11), 2083-2095.

Newbold, T., Gilbert, F., Zalat, S., El-Gabbas, A., & Reader, T. (2009). Climate-based models of spatial patterns of species richness in Egypt's butterfly and mammal fauna. *Journal of Biogeography*, 36(11), 2085-2095.

Oldham, M.J. & Brinker, S.R. (2009) Rare Vascular Plants of Ontario, 4th edn. Natural Heritage Information Centre, Ontario Ministry of Natural Resources. Peterborough, ON.

Oldham, Michael J. (2017). List of the Vascular Plants of Ontario's Carolinian Zone (Ecoregion 7E). Carolinian Canada and Ontario Ministry of Natural Resources and Forestry. Peterborough, ON. 132 pp.

Parviainen, M., Marmion, M., Luoto, M., Thuiller, W., & Heikkinen, R. K. (2009). Using summed individual species models and state-of-the-art modelling techniques to identify threatened plant species hotspots. *Biological Conservation*, 142(11), 2501-2509.

Pearson, R.G., Raxworthy, C.J., Nakamura, M. & Peterson, A.T. (2007) Predicting species distributions from small numbers of occurrence records: a test case using cryptic geckos in Madagascar. Journal of Biogeography, 34, 102–117

Peterman, W. E., Crawford, J. A., & Kuhns, A. R. (2013). Using species distribution and occupancy modeling to guide survey efforts and assess species status. *Journal for Nature Conservation*, 21(2), 114-121.

Peterson, A. T., Soberon, J., Pearson, R. G., Anderson, R. P., Martínez-Meyer, E., Nakamura, M., & Araujo, M. B. (2011). Ecological niches and geographic distributions. Princeton, NJ: Princeton University Press.

Phillips, S.J., Anderson, R.P. & Schapire, R.E. (2006) Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, 190, 231–259.

Pineda, E., & Lobo, J. M. (2009). Assessing the accuracy of species distribution models to predict amphibian species richness patterns. *Journal of Animal Ecology*, 78(1), 182-190.

Raes, N., Roos, M. C., Slik, J. W., Van Loon, E. E., & Steege, H. T. (2009). Botanical richness and endemicity patterns of Borneo derived from species distribution models. *Ecography*, 32(1), 180-192.

Rebelo, H., & Jones, G. (2010). Ground validation of presence-only modelling with rare species: a case study on barbastelles Barbastella barbastellus (Chiroptera: Vespertilionidae). *Journal of Applied Ecology*, 47(2), 410-420.

Rhoden, C. M., Peterman, W. E., & Taylor, C. A. (2017). Maxent-directed field surveys identify new populations of narrowly endemic habitat specialists. *PeerJ*, 5, e3632.

SARA (Species at Risk Act) (2002). Bill C-5, an Act Respecting the Protection of Wildlife Species at Risk in Canada. Government of Canada, Ottawa, Ontario.

Scherrer, D., Massy, S., Meier, S., Vittoz, P., & Guisan, A. (2017). Assessing and predicting shifts in mountain forest composition across 25 years of climate change. *Diversity and Distributions*, 23(5), 517-528.

Searcy, C. A., & Shaffer, H. B. (2014). Field validation supports novel niche modeling strategies in a cryptic endangered amphibian. *Ecography*, 37(10), 983-992.

Segurado, P., & Araujo, M. B. (2004). An evaluation of methods for modelling species distributions. *Journal of Biogeography*, 31(10), 1555-1568.

Soultan, A., & Safi, K. (2017). The interplay of various sources of noise on reliability of species distribution models hinges on ecological specialisation. *PloS One*, 12(11), e0187906.

Swets, J.A. (1988) Measuring the accuracy of diagnostic systems. *Science*, 240, 1285–1293.

Syphard, A. D., & Franklin, J. (2010). Species traits affect the performance of species distribution models for plants in southern California. *Journal of Vegetation Science*, 21(1), 177-189.

Thuiller, W. (2004). Patterns and uncertainties of species' range shifts under climate change. *Global Change Biology*, 10(12), 2020-2027.

Thuiller, W., Lavorel, S., Araújo, M. B., Sykes, M. T., & Prentice, I. C. (2005). Climate change threats to plant diversity in Europe. *Proceedings of the National Academy of Sciences of the United States of America*, 102(23), 8245-8250.

Tukiainen, H., Bailey, J. J., Field, R., Kangas, K., & Hjort, J. (2017). Combining geodiversity with climate and topography to account for threatened species richness. *Conservation Biology*, 31(2), 364-375.

Tulloch, A. I., Sutcliffe, P., Naujokaitis-Lewis, I., Tingley, R., Brotons, L., Ferraz, K. M. P., et al. (2016). Conservation planners tend to ignore improved accuracy of modelled species distributions to focus on multiple threats and ecological processes. *Biological Conservation*, 199, 157-171.

van Proosdij, A. S., Sosef, M. S., Wieringa, J. J., & Raes, N. (2016). Minimum required number of specimen records to develop accurate species distribution models. *Ecography*, 39(6), 542-552.

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. *Diversity and Distributions*, 15(4), 565-576.

Yu, F., Skidmore, A. K., Wang, T., Huang, J., Ma, K., & Groen, T. A. (2017). Rhododendron diversity patterns and priority conservation areas in China. *Diversity and Distributions*, 23(10), 1143-1156.

Zhang, J., Nielsen, S. E., Grainger, T. N., Kohler, M., Chipchar, T., & Farr, D. R. (2014). Sampling plant diversity and rarity at landscape scales: importance of sampling time in species detectability. *PloS One*, 9(4), e95334.

Zhang, M. G., Zhou, Z. K., Chen, W. Y., Slik, J. F., Cannon, C. H., & Raes, N. (2012). Using species distribution modeling to improve conservation and land use planning of Yunnan, China. *Biological Conservation*, 153, 257-264.

Zimmermann, N. E., Edwards, T. C., Moisen, G. G., Frescino, T. S., & Blackard, J. A. (2007). Remote sensing-based predictors improve distribution models of rare, early successional and broadleaf tree species in Utah. *Journal of Applied Ecology*, 44(5), 1057-1067.

CHAPTER 2: Incorporating older presence data in species distribution models does not negatively impact predictive performance when records are spatially accurate

ABSTRACT

Ensuring that occurrence data are accurate is a critical component to creating useful species distribution models (SDMs). While there is some prejudice against using older records in SDMs because of the belief that these records are less accurate, there is little evidence to support their indiscriminate exclusion. To test the effect of including older presence records in models alongside more recent records on predictive performance, I built SDMs for seven species of rare plants occurring in Ontario with varying proportions of older records (those from the first half of the time span of species' records) - from 10% of the total records used to build the model to 90% of the total records. I selected all records such that they were spatially accurate relative to the resolution of the environmental predictors and the model outputs. I calculated the resulting AUC and sensitivity of the models using data independent from the data used to build the models. While results varied among the species tested, generally there was no clear pattern showing that greater proportions of older records decreased AUC and sensitivity. Thus, the inclusion of spatially accurate older records in SDMs can be useful as a way to increase record sample size. This may be especially useful for threatened species, which tend to have limited recent record availability due to their rarity.

Key-words: SDM, Maxent, rare plants, spatial accuracy, record age, AUC, sensitivity

Introduction

Species distribution modelling has proliferated as a widespread and effective ecological tool in the last two decades. By making use of both species occurrence data and environmental variables across a geographic area, species distribution models (SDMs) can be used to describe the realized niche of a species when our knowledge of such critical information is incomplete (Guisan & Zimmerman 2000). Beyond this, their applications include (but are not limited to) predicting a species' distribution in the future with climate change (e.g. Thuiller 2004, Elith et al. 2010, McKenney et al 2014), predicting invasive species spread (e.g. Peterson et al. 2003, Ficetola et al. 2007, Jiménez-Valverd et al. 2011) and prescribing appropriate management (Bennett 2014), and assessing the efficacy of conservation areas' protection of biodiversity and threatened species (Loiselle et al. 2003, Koch et al. 2017, Amaral et al. 2017). Because of the important implications for species conservation that SDMs can provide, it is crucial to create models that are as accurate as possible.

Two general types of SDM procedures are those requiring both species presence and absence records and those requiring only species presence records, which are often obtained from natural history collections such as museums, herbaria, etc. True absences for species can be difficult to obtain because they require diligent search efforts to confirm species absence from a site of specified area (Hirzel et al. 2001, Li et al. 2011). Hence, presence-only methods of modelling are quite prominent in the SDM literature and there are now many algorithms to choose from (Elith et al. 2006). Of the available techniques, Maxent (Phillips et al. 2006) has been shown to accurately predict species presences (Elith et al. 2006, Elith & Graham 2009) even with few presence

records available to train the model (Hernandez et al. 2006, Pearson et al. 2007, van Proosdij et al. 2015). This can be especially useful for rare species, which tend to have few available data points in biodiversity databases and natural history collections.

Even with the generally good performance of Maxent presence-only models, modelers must be careful when selecting presence records for distribution modelling (De Giovanni et al. 2012). Occurrence records are subject to error and bias, both of which can impact model predictions. Types of error include those that are spatial, usually caused by errors in the georeferencing process or in the original description of the species' location, and those that are taxonomic because of misidentification or outdated nomenclature (Graham et al. 2004). The degree to which these errors decrease model accuracy can vary (Graham et al. 2008, Moudrý & Šímová 2012, Tulowiecki et al. 2014, Hayes et al. 2015, Costa et al. 2015, Mitchell et al. 2017, Soultan & Safi 2017) but can be accounted for using various techniques within or in addition to an SDM framework (Feeley & Silman 2010, Velásquez-Tibatá et al. 2016, Hefley et al. 2017). Similarly, the effects of several forms of geographical sampling bias of presence records have been explored in the literature including regional bias (Wolmarans et al. 2010), roadside bias (Kadmon et al. 2004), and bias related to political boundaries (Barnes et al. 2014), with several methods of bias correction available (Kramer-Schadt et al. 2013, Pardo et al. 2013, Syfert et al. 2013, Fourcade et al. 2014, Fithian et al. 2015, Stolar & Nielson 2015).

One aspect of presence record selection that has received very little attention is the age of the records used to train the model. With natural history collections acting as one of the main sources of presence information for SDMs and with these collections

typically containing records that extend from the present-day to decades past, there is great variation in record age available for model creation. There is a general sense among many modelers that older records are less reliable than more recent ones and should not be included in SDMs (e.g. Lütolf et al. 2006, Boitani et al. 2011, Tukiainen et al. 2017) or should be down-weighted (Tessarolo et al. 2017). Much of this skepticism concerning older records stems from the often lower spatial accuracy associated with these records in comparison with newer ones which benefit from the recent advances in GPS technology. Indeed, Reside et al. (2011) found that including historic species records in SDMs with coarse resolution decreased model performance compared with using only recent species records with high resolution.

While it is true that older records usually do have less geographic precision, this is not always the case. In situations where detailed field notes were kept, older records can have comparable spatial accuracy to new records. This may be especially relevant for rare species which typically have fewer records available with which to build models, thus it would be highly undesirable to discard any available data unnecessarily. For this reason, the issue of spatial accuracy and record age should not be conflated.

The goal of this study was to determine the extent to which the age of presence records used to build presence-only SDMs affects the predictive performance of resulting models when controlling for the issue of spatial accuracy of records. I built SDMs for seven rare plant species of Ontario using varying proportions of old and new data, and measured the resulting accuracy of these models using independently collected presence and absence records. All the records I used were accurate to within 100 meters. Thus, I could analyze the effect of using differing amounts of older (but still

high resolution) records on model performance. I tested the hypothesis that using increasing amounts of older data relative to newer data will result in decreased model sensitivity and AUC.

Methods

I modeled seven species of vascular plants native to Ontario that occur in woodland habitats (Table 1). These species are all considered threatened in the province, i.e. they are designated as being vulnerable (Natureserve S-rank = S3), or imperiled (S2) in Ontario (Faber-Langendoen et al. 2012). I did not include critically imperiled (S1) species because there were not enough records for any of these species. I used threatened species because these tend to have fewer records available, thus older records may be of higher value, if they can indeed be useful for predicting current occurrences. While threatened, the seven species used here were selected because they have enough records (>100) to build models while only using a subset of the records (see below). I obtained presence-only occurrence records for each species from the Ontario Natural Heritage Information Centre (NHIC) which consist of herbarium records, field surveys by government biologists, and other confirmed sightings. I used only records that had a spatial accuracy of 100 m or less, as 100m x 100m was the resolution of the environmental predictors and the resulting distribution maps. I used fourteen environmental variables as predictors for the models including climatic, topographic, soil, and surficial geology data (Table 2). These variables were minimally correlated with one another ($r < 0.7$).

I divided records for each species into “old” and “new” categories based on their year of observation. To do this, I divided the number of years the records for a species

spanned by two and categorized those that were contained in the more recent years as “new” and those that were contained in the older half as “old.” With this method, the years determined to be the old vs. new cutoffs differed among the seven species with the cutoffs ranging from 1980 to 1995 (Figure 1). Although I could have chosen one predetermined year to use as the cutoff for all species, I chose different years for each species realizing that “old” and “new” is a relative term when considering species presence records, depending on how far back in time the data stretch for a species.

I used five combinations of old and new data to build the models for each of the seven species: 50% new 50% old (50n50o), 75% new 25% old (75n25o), 90% new 10% old (90n10o), 75% old 25% new (75o25n), and 90% old 10% new (90o10n). I kept the number of records for each of these categories constant (or n-1 if there was an odd number of records) to avoid introducing a confounding factor of sample size. I created 20 models for each age category for each species with the records used for each model iteration being randomly chosen each time.

I used Maxent (Phillips et al. 2006) to build the SDMs. Maxent has been shown to create accurate SDMs in comparison with other techniques (Elith & Graham 2009) using only presence records, even when few records are available for model building (Hernandez et al. 2006, Pearson et al. 2007, van Proosdij et al. 2015). I used the raw output format which is the Maxent exponential model itself and is a measure of relative habitat suitability. (Phillips et al. 2017).

I evaluated each model using independent presence and absence data originating from two sources: 1) independent presences were obtained from the NHIC from their central holdings database and 2) independent absences (and occasionally

presences) were obtained from 2014 and 2015 field data in which botanists surveyed 156 100m x 100m cells throughout Southern Ontario predicted to be suitable for one or more plant species at-risk (see McCune 2016; McCune et al. 2017; Chapter 1). I excluded all records in the independent data that were duplicates of records used to build the original models and thus, all of the independent data used to evaluate models were different from the data used to build them. This ensures true validation of the models as opposed to simply using a subset of the original data for testing (Newbold et al. 2010). For each model, I used two methods to measure predictive performance: AUC and sensitivity. AUC is a threshold independent measure that ranges from 0 to 1 where 1 indicates a model that perfectly discriminates between presences and absences and 0.5 indicates a model that does no better than random at discriminating between the two (Fielding & Bell 1997). Sensitivity is a threshold dependent measure that calculates the percent of actual species presences correctly classified as presences by the model. Here, the threshold chosen for sensitivity analysis was that which maximizes the true positive and true negative rates of the model. Together, both these measures, AUC and sensitivity, can demonstrate the accuracy of the models.

For each species, I completed a one-way ANOVA comparing both AUC and sensitivity among SDMS built with varying proportions of new and old data. Where there was significance, I completed a post-hoc Tukey analysis to check for specific differences. I completed model building, evaluation, and statistical analysis in R 3.3.1 (R Core Team 2016) with the dismo (Hijmans et al. 2016) and rJava packages (Urbanek 2017).

Results

The mean independent AUC values for the seven species ranged from 0.66 for *Arisaema dracontium* to 0.90 for *Magnolia acuminata* while the mean sensitivity values ranged from 0.71 for *Celtis tenuifolia* to 0.99 for *M. acuminata.* Regardless of age, generally *M. acuminata* had the best performing models in terms of AUC and sensitivity while *A. dracontium* had the worst, generally below the AUC value of 0.7 widely recognized as indicating an acceptable model (Swets 1986).

The relative effect of the inclusion of older presence records on the AUC and sensitivity values of the distribution models differed among the seven species (Figures 2 and 3). For *M. acuminata* and *C. tenuifolia*, there was no significant difference among the SDMs based on varying proportions of old and new data in terms of AUC or sensitivity. Generally, for *C. dentata, A. scolopendrium var. americanum*, and *C. arietinum*, models had higher AUC values when a greater proportion of new presence records were included. The opposite effect was observed for two species, *C. florida* and *A. dracontium*, which generally had models with higher AUC values using more older records.

Using sensitivity as a measure of model performance, most species exhibited no difference among the five age categories. Only one species, *C. florida*, had models with higher sensitivity rates when more new data were used for model training. Conversely, two species*, C. dentata* and *C. arietinum*, generally had models with higher sensitivity when a greater amount of older data was used.

There was no species for which both the AUC and sensitivity of the models was higher with greater proportions of new data. Similarly, there was also no species that had significantly higher AUC and sensitivity values for models including greater amounts of old data. In fact, three species showed opposing results for the two model performance measures.

Discussion

My results indicate that using relatively older records alongside different proportions of newer records to train SDMs does not uniformly decrease the predictive performance of these models with regard to AUC or sensitivity. As long as the data are spatially accurate relative to the resolution of the environmental predictors, the age of the records has minimal influence on model accuracy. This is exemplified by the fact that none of the seven species tested here had significantly decreased AUC and significantly decreased sensitivity with increasing proportions of older data (although one or the other effect was observed for four of the species). Furthermore, two species exhibited no difference between the different age categories of data for either AUC or sensitivity and two additional species exhibited no difference for sensitivity.

One possible explanation for the lack of a significant detrimental effect of the older data on the accuracy of the distribution models could be the timescale of the environmental predictors relative to the dates of the occurrence records used to build the models. Because the output of an SDM only denotes species distribution for a specific moment in time, it is important that the temporal origin of the two input data types match as much as is feasible (Jiménez-Valverde et al. 2008, Roubicek et al. 2010, Moudrý & Šímová, 2012). However, if the values of the environmental variables

used in the models are similar between the dates of the oldest records and the dates of the newest records, then using older records should not present a problem for the purpose of creating models with high predictive performance. For example, for the species tested here two of the most important predictors in terms of both percent contribution and permutation importance (measures indicating the amount each of the environmental variables contributes to the fitting of the model) for the Maxent models were soil texture and surficial geology. Neither of these environmental predictors would have changed across the study area in any appreciable way over the timescale of the occurrence records (<100 years). In situations where this is not true (e.g. using forest cover as an environmental predictor in regions where large amounts of deforestation have occurred) modelers should exhibit caution with record selection.

One concern with using older records in SDMs is the strong possibility that the species is no longer present at the historical locations on record and thus the SDM could be biased towards a species' potential distribution rather than a realized distribution. For certain applications, such as highlighting potential areas for species translocations, this result may be desirable (Guisan et al. 2013). However, even in other situations where the goal is to obtain a representation of the realized distribution (for the purpose of discovering new populations, for example), a species' absence at all or some of the locations where historical occurrences were found may not be harmful to the model output. A species may be absent from its historical locations not because of changes in the environmental variables used to build the model, but because of other factors such as poor dispersal ability and biotic interactions (Araujo & Guisan 2006, Gogul-Prokurat 2011, Wisz et al. 2013) all of which are infrequently used in ecological

niche modeling. In this case and assuming the current realized niche corresponds to the historical realized niche, the environmental conditions present at the known historical occurrence points would still be representative of the species' ecological niche and thus these records could still be useful to include in the model.

The critical component to making use of older records in SDMs is ensuring their spatial accuracy relative to the resolution of the environmental predictors. Indeed, when older records are not spatially accurate, their inclusion in SDMs has been shown to decrease model predictive performance (Reside et al. 2011). Similarly, including coarse resolution species data regardless of age can decrease model accuracy even with an appreciable increase in sample size (Moudrý & Šímová, 2012). However, through a fairly simple data cleaning procedure, it is possible to remove these high uncertainty occurrences. This will undoubtedly greatly limit the amount of older data that be used for model building since older records were collected in a time before GPS technology. For example, in this study up to 85% of older records were deleted because of inadequate spatial accuracy compared to up to 23% of newer records. Even with this limitation, however, the resulting increase in sample size of total records available for model building makes the inclusion of older data valuable.

The ability to use older occurrence records for species distribution modeling is especially important for threatened taxa. These species typically have few recent presence records available for use in modeling due to their rarity. Including older, spatially accurate records for rare species could help to increase the number of presences available for building presence-only SDMs to an acceptable level (minimum sample sizes vary depending on modelling method and taxa used: Stockwell &

Peterson 2002, Pearson et al. 2007, van Proosdij et al. 2016) which is critical given that using more occurrence records typically leads to more accurate models (e.g. Hernandez et al. 2006, Wisz et al. 2008, Loiselle et al. 2008, Feeley et al. 2011, Moudrý & Šímová, 2012). Additionally, rare species are some of the most important to create accurate models for so that additional populations can be found if they exist and their regional and global conservation status can be better understood and qualified (Guisan et al. 2013). In these cases, the exclusion of older records may do more harm than good. Even for species that are not considered threatened, the potential to use older records may be highly beneficial when there are few records of recent origin in biodiversity databases, which is often the case for plant species (Meyer et al. 2016), or when older museum records have better spatial coverage (Boakes et al. 2010).

Tables

Table 1.2: Plant species included in the analysis, including their S-Rank (NatureServe conservation status) in Ontario, the total number of presence records available from the NHIC with which to build models, and the number of independent presence and absence records available with which to test the accuracy of the models.

Species	Common Name	S-Rank	Number of Cells	Number of Independent Records
Arisaema dracontium	Green Dragon	S3	105	250
Asplenium scolependrium var. americanum	Hart's-tongue Fern	S3	159	139
Castanea dentata	American Chestnut	S2	208	409
Celtis tenuifolia	Dwarf Hackberry	S2	393	158
Cornus florida	Flowering Dogwood	S2	659	236
Cypripedium arietinum	Ram's-head Lady's-slipper	S3	131	68
Magnolia acuminata	Cucumber Tree	S2	217	167

Table 2.2: Environmental variables used in the models for all species.

Variable	Unit	Source	Reference/Access
Elevation	meters	Canadian Digital Elevation Model	http://geogratis.gc.ca/
Slope	degrees	Canadian Digital Elevation Model	http://geogratis.gc.ca/
Aspect	unitless	Canadian Digital Elevation Model	http://geogratis.gc.ca/
Soil Texture	Categorical, 24 categories	Soil Survey Complex, Ontario Ministry of Agriculture	https://www.ontario.ca/page/land-information-ontario
Soil Drainage	Categorical, 9 categories	Soil Survey Complex, Ontario Ministry of Agriculture	https://www.ontario.ca/page/land-information-ontario
Surficial Geology	Categorical, 40 categories	Canada Forest Service	https://www.ontario.ca/page/land-information-ontario
Isothermality	Percentage	Canada Forest Service	McKenney et al. 2011
Mean temperature of wettest quarter	°C	Canada Forest Service	McKenney et al. 2011
Annual precipitation	mm	Canada Forest Service	McKenney et al. 2011
Precipitation seasonality	percentage	Canada Forest Service	McKenney et al. 2011

Precipitation of warmest quarter	mm	Canada Forest Service	McKenney et al. 2011
Total precipitation for growing season	mm	Canada Forest Service	McKenney et al. 2011
Annual mean temperature	°C	Canada Forest Service	McKenney et al. 2011
Mean temperature of growing season	°C	Canada Forest Service	McKenney et al. 2011

Figures

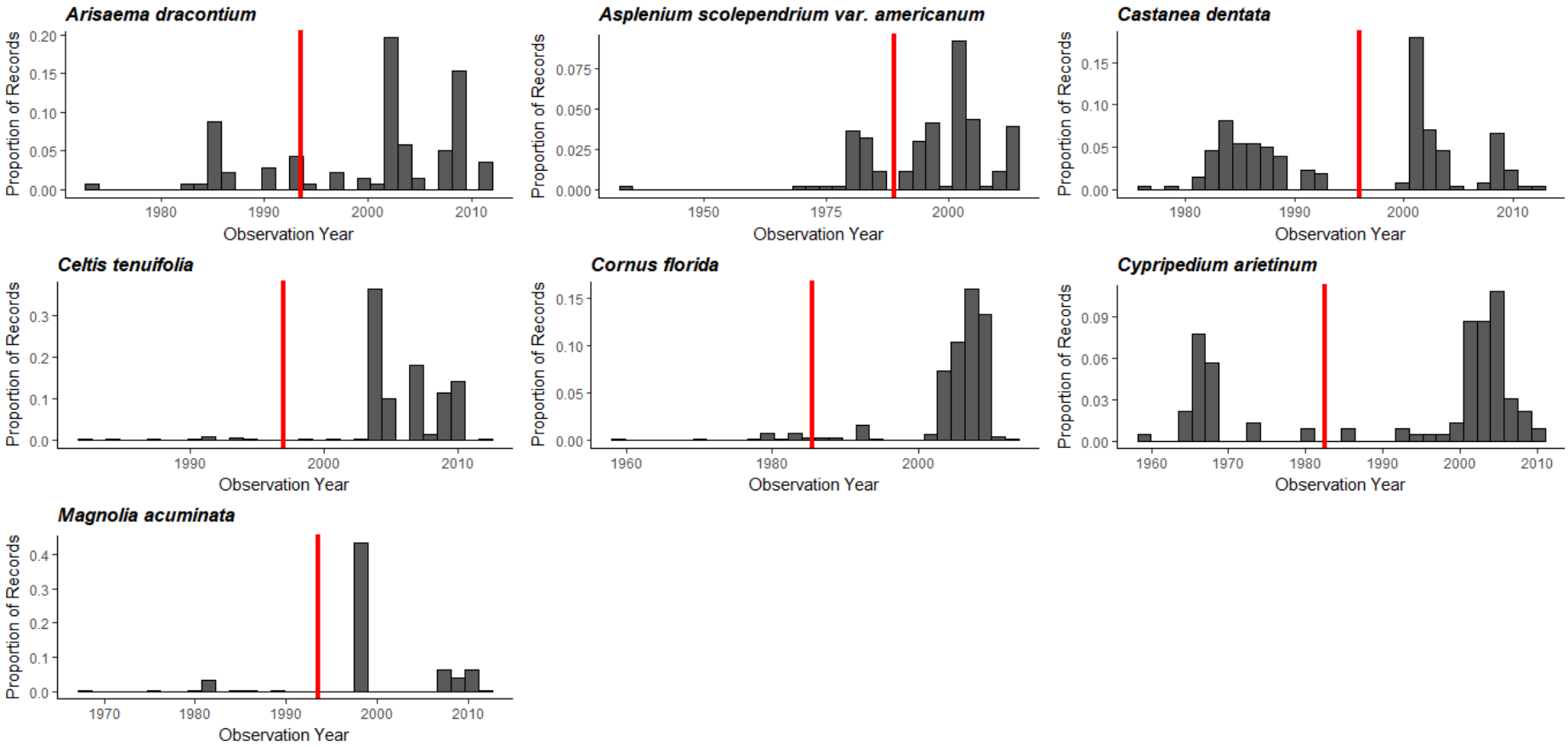


Figure 1.2: Distribution of records for each species by year of observation. Records to the left of the red line are considered “old” while records to the right are considered “new.”

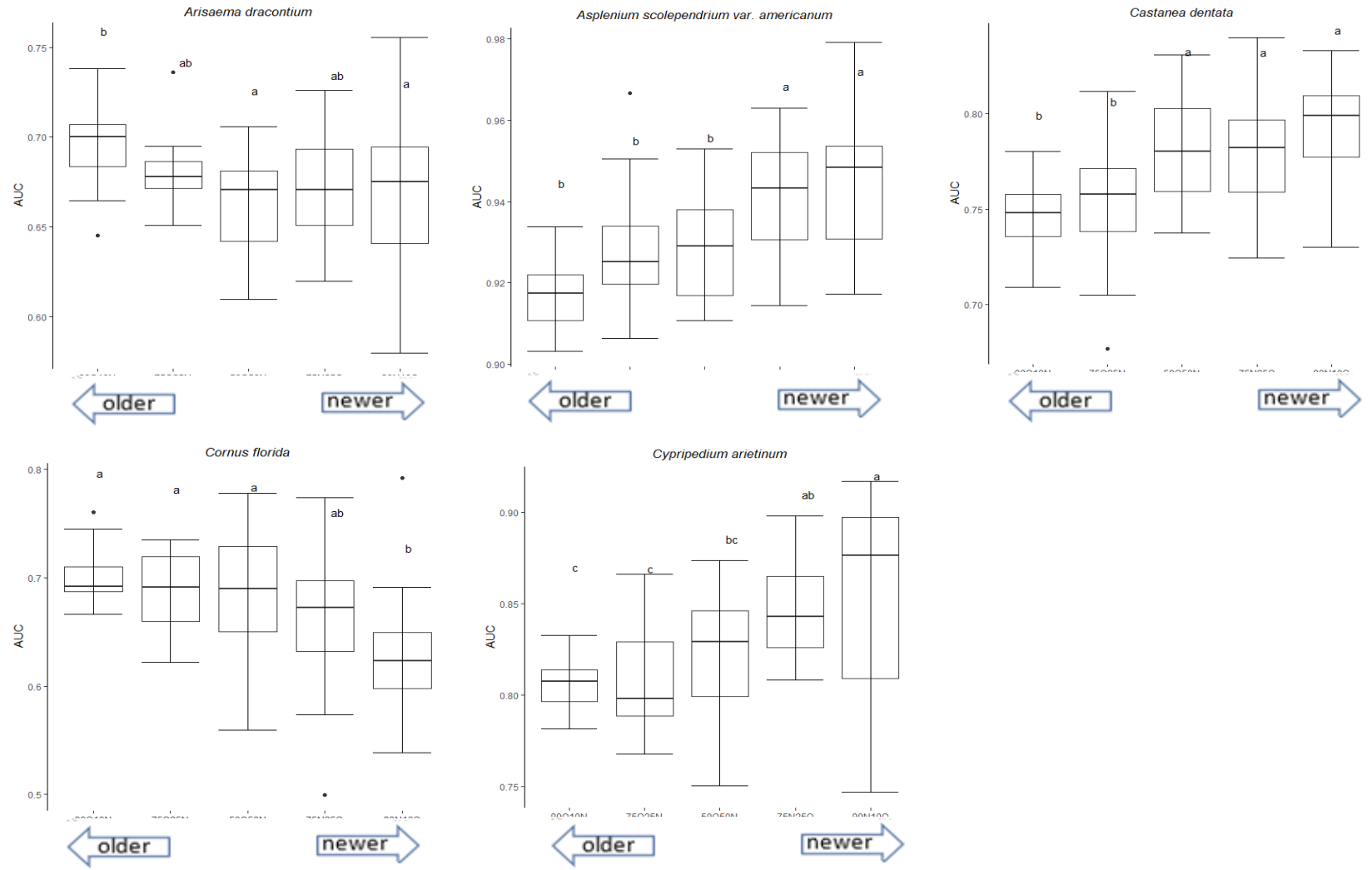


Figure 2.2: Boxplots showing the distribution of AUC values models built using different proportions of old and new occurrence records for species where there were significant differences between the model types. Letters denote significant differences between age categories.

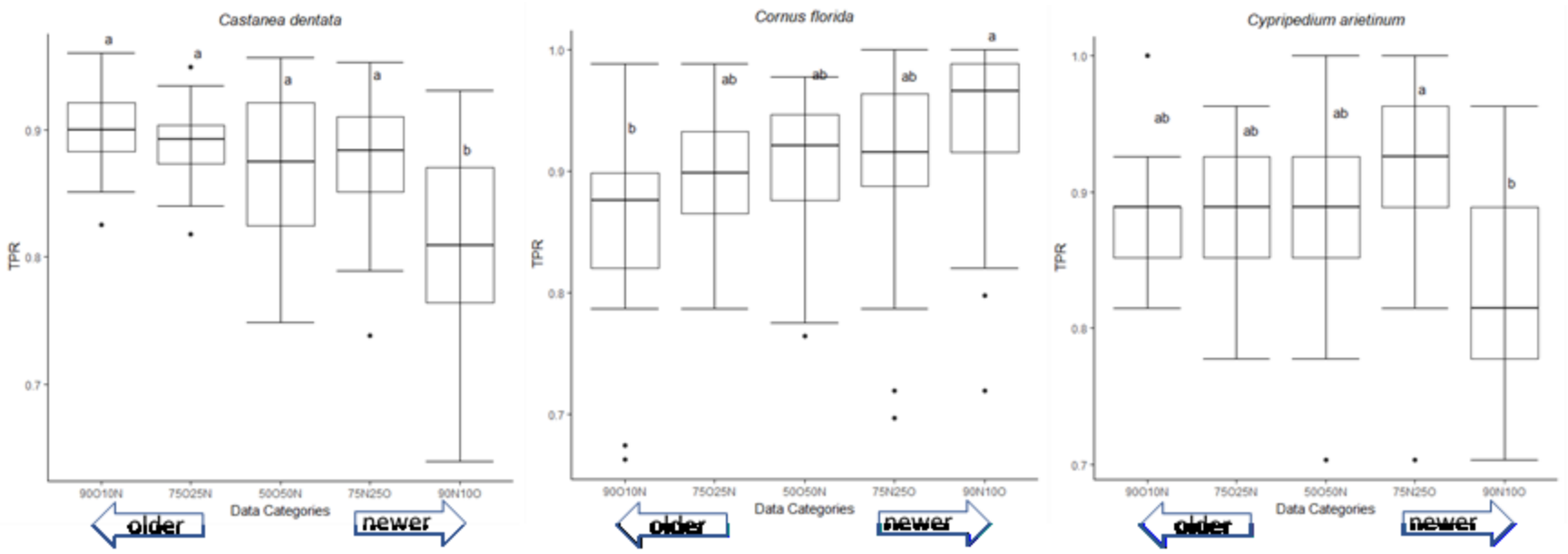


Figure 3.2: Boxplots showing the distribution of sensitivity (true positive rate, TPR) values for models built using different proportions of old and new occurrence records for species where there were significant differences between the model types. Letters denote significant differences between age categories.

References

Amaral, A. G., Munhoz, C. B., Walter, B. M., Aguirre-Gutiérrez, J., & Raes, N. (2017). Richness pattern and phytogeography of the Cerrado's herb-shrub flora and implications for conservation. *Journal of Vegetation Science*, 28, 1-15.

Araujo, M. B., & Guisan, A. (2006). Five (or so) challenges for species distribution modelling. *Journal of Biogeography*, 33(10), 1677-1688.

Barnes, M. A., Jerde, C. L., Wittmann, M. E., Chadderton, W. L., Ding, J., *et al.* (2014). Geographic selection bias of occurrence data influences transferability of invasive *Hydrilla verticillata* distribution models. *Ecology and Evolution*, 4(12), 2584-2593.

Bennett, J. R. (2014). Comparison of native and exotic distribution and richness models across scales reveals essential conservation lessons. *Ecography*, 37(2), 120-129.

Boitani, L., Maiorano, L., Baisero, D., Falcucci, A., Visconti, P., & Rondinini, C. (2011). What spatial data do we need to develop global mammal conservation strategies?. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 366(1578), 2623-2632.

Costa, H., Foody, G. M., Jiménez, S., & Silva, L. (2015). Impacts of species misidentification on species distribution modeling with presence-only data. *ISPRS International Journal of Geo-Information*, 4(4), 2496-2518.

De Giovanni, R., Bernacci, L. C., de Siqueira, M. F., & Rocha, F. S. (2012). The real task of selecting records for ecological niche modelling. *Natureza & Conservação*, 10, 139-144.

Elith, J., & Graham, C. H. (2009). Do they? How do they? WHY do they differ? On finding reasons for differing performances of species distribution models. *Ecography*, 32(1), 66-77.

Elith, J., Graham, C. H., Anderson, R. P., Dudík, M., Ferrier, S., Guisan, A., *et al*. (2006). Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, 129-151.

Elith, J., Kearney, M., & Phillips, S. (2010). The art of modelling range-shifting species. *Methods in Ecology and Evolution*, 1(4), 330-342.

Faber-Langendoen, D., J. Nichols, L. Master, K. Snow, A. Tomaino, R. Bittman, G. Hammerson, B. Heidel, L. Ramsay, A. Teucher, and B. Young. (2012). NatureServe Conservation Status Assessments: Methodology for Assigning Ranks. NatureServe, Arlington, VA.

Feeley, K. J., & Silman, M. R. (2011). Keep collecting: accurate species distribution modelling requires more collections than previously thought. *Diversity and Distributions*, 17(6), 1132-1140.

Fithian, W., Elith, J., Hastie, T., & Keith, D. A. (2015). Bias correction in species distribution models: pooling survey and collection data for multiple species. *Methods in Ecology and Evolution*, 6(4), 424-438.

Ficetola, G. F., Thuiller, W., & Miaud, C. (2007). Prediction and validation of the potential global distribution of a problematic alien invasive species—the American bullfrog. *Diversity and Distributions*, 13(4), 476-485.

Fielding, A.H. & Bell, J.F. (1997) A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environmental Conservation*, 24, 38–49.

Fourcade, Y., Engler, J. O., Rödder, D., & Secondi, J. (2014). Mapping species distributions with MAXENT using a geographically biased sample of presence data: a performance assessment of methods for correcting sampling bias. *PloS One*, 9(5), e97122.

Gogol-Prokurat, M. (2011). Predicting habitat suitability for rare plants at local spatial scales using a species distribution model. *Ecological Applications*, 21(1), 33-47.

Graham, C. H., Elith, J., Hijmans, R. J., Guisan, A., Townsend Peterson, A., Loiselle, B. A., & NCEAS Predicting Species Distributions Working Group. (2008). The influence of spatial errors in species occurrence data used in distribution models. *Journal of Applied Ecology*, 45(1), 239-247.

Graham, C. H., Ferrier, S., Huettman, F., Moritz, C., & Peterson, A. T. (2004). New developments in museum-based informatics and applications in biodiversity analysis. *Trends in Ecology & Evolution*, 19(9), 497-503.

Guisan, A., Tingley, R., Baumgartner, J. B., Naujokaitis-Lewis, I., Sutcliffe, P. R., *et al*. (2013). Predicting species distributions for conservation decisions. *Ecology Letters*, 16(12), 1424-1435.

Guisan, A., & Zimmermann, N. E. (2000). Predictive habitat distribution models in ecology. *Ecological Modelling*, 135(2-3), 147-186.

Hayes, M. A., Ozenberger, K., Cryan, P. M., & Wunder, M. B. (2015). Not to put too fine a point on it—does increasing precision of geographic referencing improve species distribution models for a wide-ranging migratory bat? *Acta Chiropterologica*, 17(1), 159-169.

Hefley, T. J., Brost, B. M., & Hooten, M. B. (2017). Bias correction of bounded location errors in presence-only data. *Methods in Ecology and Evolution*. 8, 1566-1573.

Hernandez P.A., Graham C.H., Master L.L., Albert D.L. (2006) The effect of sample size and species characteristics on performance of different species distribution modeling methods. *Ecography*, 29, 773–785

Hijmans, R.J., Phillips S., Leathwick J., and Elith J. (2017). dismo: Species Distribution Modeling. R package version 1.1-4. https://CRAN.R-project.org/package=dismo

Hirzel, A. H., Helfer, V., & Metral, F. (2001). Assessing habitat-suitability models with a virtual species. *Ecological Modelling*, 145(2-3), 111-121.

Jiménez-Valverde, A., Lobo, J. M., & Hortal, J. (2008). Not as good as they seem: the importance of concepts in species distribution modelling. *Diversity and Distributions*, 14(6), 885-890.

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. *Biological Invasions*, 13(12), 2785-2797.

Kadmon, R., Farber, O., & Danin, A. (2004). Effect of roadside bias on the accuracy of predictive maps produced by bioclimatic models. *Ecological Applications*, 14(2), 401-413.

Koch, R., Almeida-Cortez, J. S., & Kleinschmit, B. (2017). Revealing areas of high nature conservation importance in a seasonally dry tropical forest in Brazil: Combination of modelled plant diversity hot spots and threat patterns. *Journal for Nature Conservation*, 35, 24-39.

Kramer-Schadt, S., Niedballa, J., Pilgrim, J. D., Schröder, B., Lindenborn, (2013). The importance of correcting for sampling bias in MaxEnt species distribution models. *Diversity and Distributions*, 19(11), 1366-1379.

Li, W., Guo, Q., & Elkan, C. (2011). Can we model the probability of presence of species without absence data?. *Ecography*, 34(6), 1096-1105.

Loiselle, B. A., Howell, C. A., Graham, C. H., Goerck, J. M., Brooks, T., Smith, K. G., & Williams, P. H. (2003). Avoiding pitfalls of using species distribution models in conservation planning. *Conservation Biology*, 17(6), 1591-1600.

Loiselle, B. A., Jørgensen, P. M., Consiglio, T., Jiménez, I., Blake, J. G., Lohmann, L. G., & Montiel, O. M. (2008). Predicting species distributions from herbarium collections: does climate bias in collection sampling influence model outcomes?. *Journal of Biogeography*, 35(1), 105-116.

Lütolf, M., Kienast, F., & Guisan, A. (2006). The ghost of past species occurrence: improving species distribution models for presence-only data. *Journal of Applied Ecology*, 43(4), 802-815.

McCune, J. L. (2016). Species distribution models predict rare species occurrences despite significant effects of landscape context. *Journal of Applied Ecology*, 53(6), 1871-1879.

McCune, J. L., Van Natto, A., & MacDougall, A. S. (2017). The efficacy of protected areas and private land for plant conservation in a fragmented landscape. *Landscape Ecology*, 32(4), 871-882.

McKenney, D. W., Pedlar, J. H., Lawrence, K., Papadopol, P., & Campbell, K. (2014). Hardiness Zones and Bioclimatic Modelling of Plant Species Distributions in North America. In Proceedings of the 2014 Annual Meeting of the International Plant Propagators Society 1085 (pp. 139-148).

Mitchell, P. J., Monk, J., & Laurenson, L. (2017). Sensitivity of fine-scale species distribution models to locational uncertainty in occurrence data across multiple sample sizes. *Methods in Ecology and Evolution*, 8(1), 12-21.

Moudrý, V., & Šímová, P. (2012). Influence of positional accuracy, sample size and scale on modelling species distributions: a review. *International Journal of Geographical Information Science*, 26(11), 2083-2095.

Newbold, T., Reader, T., El-Gabbas, A., Berg, W., Shohdi, W. M., *et al*. (2010). Testing the accuracy of species distribution models using species records from a new field survey. *Oikos*, 119(8), 1326-1334.

Pardo, I., Pata, M. P., Gómez, D., & García, M. B. (2013). A novel method to handle the effect of uneven sampling effort in biodiversity databases. *PloS One*, 8(1), e52786.

Pearson, R.G., Raxworthy, C.J., Nakamura, M. & Peterson, A.T. (2007) Predicting species distributions from small numbers of occurrence records: a test case using cryptic geckos in Madagascar. *Journal of Biogeography*, 34, 102–117

Peterson, A. T., Papes, M., & Kluza, D. A. (2003). Predicting the potential invasive distributions of four alien plant species in North America. *Weed Science*, 51(6), 863-868.

Phillips, S. J., Anderson, R. P., Dudík, M., Schapire, R. E., & Blair, M. E. (2017). Opening the black box: an open-source release of Maxent. *Ecography*, 40(7), 887-893.

Phillips, S.J., Anderson, R.P. & Schapire, R.E. (2006) Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, 190, 231–259.

R Core Team (2016) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna

Reside, A. E., Watson, I., VanDerWal, J., & Kutt, A. S. (2011). Incorporating low-resolution historic species location data decreases performance of distribution models. *Ecological Modelling*, 222(18), 3444-3448.

Roubicek, A. J., VanDerWal, J., Beaumont, L. J., Pitman, A. J., Wilson, P., & Hughes, L. (2010). Does the choice of climate baseline matter in ecological niche modelling?. *Ecological Modelling*, 221(19), 2280-2286.

Soultan, A., & Safi, K. (2017). The interplay of various sources of noise on reliability of species distribution models hinges on ecological specialisation. *PloS One*, 12(11), e0187906.

Stockwell, D. R., & Peterson, A. T. (2002). Effects of sample size on accuracy of species distribution models. *Ecological Modelling*, 148(1), 1-13.

Stolar, J., & Nielsen, S. E. (2015). Accounting for spatially biased sampling effort in presence-only species distribution modelling. *Diversity and Distributions*, 21(5), 595-608.

Swets, J.A. (1988) Measuring the accuracy of diagnostic systems. *Science*, 240, 1285–1293.

Syfert, M. M., Smith, M. J., & Coomes, D. A. (2013). The effects of sampling bias and model complexity on the predictive performance of MaxEnt species distribution models. *PloS One*, 8(2), e55158.

Tessarolo, G., Ladle, R., Rangel, T., & Hortal, J. (2017). Temporal degradation of data limits biodiversity research. *Ecology and Evolution*, 7(17), 6863-6870.

Thuiller, W. (2004). Patterns and uncertainties of species' range shifts under climate change. *Global Change Biology*, 10(12), 2020-2027.

Tulowiecki, S. J., Larsen, C. P., & Wang, Y. C. (2015). Effects of positional error on modeling species distributions: a perspective using presettlement land survey records. *Plant Ecology*, 216(1), 67-85.

Urbanek, S. (2017). rJava: Low-Level R to Java Interface. R package version 0.9-9. https://CRAN.R-project.org/package=rJava

van Proosdij, A. S., Sosef, M. S., Wieringa, J. J., & Raes, N. (2016). Minimum required number of specimen records to develop accurate species distribution models. *Ecography*, 39(6), 542-552.

Velásquez-Tibatá, J., Graham, C. H., & Munch, S. B. (2016). Using measurement error models to account for georeferencing error in species distribution models. *Ecography*, 39(3), 305-316.

Wisz, M. S., Hijmans, R. J., Li, J., Peterson, A. T., Graham, C. H., Guisan, A., & NCEAS Predicting Species Distributions Working Group. (2008). Effects of sample size on the performance of species distribution models. *Diversity and Distributions*, 14(5), 763-773.

Wisz, M. S., Pottier, J., Kissling, W. D., Pellissier, L., Lenoir, J., *et al*. (2013). The role of biotic interactions in shaping distributions and realised assemblages of species: implications for species distribution modelling. *Biological Reviews*, 88(1), 15-30.

Wolmarans, R., Robertson, M. P., & van Rensburg, B. J. (2010). Predicting invasive alien plant distributions: how geographical bias in occurrence records influences model performance. *Journal of Biogeography*, 37(9), 1797-1810.