Estimating the number of undiscovered rare plant occurrences in Southern Ontario

By Elise S. Urness

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

ABSTRACT

We often do not know the total number of extant populations of species of conservation concern. Species distribution models (SDMs) can be used to predict the probability of species' occurrence. We tested the use of validated SDMs to estimate the number of occurrences of rare plant species across Southern Ontario. We built SDMs for six rare species using known occurrence records and then surveyed 282 new sites and used presence/absence records from these sites to predict probability of occurrence based on the SDM output. We summed these probabilities to estimate the number of occurrences on the landscape. We then used simulation exercises to estimate the likelihood that our sample size was large enough to make a confident estimate. Simulation results showed that the true number of extant occurrences can be overestimated with fewer than 1,000 SDM-directed survey sites. Therefore, our estimates may be overestimates of the true number of extant occurrences, and more surveys will be required to obtain more accurate estimates. This technique for estimating the number of remaining rare species occurrences will inform researchers and managers as they prioritize time and money towards decisions around species recovery and protection, and where to allocate additional survey effort.

ACKNOWLEDGEMENTS

First and foremost, I would like to thank God for giving me this amazing opportunity as well as providing me with the mental strength and emotional fortitude to reach the end. I also thank my husband, Jonathan Dean Urness, for his unending support and encouragement throughout and beyond this graduate degree.

A much deserved thank you to Dr. Joseph Bennett, the Supervisor I would never have known that I needed. Dr. Bennett conceived the study, planned the analytical framework for occurrence estimates, and edited the code and paper. I appreciate his continued enthusiasm for my thesis research, and ever encouraging meetings when I was lost or stuck on R code issues. There are many qualities that Dr. Bennett exemplified that I hope to take with me into future supervisory roles, namely his patience, enthusiasm and diplomacy, always lifting people up and making them feel capable and inspired.

Thank you to Dr. Jenny McCune and Hanna Rosner-Katz for their work that precedes this thesis project, for building MaxEnt SDMs and performing targeted field surveys. And Jenny for starting the project back in 2014, for writing the R code for the focal species occurrence estimates, and her continued correspondence, coaching and collaboration throughout the analysis and finalization of the findings. Thank you to Taylor Radu (undergraduate research assistant) for her commitment to the entire field season and her assistance in conducting field surveys. I am also grateful for the opportunity to spend the summer (2018) in the field with both Jenny and Hanna learning hundreds of species by name; what an amazing wealth of knowledge and experience they shared with me.

Without Dr. Shaun Coutts this project would not have been possible. Thank you, Dr. Coutts, for your advanced R coding and modelling skills and your willingness to coach a beginner coder such as myself. Dr. Coutts wrote the R code for the simulation of presences, random and informed sampling and occurrences estimates. Thank you to Ben Hilna for additional R code support and troubleshooting, for his patience and willingness to teach.

I acknowledge my committee members in their advisory roles; Risa Sargent (University of Ottawa), Tyler Smith (Agriculture Canada), and Jenny McCune (University of Lethbridge). An additional thank you to Tyler Smith for his crucial support in helping to identify and verify identification of our unknown grass and sedge samples collected during the 2018 field season.

I thank the Ontario Ministry of the Environment, Conservation and Parks (MECP), specifically the Species at Risk Stewardship Program for funding this research and the Hibiscus Millennium Project Bursary for additional financial support. I also acknowledge funding from Jenny's Liber Ero fellowship for funding the first two years of surveys.

Finally, I would be remiss not to thank the students and faculty members of the Geomatics and Landscape Ecology Lab. The comradery and support from my fellow students were paramount to not only my success as a master's student but also my enjoyment of the entire grad student experience. Thank you for your collective encouragement, consolation, continued friendship.

AUTHOR CONTRIBUTIONS

Chapter Two: Estimating the number of undiscovered rare plant occurrences in Southern Ontario

Elise S. Urness, Jenny L. McCune, Shaun R. Coutts and Joseph R. Bennett

Though I am the primary author of this master's thesis, this research is a part of a collaborative effort which began years before I started graduate work under Dr. Joseph Bennett. The project was conceived by McCune, initially intending to test and use SDMs to target field surveys for rare woodland plants. McCune built the initial species distribution models (SDMs) used to direct future surveys, I helped to collect, digitize and interpret data. Bennett suggested using the SDMs and survey results to test whether we could use these to estimate the number of extant occurrences. McCune, Coutts and Bennett wrote the R code for data analysis. Bennett acted as the principal supervisor and advisor on the project. I was responsible for all the writing. All co-authors provided comments and feedback on the manuscript.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Figure 10: Informed sampling of simulated presences (200-6400) of *A. dracontium*, with 50 iterations at each sample size (10 – 1,000,000). The cells with a probability of containing a presence lower than 0.0158 were removed from occurrence estimates. The horizontal dashed red line is the known number of presences assigned to the landscape.

Figure 11: Random (blue) and informed (red) sampling of simulated presences (200 – 6400) of *A. dracontium, w*ith 50 sampling iterations at each sample size (10 – 1,000,000). No filter was used to obtain these occurrence estimates. The horizontal dashed red line is the known number of presences assigned to the landscape.

Figure 12: Panels A – F delineate the simulations by the number of presences assigned the landscape shown next to each letter. The mean number of observed presences for both informed (blue) and random (red) sampling techniques are depicted along the y-axis. No filter was used to obtain these occurrence estimates. The horizontal dashed red line is the known number of presences assigned to the landscape.

Figure 13: Panels A - F show the mean estimated number of occurrences at each number of presences, informed vs random sampling. No filter was used to obtain these estimates of occurrence.

LIST OF APPENDICES

Appendix Figure 19: Informed sampling of simulated presences (200-6400) of *F. quadrangulata*, with 50 iterations at each sample size (10 – 1,000,000). The cells with a probability of containing a presence lower than 0.02 were removed from occurrence estimates. The horizontal dashed red line is the known number of presences assigned to the landscape.

Appendix Figure 20: Random (blue) and informed (red) sampling of simulated presences (200 – 6400) of *F. quadrangulata, w*ith 50 sampling iterations at each sample size (10 – 1,000,000). No filter was used to obtain these occurrence estimates. The horizontal dashed red line is the known number of presences assigned to the landscape.

Appendix Figure 21: Panels A - F delineate the simulations by the number of presences assigned the landscape shown next to each letter. The mean number of observed presences for both informed (blue) and random (red) sampling techniques are depicted along the y-axis. No filter was used to obtain these occurrence estimates. The horizontal dashed red line is the known number of presences assigned to the landscape.

Appendix Figure 22: Panels A - F show the mean estimated number of occurrences at each number of presences, informed vs random sampling. No filter was used to obtain these estimates of occurrence.

Appendix Figure 23: *F. quadrangulata* SDM used as landscape for simulation. Cropped using the same dimensions as for *A. dracontium* to contain 2.6 million cells. The SDM is depicted as a heat map showing suitable sites in red/orange and less suitable sites in blue.

Appendix 24: R code used for simulations

CHAPTER ONE: INTRODUCTION

The threats to species at risk are largely human caused and widespread including habitat loss, degradation, fragmentation, invasive species, disease and insect outbreaks (Bickerton & Thompson-Black 2010; Boland et al., 2012; Faber-Langendoen et al., 2012; Donley et al., 2013). Habitat loss due to land use change is considered the principal threat to species at risk (Kerr & Cihlar, 2004; Kerr & Deguise, 2004; Pimm et al., 2014) and occurs most rapidly and extensively in areas with the highest human population densities (Cardillo et al., 2004; Coristine et al., 2018; Kerr & Currie, 1995). Humans have modified more than 80% of the world's terrestrial surfaces (Sanderson et al., 2002; Tulloch et al., 2016). Even in areas where human population densities are lower, ecosystems are still subject to anthropogenic influence. These increasing pressures threaten the persistence of our national and global biodiversity into the future.

The value of nature as a whole is incalculable, making it difficult to estimate the distinct consequences of its losses (Balmford et al., 2011). Biodiversity is worth more than solely its intrinsic value, and its loss is more than simply the loss of ecosystem services (Balmford et al., 2011). In an essay on the significance of an individual species, Aldo Leopold wrote "To keep every cog and wheel is the first precaution of intelligent tinkering" (Leopold, 1949). Whether you consider their inherent value, or the roles they play in maintaining healthy ecosystems and supporting human populations, species on the brink of extinction are worth conserving.

What is being done?

There are several mechanisms by which species at risk are protected, one being through the protection of their habitat in the form of protected area networks. However, Heywood (2019) claims that, on a global scale, protected areas are not living up to their intended potential in their capacity to conserve biodiversity. It is apparent that current approaches are inadequate, as protected areas are often not very representative of where species at risk are (Havens et al., 2014; Heywood, 2017).

It is becoming increasingly important that we consider conservation on privately owned land as a supplementary approach to our current deficient efforts to protect land. While it is often assumed that protected areas contain more biodiversity than private lands due in part to their strategic placement and management, this is not always the case (Rayner et al., 2014; Richart & Hewitt, 2008). McCune et al. (2017) found that native plant species richness tended to be higher on privately-owned sites in Southern Ontario. They also found that privately-owned sites had a higher likelihood of supporting a species of conservation concern than sites within protected areas when controlling for landscape context, land use history and forest age.

Biological surveys on private land are far less common than those conducted on protected lands (Hilty & Merenlender, 2003; Wilcove et al., 2004). This disparity presents a challenge for comparing efficacy of protected areas and private lands for conservation (McCune et al., 2017). Few surveys being done on private land naturally leads to an incomplete understanding of how many occurrences exist on the landscape, and subsequently to their protection being under-explored. This could indicate that we are not only failing to protect

many of these rare species occurrences, but also, we do not know how many populations of rare species are extant.

Systematic prioritization has become a common strategy to maximize efficient conservation efforts, resources, and spending (Bennett et al., 2015; Di Fonzo et al., 2016; Tulloch et al., 2015). These methods prioritize rare species by their extinction risk and other factors such as the potential cost of management, and concentrate attention on conservation actions designed to protect them (Possingham et al., 2002). A comprehensive understanding of the location and size of populations of rare species is fundamental to effective prioritization for the purpose of conservation planning (Elith et al., 2006).

Rare species pose a challenge in that there are often very little data on their population locations, making it difficult to estimate their true abundance and distribution (McDonald, 2004). Estimating the proportion of suitable sites occupied by a rare species is a necessary step toward effective monitoring programs and often less expensive than abundance estimations and comprehensive species inventories (Mackenzie et al., 2003).

Species Distribution Models

Species distribution models (SDMs) have become popular tools for ecologists to understand the underlying environmental and geospatial variables that make a site relatively suitable or unsuitable for a given species (Guisan et al., 2013). These models are designed to correlate environmental variables with known occurrence records to provide a measure of species' occupancy potential across a region (Elith & Graham, 2009; Hernandez et al., 2006). There are many applications of SDMs in the literature, including use in climate change and

invasive species research, evolutionary biology, and epidemiology (Elith & Graham, 2009). SDMs and ecological niche models have also been used to determine and prioritize areas for species reintroductions (Martinez-Meyer et al., 2006).

SDMs have been used to map distributions and predict occurrences of rare species (Boetsch et al., 2003; Dunk et al., 2004; Engler et al., 2004; Gogol-Prokurat, 2011; Guisan et al., 2006; Williams et al., 2009). However, it does not appear that SDMs have previously been employed to estimate how many occurrences or populations exist of a given species in a region of interest. We define 'occurrence' in this study as a location at which a species is present, regardless of abundance. In many cases 'occurrence' is interchangeable with population, however multiple occurrences within a 1 km distance of each other are sometimes considered the same population, or same 'element occurrence' (NatureServe, 2002).

Study purpose

In this study we used a maximum entropy modelling (MaxEnt) approach to estimate the extent of suitable habitat across our study region (Elith & Graham, 2009) and subsequently calculate occurrence estimates. Modelling species distributions for rare species can be difficult due to limited occurrence records (Williams et al., 2009). We used MaxEnt to build SDMs because it performs well with few species' occurrence records as well as presence-only data (Hernandez et al., 2006; Pearson et al., 2007; Phillips et al., 2006; Van Proosdij et al., 2016). The specific objectives of this research are: 1) to estimate the number of remaining occurrences of selected rare plant species, and 2) to examine the confidence of these estimates using a simulation process. The simulation process allows us to assign presences to a landscape,

subsequently sample them, calculate estimates, and compare estimates to the known number of presences. Estimating the true number of occurrences is important in order to more accurately assess the conservation status of species and prioritize them for conservation. The results of these tests therefore have the potential to contribute to species at risk prioritization provincially, nationally and globally.

A previous study in our study region found that targeted sampling, informed by species distribution models (SDMs), effectively identified previously unrecorded occurrences of several rare woodland plant species in Southern Ontario (McCune, 2016; Rosner-Katz et al., accepted). Building on these methods, we used SDMs followed by testing with independent presence and absence data to predict the probability of selected rare species being present at a given site, and then summed the probabilities to estimate the total number of sites predicted to contain a selected species. The focus of this study is to provide tools for assessing rarity and risk of extinction by estimating the number of rare plant species occurrences that have yet to be discovered. We define 'rare' as those species which are considered vulnerable, imperiled, or critically imperiled at the provincial level in Ontario (S-rank S3, S2, or S1 respectively; Faber-Langendoen et al., 2012).

Because we were able to detect previously unrecorded occurrences of rare woodland plants using a targeted sampling method at sites identified as suitable, we concluded that there are more occurrences of several species than formerly thought. This does not mean that these species are not threatened. It tells us that knowing the landscape composition and environmental variables associated with sites where the species are known to occur can help us

predict potential population trends as the landscape is modified by human use and climate change.

CHAPTER TWO: ESTIMATING THE NUMBER OF UNDISCOVERED RARE PLANT OCCURRENCES IN SOUTHERN ONTARIO

1 | INTRODUCTION

In Canada there are over 500 species listed under the federal *Species at Risk Act (SARA)* (Government of Canada, 2019). Most of these species occur in Canada's southern regions where land use change is highest (Kerr & Cihlar, 2004). The protection and recovery of species at risk in such areas are critical to slow down the current rapid rate of species extinction worldwide (Myers et al., 2000).

Determining conservation status for species at risk requires an assessment of the number of individuals, number of populations, their geographic locations, changes in the population size over time and threats to the species' persistence (COSEWIC, 2017). A species of conservation concern requires some form of status assessment and formal listing to be acknowledged as 'at risk'. This poses a challenge because rare species often lack complete data on population sizes and locations, and on trends in populations over time (McCune, 2016; Schemske et al., 1994). In response to small and/or declining populations, many of these recovery strategies advise some form of inventory, along with surveying and monitoring of known and newly discovered occurrences (Bickerton & Thompson-Black, 2010; Boland et al., 2012; Donley et al., 2013; Faber-Langendoen et al., 2012). These surveys are often expensive, and may be biased towards publicly owned lands, which may be easier for surveyors to access (Hilty & Merenlender, 2003). In some cases, increased survey effort as a result of species being

listed by COSEWIC has led to the discovery of several previously unknown populations (Favaro et al., 2014.) This shows that many undiscovered populations of species at risk remain unprotected, because they remain unreported.

SDMs, also known as ecological niche models, correlate geospatial variables with known occurrence records to provide a measure of habitat suitability across a region (Elith & Graham, 2009; Hernandez et al., 2006). SDMs allow researchers to model the probability of species presence using the relationship between presence/absence species data and associated environmental variables to predict a species distribution at unsurveyed sites over a continuous area. While species distribution maps are increasingly being used to set conservation priorities based on modeled distributions (Guisan et al., 2013; Schuster et al., 2019), their utility for predicting numbers of occurrences is underexplored to date.

In this study we use species distribution models (SDMs) combined with independent presence and absence data to predict how many occurrences of rare plants remain undiscovered in Southern Ontario. Analyses using both presence and absence data are becoming increasingly useful to inform biodiversity conservation planning, even for rare species with few data (Bayley & Peterson, 2001). The scarcity of data regarding rare species results in difficulty applying the usual statistical approaches to modeling distributions (Engler et al., 2004). However, although the distributions of rare plants may be difficult to model, plants offer the advantages of sessility and a relatively limited dispersal ability compared to vagile organisms. Their presence at a given site offers insights into the species tolerance to both current and past site conditions and their ability to persist at the site (Dunk et al., 2004). There have been examples of past successes, whereby researchers have used SDMs to predict rare plant species

habitat and subsequently identified new occurrences (e.g. Fois et al., 2018; McCune, 2016; Rhoden et al., 2017).

The lack of data and often poor quality of available data on rare species leads to uncertainty in the species prioritization process (Langford et al., 2011). This gap in the knowledge base provides clear rationale for this project, aimed at developing methods for predicting species occurrence on the landscape while accounting for sampling biases and uncertainties (Arponen, 2012). The objectives of this research are: 1) to estimate the number of remaining occurrences of six rare plant species in the study area, and 2) to examine the confidence of these estimates using a simulation process. Estimating the true number of occurrences is important in order to more accurately assess the conservation status of species and prioritize for which species more field surveys are necessary to locate undiscovered populations. The results of these tests therefore have the potential to contribute to species at risk prioritization provincially, nationally and globally.

2 | METHODS

2.1 | Southern Ontario Case Study

We chose Southern Ontario for this study (Figure 1) for three main reasons: 1) Southern Ontario is one of Canada's biodiversity hotspots (Coristine et al., 2018; Kerr & Cihlar, 2004). 2) Southern Ontario is among the least protected ecoregions in the country (Coristine et al., 2018). 3) Southern Ontario is highly developed (agriculture, industry, urban, suburban, and rural) with limited remaining forested land, most of which tends to be relatively small in area and privately

owned. As the remaining forest patches tend to be small and isolated, this region provides a prime example of a context where protection of small areas is paramount (Goefroid & Koedam, 2003).

Southern Ontario has a very high number of species at risk coinciding with heavily developed and densely human populated areas. This region is of both national and provincial ecological importance as over 40% of Canada's plant species are known to occur in this small southern region (Oldham, 2017). It is also home to 35% of Canada's human population, though it represents a mere 1% of Canada's terrestrial area.

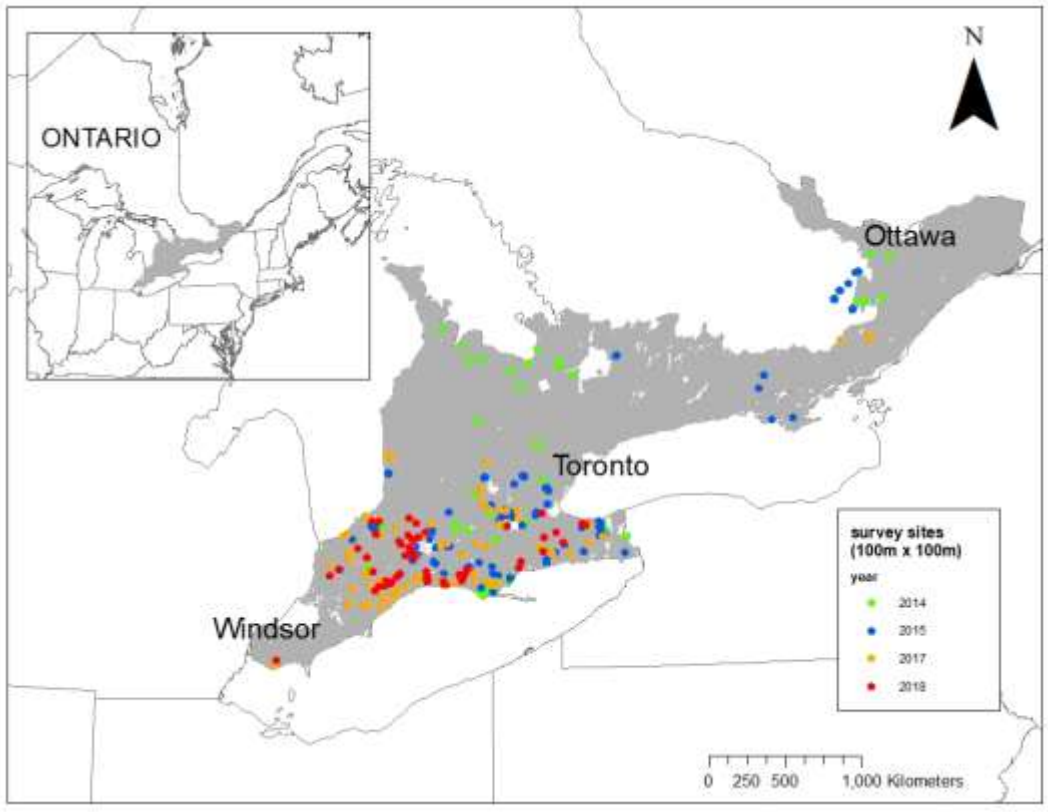


Figure 1: Study area map of Southern Ontario showing survey sites over four field seasons. The study extent is shown in grey.

2.2 | Focal species

McCune (2016) and Rosner-Katz et al. (accepted) built SDMs for 41 species to test the efficiency of SDMs to target field surveys for rare plant species. We chose six of these species, hereby referred to as our 'focal' species, based on the high number of new occurrences detected over four field survey seasons (Table 1).

Table 1: Focal species conservation status and our current understanding of how many populations exist in Southern Ontario. *Natural History Information Centre (NHIC)

SPECIES (SCIENTIFIC NAME)	COMMON NAME	NATIONAL STATUS (COSEWIC)	ONTARIO STATUS (SRANK)	Field survey records (2014 – 2018)	records of occurrences (field surveys + *NHIC presences)
Arisaema dracontium	green dragon	Special Concern	S3 - vulnerable	19	98
Lithospermum latifolium	American gromwell	N/A	S3 - vulnerable	9	22
Symphyotrichum prenanthoides	Crooked-stem aster	Special Concern	S2 - imperiled	11	30
Fraxinus quadrangulata	blue ash	Threatened	S2 - imperiled	5	51
Cornus florida	Eastern flowering dogwood	Endangered	S2 - imperiled	14	387
Castanea dentata	American chestnut	Endangered	S2 - imperiled	8	160

Arisaema drac0ontium (green dragon) is a perennial herb that grows in moist deciduous woods along rivers, creeks and seasonal floodplains (Boles et al., 2000). *A. dracontium* produces a single compound leaf at the end of a short stalk 15 – 90 cm tall (Donley et al., 2013). Because its climatic tolerances limit its range to the heavily populated region of Southern Ontario, *A. dracontium* is confined to increasingly small and isolated habitat fragments (Rothfels & Smith, 2003). Several of these isolated populations have very few individuals, which exacerbates their risk of extinction (Donley et al., 2013).

Lithospermum latifolium (American gromwell) is a tall, hairy perennial herb that grows along shaded riverbanks and forested floodplains in deciduous woodlands (Reznicek et al. 2011). As for many species at risk in Ontario*, L. latifolium's* range extends south into the United States. It is uncommon in its range and listed as endangered in the states of Maryland and Pennsylvania (USDA, 2019). There are few known occurrences of this species and it appears to be declining range wide for unclear reasons, one of which could be that its habitat is often prime for development (NatureServe, 2019).

Symphyotrichum prenanthoides (Crooked-stem aster) is a perennial herb with zig-zagging stems. *S. prenanthoides* occurs in a wide range of habitats including floodplains, along creeks, edges of woods in riparian forests, and even along some roadsides as it is tolerant to moderate disturbance (ECCC, 2018). It is distinguishable from other asters by its clasping leaf-bases attached to the stem. The major threats to *S. prenanthoides* include habitat loss from invasive/non-native alien species which occur at most of its sites, use of herbicide on roadsides, residential development, livestock grazing, and logging (ECCC, 2018).

Fraxinus quadrangulata (blue ash) is a medium-sized tree growing up to 20 metres with a straight and slender trunk between 15 and 25 centimetres in diameter (COSEWIC, 2014). It has opposite and compound leaves with up to 11 leaflets. *F. quandrangulata* grows most commonly in deciduous forests along floodplains, sandy beaches and even on limestone outcrops associated with Lake Erie (COSEWIC, 2014). *F. quadrangulata* populations extend south as far as Georgia, United States. In Canada its range is restricted to within Ontario's southwestern region in the Carolinian forest zone (COSEWIC, 2014). *F. quadrangulata* is threatened by loss and fragmentation of its habitat, along with browsing by white-tailed deer and the potential infestation of the Emerald Ash Borer (COSEWIC, 2014).

Cornus florida (Eastern flowering dogwood) is a small deciduous tree native to eastern North America (Bickerton & Thompson-Black, 2010). It occurs most commonly in habitats ranging from open dry-mesic oak-hickory woodlands to moderately moist maple-beech deciduous or mixed forests (Bickerton & Thompson-Black, 2010). The primary threat to this species is the dogwood anthracnose fungus infection, causing population declines at an estimated rate of 7 percent annually (Bickerton & Thompson-Black, 2010).

Castanea dentata (American chestnut) is a large deciduous tree with smooth dark bark that has broad, flat-topped ridges and can grow up to 30 m tall (Boland et al., 2012). *C. dentata* was once a dominant tree species in northeastern North America, comprising 25 percent of the eastern deciduous forest in the United States before the introduction of the devastating fungal pathogen chestnut blight in 1904 (Boland et al., 2012). Chestnut blight was the main cause of the decline of *C. dentat*a, which now persists as remnant populations of individuals throughout its range (Boland et al., 2012).

2.3 | Initial records and models

We used MaxEnt (Phillips et al., 2006) to build SDMs because it performs well with few species' occurrence records as well as presence-only data (Hernandez et al., 2006; Pearson et al., 2007; Van Proosdij et al., 2016). MaxEnt is a machine learning method that is centered on the principle of *maximum entropy*. For the purposes of species distribution modeling, this principle ensures that the range of environmental conditions over which habitat is predicted to be suitable for a species is constrained only as much as is warranted by the available data (Phillips et al., 2006; Elith et al., 2011; Merow et al., 2013). The MaxEnt program is used to estimate the range of tolerable environmental conditions for a species based on the known occurrences without unjustifiably narrowing the species range by extrapolating beyond the known occurrences (McCune, 2019).

We obtained presence records from Ontario's Natural History Information Center (NHIC) and excluded any records with spatial uncertainty greater than 100m (McCune, 2016). These records were used to build SDMs based on climatic, topographic, soil, forest contiguity, land cover, and surficial geology variables (Appendix table 2), at a 100 x 100m resolution. The extent of the models included all Southern Ontario (Figure 1). MaxEnt ranked each 100m x 100m grid cell with a score from 0 to 100 and refers to these values as the cumulative outputs (Merow et al., 2014). Each cell was ranked based on the percentage of cells in the study area that have a cumulative output value equal to or less than that cell's value (Merow et al., 2013). For the purposes of this study I will refer to MaxEnt's cumulative output values as the habitat suitability score.

2.4 | Field surveys

A total of 282 cells were surveyed over 4 field seasons: 51 cells in 2014, 105 cells in 2015, 70 cells in 2017, and 56 cells in 2018. The survey sites were selected non-randomly based on predicted habitat suitability of one or several plant species of conservation concern, while trying to survey at least 10 'suitable' cells for each focal species (Rosner-Katz et al., accepted). SDMs were built for a total of 41 plant species (Rosner-Katz et al., accepted), though in this study we examined the six species for which we observed the most new presences. We excluded absence records when surveys were conducted outside of the timeframe in which each species is known to be the most visible (flowering, fruiting, or vegetative) to minimize the potential for false absences. These surveyed grid cells were within remnant forest patches, the majority of which (n= 221) were privately owned woodlots; the others were on lands owned by the Province of Ontario, Conservation Authorities, or private Land Trusts. During the years 2014, 2015, and 2017 landowners were contacted in person to ask permission to conduct a survey on their property. Prior to the 2018 field season, targeted letters were sent out via mail to private landowners whose land contained suitable forest habitat.

At each site the lead surveyor (J.L. McCune, H. Rosner-Katz or K. Tisshaw) navigated to the centre of the focal cell using a handheld GPS unit. We used a compass and laser rangefinder to mark out 50m lines in each cardinal direction by temporarily placing flagging tape (Figure 2). We systematically searched each quadrant of the cell, and all vascular plant species present were recorded. Each survey lasted approximately 3-10 person hours.

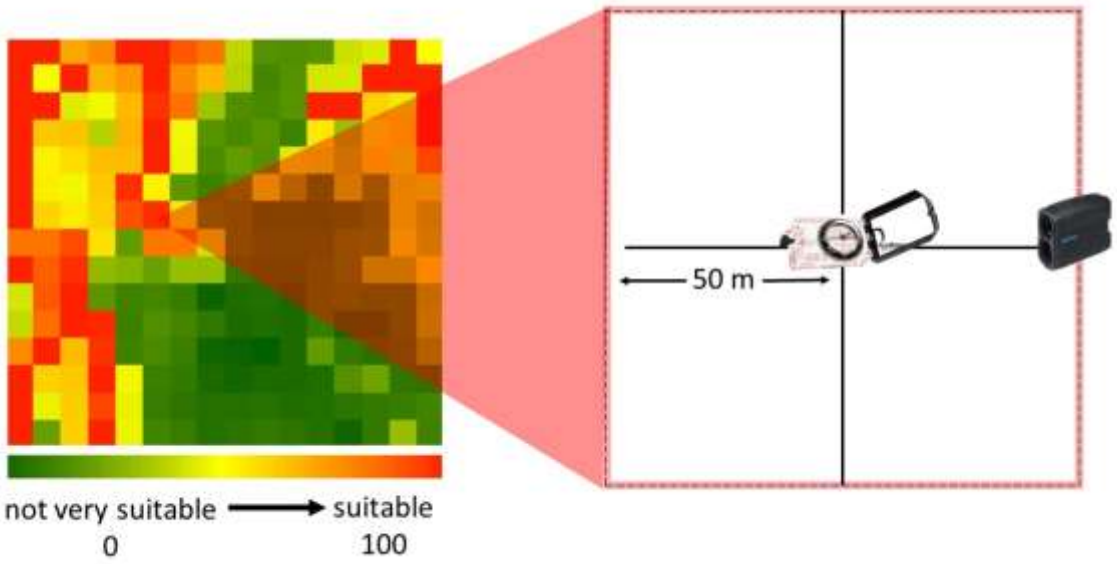


Figure 2: Field survey technique. The focal cell on the right depicts a cell targeted for a rare plant survey using the methods described above. A compass (centre of targeted cell and rangefinder (right middle of targeted cell) were used to delineate survey quadrants. The grid on the left shows the SDM habitat suitability scores of 100m x 100m cells.

2.5 | Choosing the best SDM

Eight SDM versions were created for each of the six focal species (Appendix Table 1). All model versions were built with 14 environmental predictor variables including topographic, climatic, soil, and geology predictors (elevation, slope, aspect, soil texture, soil drainage, surficial geology, isothermality, mean temperature of wettest quarter, annual precipitation, precipitation seasonality, precipitation of the warmest quarter, total precipitation for growing season, annual mean temperature, and mean temperature of growing season). In addition, some SDM versions included forest contiguity (the number of forested cells in the 81-cell neighbourhood of each cell) and/or land cover (Southern Ontario Land Resource Information System data including 25 categories) or both. We added these two variables because a pilot

study (McCune, 2016) showed that some species responded to landscape context or land cover in addition to climatic, edaphic and topographic variables and because the 'best' model version was different for different species (Rosner-Katz et al., accepted). All variables were resampled to a resolution of 100m x 100m. We tested for multicollinearity between environmental predictors (cf. Merow et al., 2014). When two variables were correlated at $r < 0.7$ we used the more generalized variable (e.g. there was a correlation between mean annual temperature and mean temperature of the warmest quarter, therefore the first was retained).

To choose the "best" SDM for each species, we evaluated each model version for the highest AUC (area under the curve) and TPR (true positive rate). We used independent presence/absence records (from our four years of field surveys, and additional independent presences from the NHIC) to test each model's performance. AUC is a threshold independent measure which has a value from 0 to 1. The closer the AUC is to 1 the better the model is at discriminating between presences and absences. An AUC of 0.5 indicates a model that does no better than random at discriminating between the two (Fielding & Bell, 1997). TPR, as a sensitivity measure, is dependent upon a chosen threshold for classifying a cell as 'suitable'. We chose thresholds for each SDM to ensure that 90% of the occurrences used to build the SDM would be predicted to be suitable, a 10% total omission rate. Evaluating which model has the highest TPR will indicate the percent of actual presences that are correctly classified as presences by the model. We evaluated the models in R (using the evaluate() function in the 'dismo' package), testing the models using all the presence and absence records from all 4 survey years. We first chose the model with the highest AUC. If there were two model versions

that tied for highest AUC then we chose the model with the highest TPR. For all analyses we used R version 3.6.2.

2.6 | Estimating occurrences

The MaxEnt predicted habitat suitability score cannot be directly interpreted as the probability of occurrence (Gogol-Prokurat, 2011). Therefore, we used a generalized linear model (GLM) with binomial link function to convert habitat suitability, as predicted by the best performing SDM, to probability of presence (cf. Rosner-Katz et al., accepted). We used presence/absence records from field surveys as the binary response variable, and the MaxEnt cumulative output measuring relative habitat suitability (0-100) as the continuous predictor. We first confirmed that the GLM performed better than an intercept-only model of presence/absence. Then, we predicted the probability of occurrence along with upper and lower confidence limits for each of the 7.2 million cells across the region based on the fitted GLM.

To estimate the total number of remaining occurrences across all 7.2 million 100m x 100m cells in the study region, we used the linearity of expectation property of summed random variables whereby independent probabilities can be summed to obtain an aggregate expected value (Ross, 1994). Specifically, we summed the predicted probabilities across all cells from the GLMs predicted probability of presence outlined above. We also estimated potential ranges of remaining occurrences using the upper and lower 95% confidence limits based on upper and lower confidence limits of occurrence probabilities for cells.

2.7 | Testing for the effect of filters

After summing the probabilities across the study region, we tested the effect of two filters designed to remove likely non-suitable cells by setting the model predictions to zero where unsuitable land occurred (Engler et al., 2004). The first filter removed cells with no forest. We used a binary forest raster, derived from the SOLRIS (Southern Ontario Land Resource Information System) 'wooded' layer accessed through Scholars GeoPortal. This is a generous depiction of forest areas as it over-estimates forest cover by assigning a '1' to all cells that have some forest in them. This filter decreases the number of cells potentially harbouring an occurrence from 7.2 million cells to 3.6 million cells. All cells that had 0 forest cover were removed and cells with at least some forest were retained. Cells with even small amounts of forest were retained because plants, even rare species, can persist in small patches (Bennett & Arcese, 2013).

The second filter eliminated cells with the 'lowest' probability (Table 2). We determined the cell with the lowest probability of presence that actually contained a presence and assumed that to be the lowest possible probability that may harbour an occurrence. For each respective species we removed all cells with a probability of presence lower than these values (see Figure 4 for methods overview). Note that the lowest probability cell where a species is present is different for each species (Table 2).

Table 2: Filter 2 lowest probabilities

SPECIES NAME	Filter 2: lowest probability
Arisaema dracontium	0.0158
Lithospermum latifolium	0.0569
Symphyotrichum prenanthoides	0.0186
Fraxinus quadrangulata	0.0200
Cornus florida	0.0154
Castanea dentata	0.0472

SIMULATION

2.8| Using SDMs to simulate presences

Theoretically (if the SDMs and GLMs are correctly specified), using the linearity of expectation assumption we should be able to extrapolate the number of occurrences. However, there is no way of testing this assumption using 'true' species occurrences, since the number of actual occurrences for all focal species is unknown. Thus, we used a simulation to test the accuracy of our approach. We took the 'best' SDM for the species with the most independent presences collected during our field surveys, *A. dracontium*, and then simulated several estimates of occurrence by varying the sample size and the number of simulated presences assigned to the landscape (see '2.9 sampling regimes' for details). For the sake of computational efficiency, the SDMs (landscapes used for simulations) were cropped from 7.2 million cells to the size of 2.6 million cells in the area of Southern Ontario where the species is most likely to occur (Figure 3). This cropped SDM raster will hereafter be referred to as the 'landscape'. The blank sections of the SDMs shown in figure 3 and Appendix figure 23 for *A.*

dracontium and *F. quadrangulata* are a result of missing data from the environmental predictor variables used to build the models (i.e. soil survey data from the Ontario ministry of agriculture is missing within city boundaries).

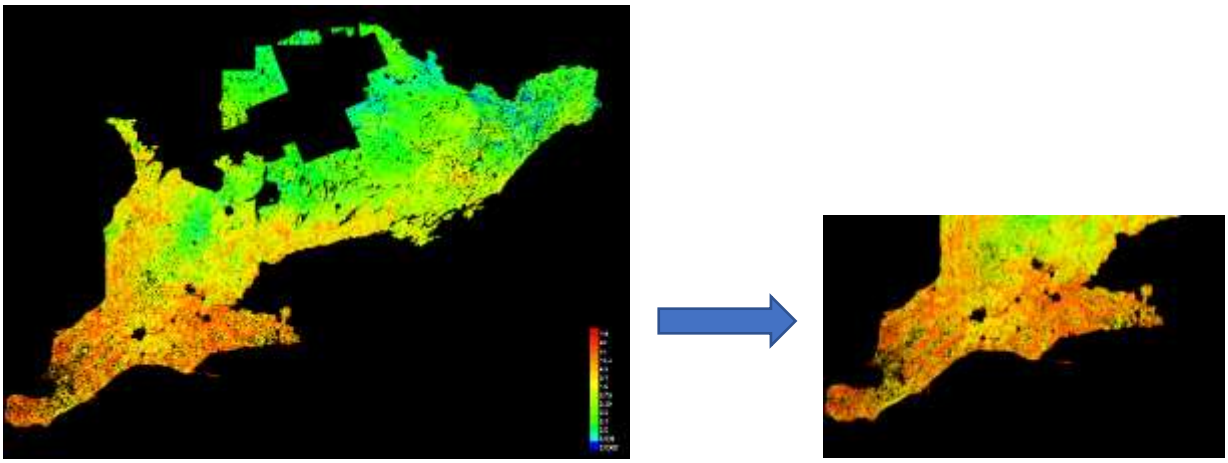


Figure 3: *A. dracontium* SDM used as landscape for simulation. Cropped landscape (right) shows 2.6 million cells. SDMs depicted as heat maps showing suitable sites in red/orange and less suitable sites in green/blue. The blank sections of the SDM raster are a result of missing environmental predictor variable data used to build the models (i.e. soil survey data is missing from cities).

We also ran our simulation process for *F. quadrangulata* to compare estimation trends for a species that has less suitable habitat (Appendix Figures 14 - 23). We chose *F. quadrangulata* as a species for comparison because there are far fewer suitable sites for *F. quadrangulata* within the range of our SDMs (see Appendix Figure 23). We assume that any general patterns of over- or under-estimation due to rarity of the species being sampled (and resulting uncertainty in models) would be similar for other focal species. In the same way that the predicted probabilities were calculated for each focal species, we predicted the probability of species occurrence across each cell in the landscape creating a probability raster map.

Simulated presences were assigned to the landscape based on this probability map, and then presences were removed using a Bernoulli draw for each cell based on assigned occurrence probabilities, until an assigned number of presences on the landscape was retained. To ensure that presences were not assigned to cells which would later be filtered out, we set the probability of occurrence in each filtered cell to 0 (using criteria described in 2.7 above), and then assigned presences using the probability of occurrence as a weight. We also ran a test to confirm that no presences had been assigned to cells with no forest by filtering our presence records using the forest filter and checking that none of the presences fell in cells with a '0' value indicating no forest. We tested the following range of 'true' presences assigned to the landscape: 200, 400, 800, 1600, 3200, and 6400.

We simulated field surveys of the resulting simulated landscapes (one for each of the six levels of simulated presences) by sampling 10 to 1,000,000 cells, increasing from 10 by an order of magnitude each time. We used two sampling strategies: random and informed (see section 2.9, below). We recorded the number of simulated presences and absences captured in each sample and then used them to build a GLM, with presence/absence as the response variable and the SDM habitat suitability score as the predictor. If the sample detected fewer than three simulated presences, we did not construct a GLM because we assumed that this was too few to produce a useful model. Using the GLMs, we summed predicted probabilities to estimate the total number of occurrences. We note that three occurrences is a small number for creating a GLM (even with a single predictor variable); however, we used this number because field surveys for rare species often detect very few new populations, and researchers may wish to create models even with such limited data.

2.9 | Sampling Regimes

We used two sampling methods: random and informed. First, at each sample size, we sampled the landscape randomly without replacement. Random sampling is often encouraged for statistical analysis. However, surveyors for rare species are unlikely to sample randomly, and are instead more likely to search predominantly suitable habitats to be more efficient with limited resources. Indeed, the data we used to calculate the focal species estimates came from previous targeted field surveys (Rosner-Katz et al., accepted). These surveys could not be directly mimicked via simulation. Some cells were surveyed based on the highest suitability for a single species, while some cells were targeted due to relatively high suitability for multiple species (Rosner-Katz et al., accepted). Thus, we also performed informed sampling using the probability from the GLM as a weight to choose samples in a binomial draw (without replacement) from candidate cells. Cells were removed from the available pool of cells once they had been sampled, thus reassigning weights and reshuffling the remaining cells.

2.10 | Testing for the effect of filters

Following the methods used to obtain the focal species estimates based on our real survey data, we applied the same two filters when calculating the total occurrence estimates for *Arisaema dracontium* and *Fraxinus quadrangulata* based on our simulated sampling data (Figure 4). Since the filters were applied after the SDMs were built and the probabilities were assigned to each cell, the same order of operations was used when simulating presences, sampling and estimating. The first filter (described in section 2.7) removed the cells with no

forest. The second filter removed cells with a probability of presence lower than the lowest probability cell that contained a presence (Table 2).

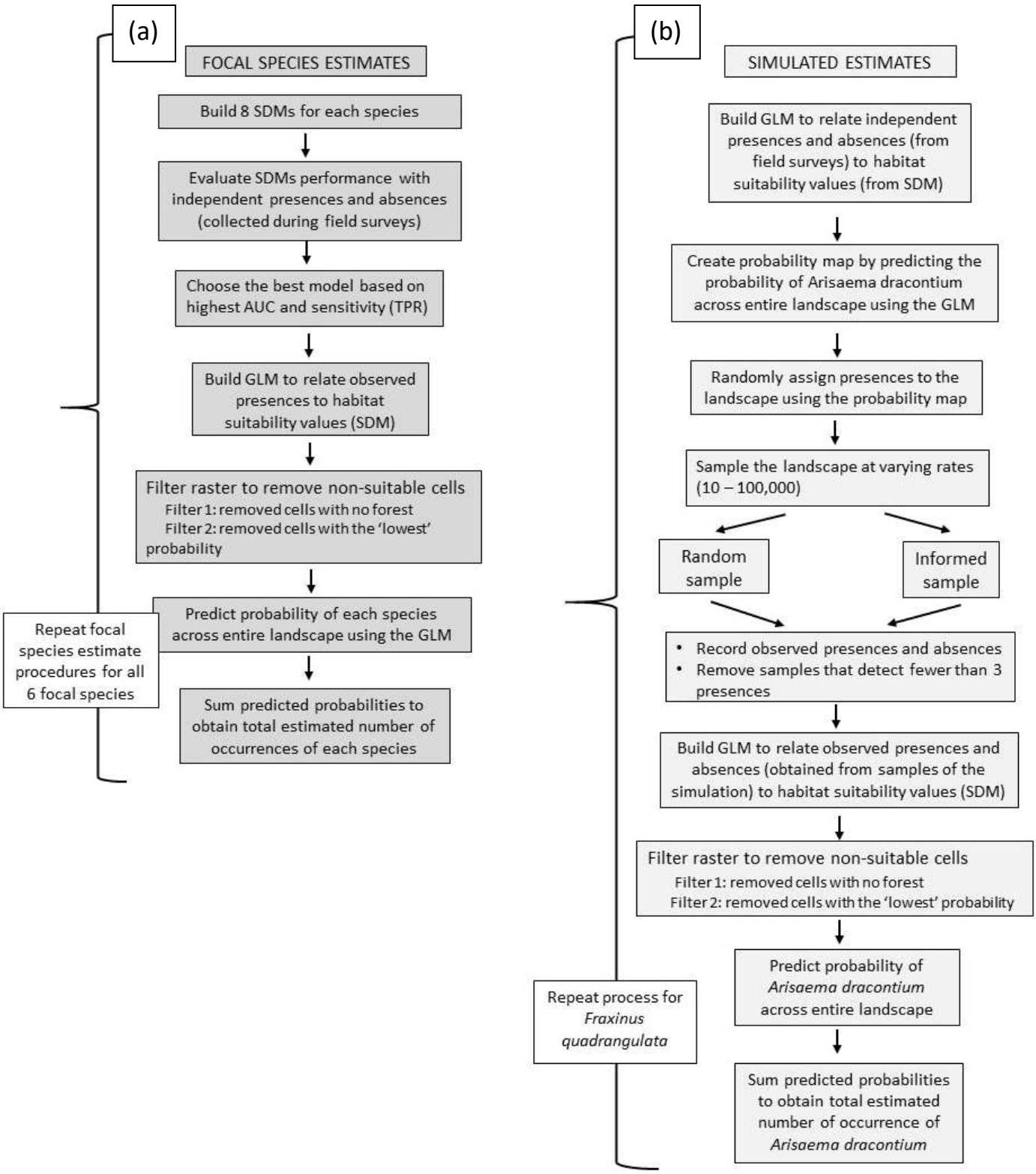


Figure 4: Flowchart of focal species estimation and simulation methods. (a) Methods of focal species estimation process. (b) Methods of presence/absence simulation and simulated sampling.

3 | RESULTS

3.1 | Focal species population estimates

Estimates of the number of occurrences varied among species and appeared to be related to the number of surveyed presences and absences. More survey presences typically meant higher occurrence estimates. The “best” model chosen for each species was the SDM with the highest AUC which, in the case of all six focal species, was also the model with the highest TPR, or at least tied with another model for highest TPR (Appendix Table 1). All GLMs used to relate presence/absence records to the habitat suitability score performed better than a null model using the criterion of ΔAIC <2 (Burnham & Anderson, 2002). The predictor variables included in each of the ‘best’ SDM versions for each focal species are shown in Appendix Table 1.

Filters

We calculated the final occurrence estimates using both the forest filter and the lowest probability filter (Table 3). In every case these estimates of the number of occurrences of each focal species in Southern Ontario is considerably greater than the best current record of the number of occurrences (Table 3).

Table 3: Focal species occurrence estimates with no filter, forest filter (removed all non-forested cells), and lowest filter (removed cells with a probability lower than the lowest probability cells to harbour a presence). Lower and higher estimates calculated using 95% confidence intervals for individual cells. SDM performance for each species is shown as area under the curve (AUC) and true positive rate (TPR) values. Field survey presences and absences are shown, along with current records of the number of occurrences in Southern Ontario for each species. The number of absences varies, and is less than 282 minus # of presences, due to the removal of absence records recorded outside of the timeframe in which each species is known to be most visible (flowering, fruiting, or vegetative) to minimize the potential for false absences. *Natural Heritage Information Centre (NHIC)

SPECIES NAME	**SDM AUC**	**SDM TPR**	**# survey pres**	**# survey abs**	**SUM OF PREDICTED PROBABILITIES: No filter (mean)**	**Forest filter (mean)**	**Lowest filter mean [lower, higher]**	**# of known occurrences (field surveys + *NHIC presences)**
Arisaema dracontium	0.87	0.95	19	250	62166	33931	5840 [2056, 9624]	98
Lithospermum latifolium	0.92	1	8	243	26129	16496	1859 [517, 3767]	22
Symphyotrichum prenanthoides	0.91	0.5	10	226	2157	5572	5206 [1311, 8901]	30
Fraxinus quadrangulata	0.90	1	5	275	70425	269	251 [18, 484]	51
Cornus florida	0.84	0.85	13	245	160226	12507	12344 [3457, 21231]	387
Castanea dentata	0.87	0.86	7	250	65609	3957	2707 [406, 5008]	160

3.2 | *Arisaema dracontium* simulation results

3.2.1 | Random sampling

At low sample sizes there was consistent over-estimation of the number of occurrences in the simulated data. When we assigned the 'true' number of presences to the landscape as 200, on such a large landscape (2.6 million cells), it was very difficult to detect any presences especially at low sample sizes. The mean estimates of occurrence, when extrapolating from GLMs, approached the 'true' number of presences when there were 1,600 or more presences on the landscape and the simulation took at least 10,000 random samples (Figure 5).

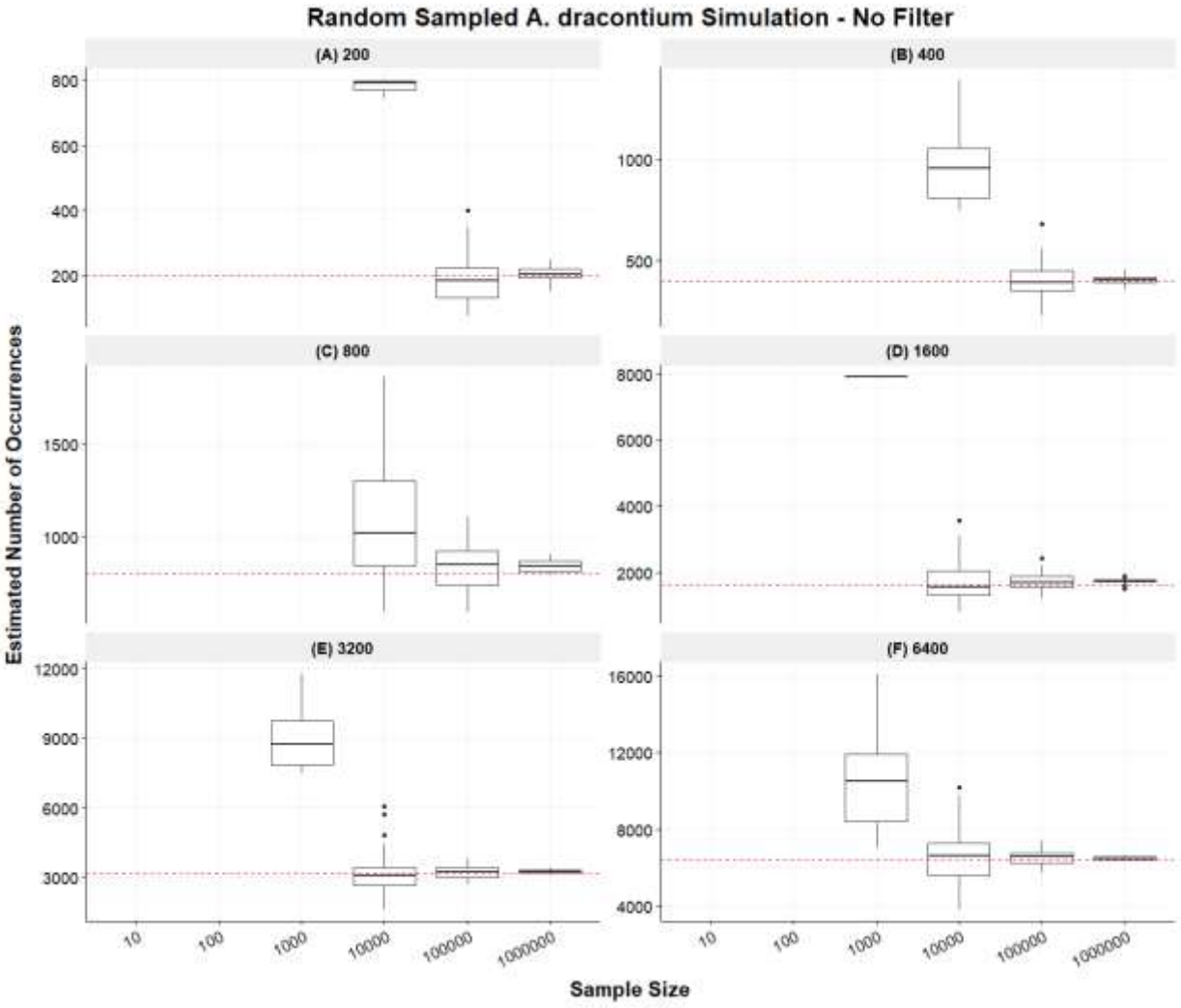


Figure 5: Random sampling of simulated presences (200-6400) of *A. dracontium, w*ith 50 sampling iterations at each sample size (10 – 1,000,000). No filter was used to obtain these occurrence estimates. The horizontal dashed red line is the known number of presences assigned to the landscape.

Forest filter

After filtering out cells with no forest, the mean estimates of total predicted presences leveled out below the 'true' number of presences at large sample sizes, regardless of the 'true' number of presences (Figure 6). At the largest sample sizes, the predicted presence values are approximately 25-55% lower than the actual number of presences.

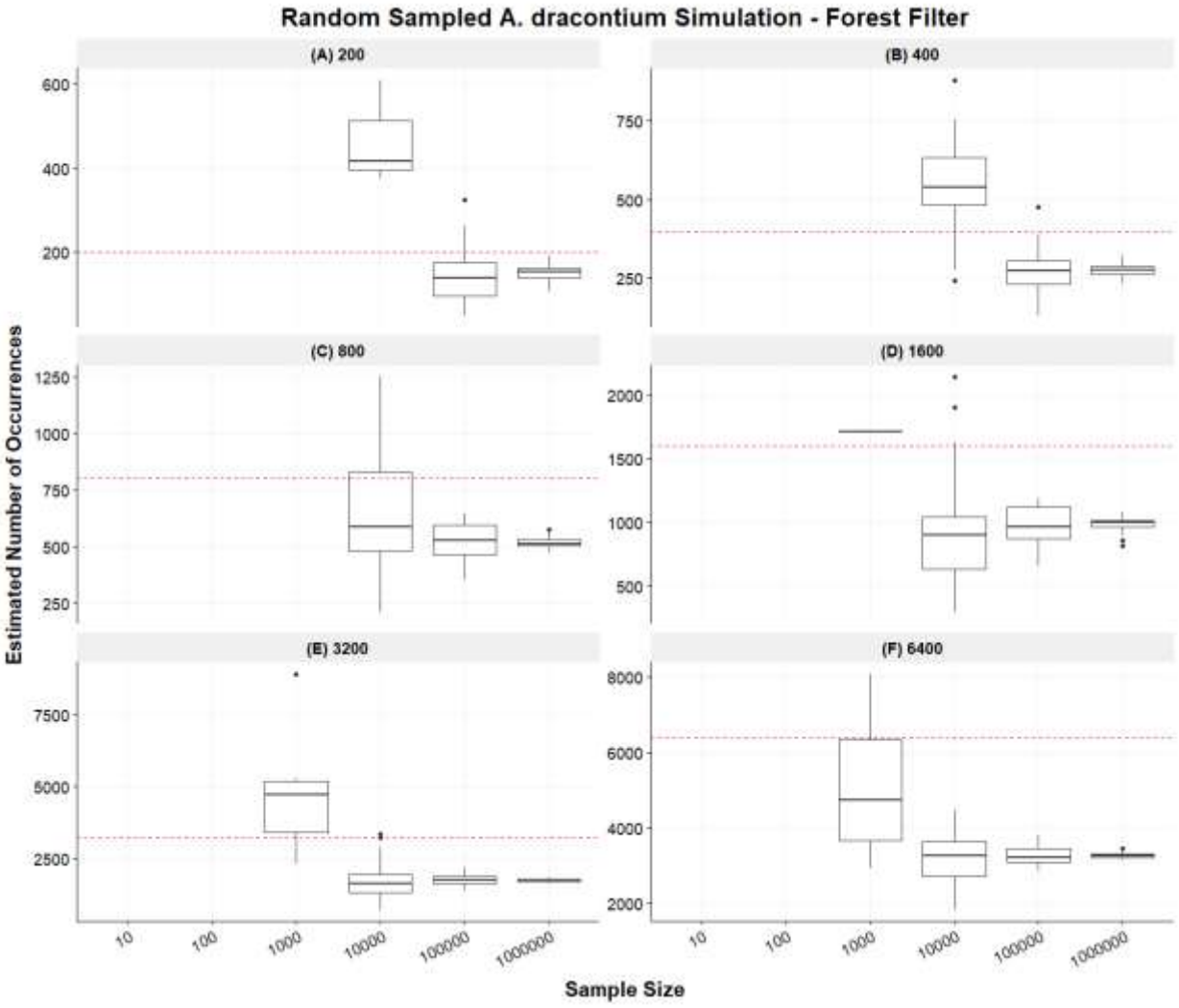


Figure 6: Random sampling of simulated presences (200-6400) of *A. dracontium*. Non-forested cells were removed prior to estimating occurrence numbers. The horizontal red line is the known number of presences assigned to the landscape.

Lowest probability filter

After filtering out cells with a probability of presence lower than the lowest probability cell that contained a presence, the effect of the filter appears even stronger (Figure 7). The predicted presence values are approximately 37-72% lower than the actual number of presences (Figure 7).

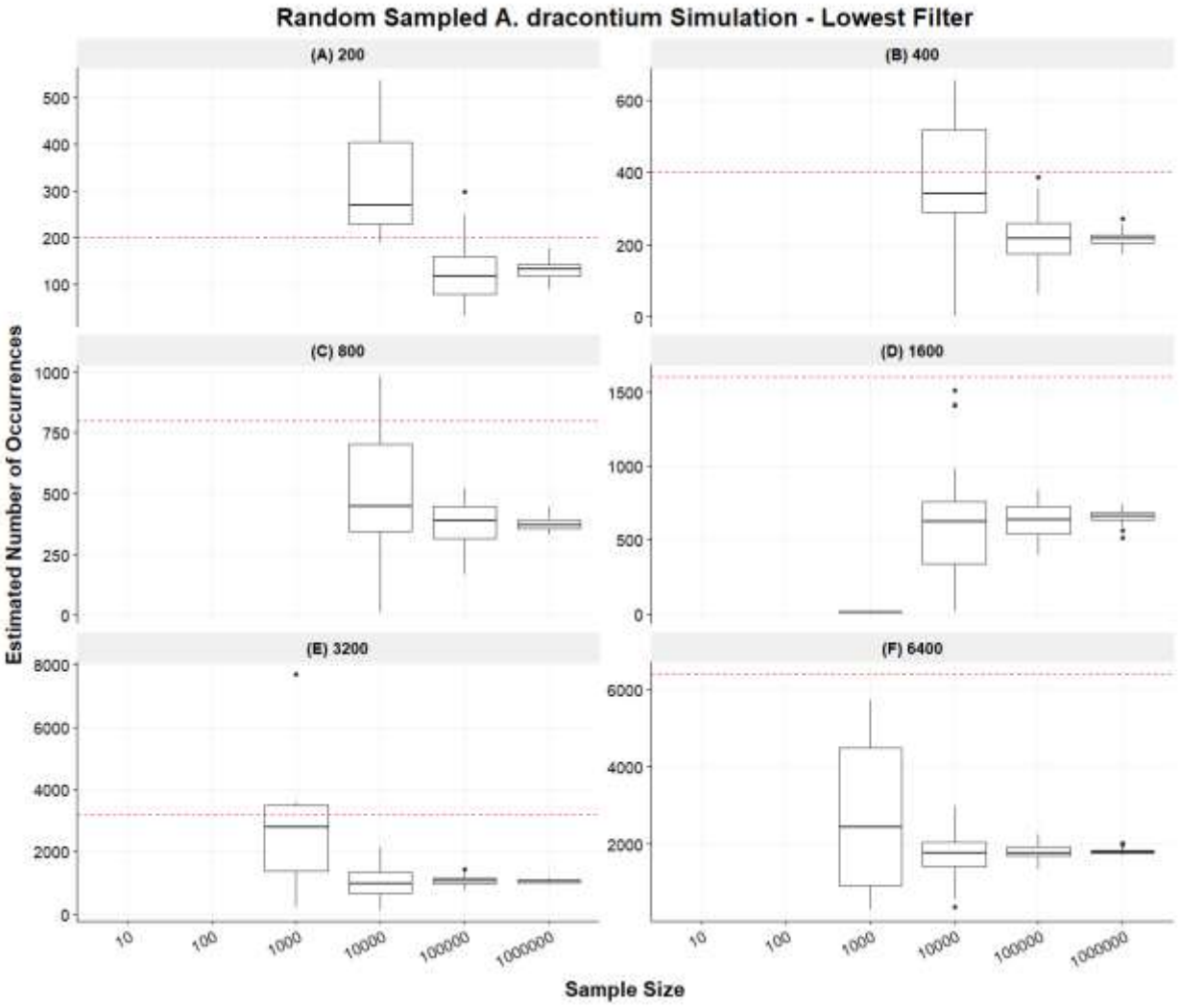


Figure 7: Random sampling of simulated presences (200-6400) of *A. dracontium* SDM, with 50 iterations at each sample size (10 – 1,000,000). The cells with a probability of containing a presence lower than 0.0158 were removed from occurrence estimates. The horizontal dashed red line is the known number of presences assigned to the landscape.

3.2.2 | Informed sampling

Sampling the landscape using an informed sampling method resulted in the same general trend of overestimation at low sample sizes as with random sampling. The difference is that the mean estimates leveled out at the 'true' number of occurrences at sample sizes of approximately 1,000, an order of magnitude lower than the sample size of 10,000 required

before estimates level out at the 'true' number of occurrences seen in the randomly sampled simulations. The informed sampling detects more presences, as it is designed to do so, allowing us to obtain accurate predictions at lower sample sizes (Figure 8).

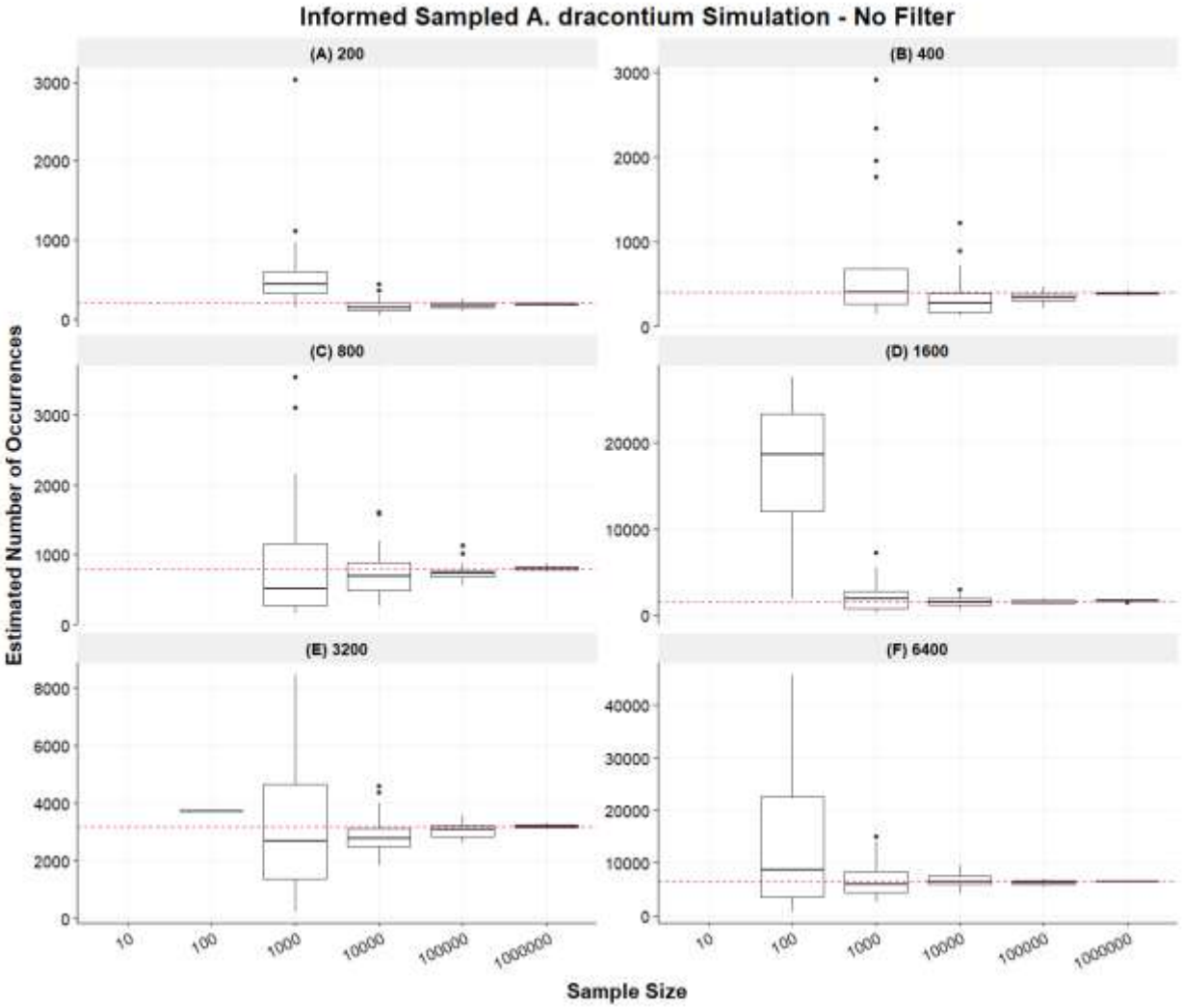


Figure 8: Informed sampling of simulated presences (200-6400) of *A. dracontium*, with 50 iterations at each sample size (10 – 1,000,000). No filter was used to obtain these occurrence estimates. The horizontal dashed red line is the known number of presences assigned to the landscape.

Forest filter

The results are similar with informed sampling as they are with random sampling after filtering out cells with no forest. The total estimated number of occurrences leveled out below the 'true' number of presences regardless of that 'true' number (Figure 9). The rate at which the estimated occurrence values decrease after applying this filter is comparable to that of the randomly sampled landscape predictions.

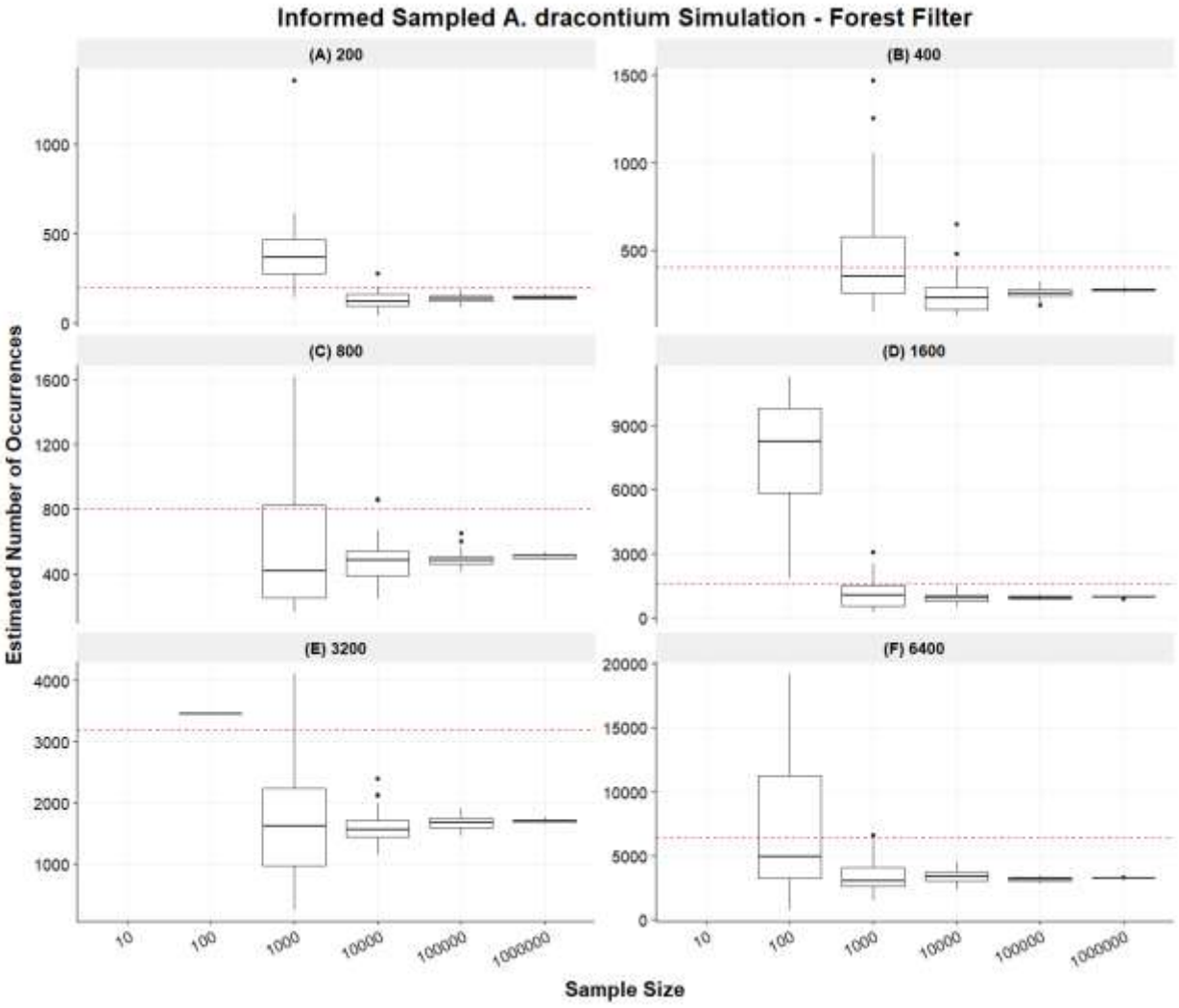


Figure 9: Informed sampling of simulated presences (200-6400) of *A. dracontium*, with 50 iterations at each sample size (10 – 1,000,000). Non-forested cells were removed prior to estimating occurrence numbers. The horizontal dashed red line is the known number of presences assigned to the landscape.

Lowest probability filter

The effect of applying the lowest probability filter to the informed sampling results shows the same trend of underestimating occurrences at larger sample sizes as with random sampling (Figure 10).

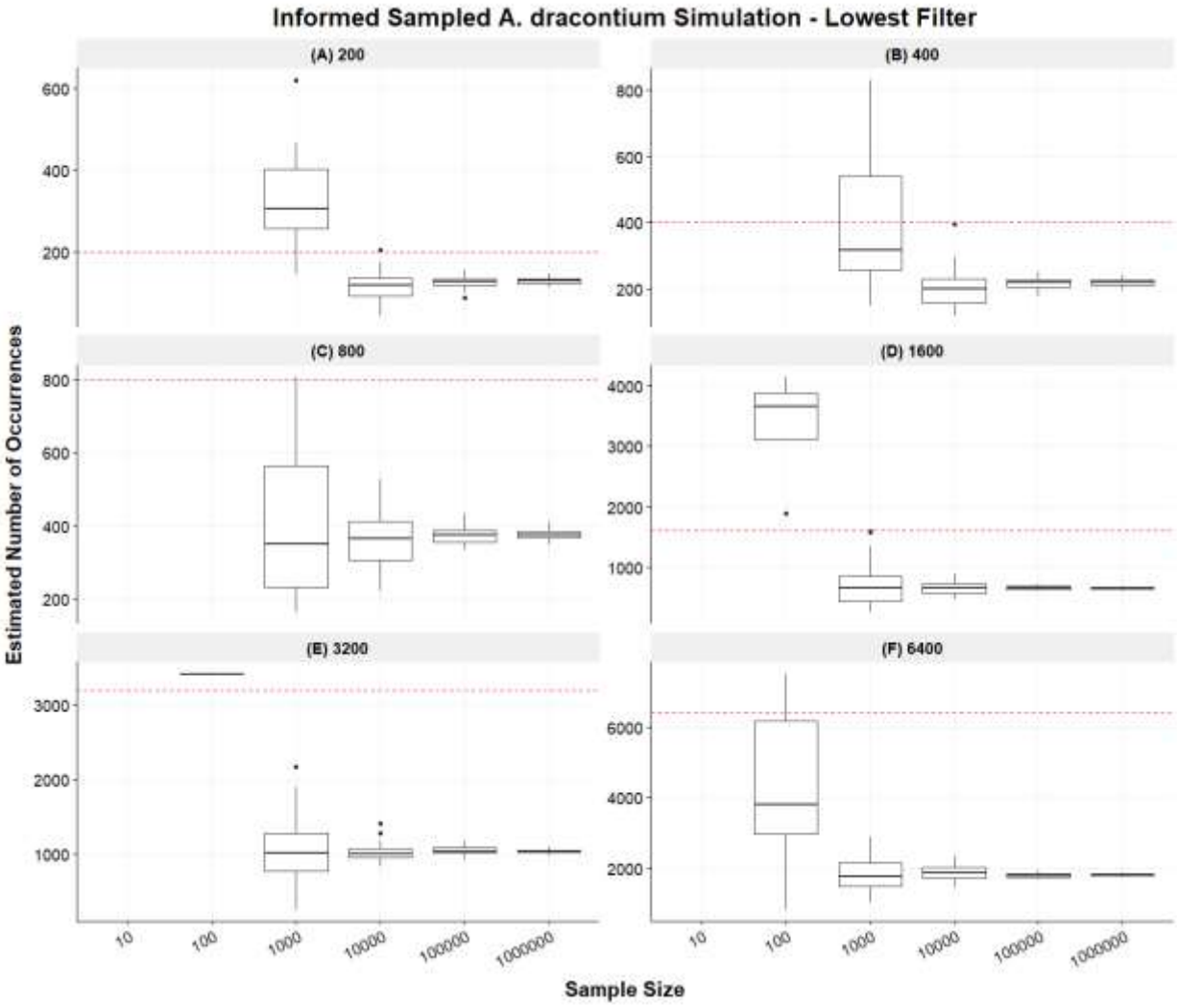


Figure 10: Informed sampling of simulated presences (200-6400) of *A. dracontium*, with 50 iterations at each sample size (10 – 1,000,000). The cells with a probability of containing a presence lower than 0.0158 were removed from occurrence estimates. The horizontal dashed red line is the known number of presences assigned to the landscape.

3.3 | Comparing random to informed sampling

The informed sampling resulted in a mean estimated number of presences closer to the actual number at an order of magnitude lower simulated sample size (Figure 11). By using the informed sampling technique, we detected more presences in our sample than by using the random sampling

technique (Figure 12). At simulated sample sizes <1000, random sampling did not find the necessary number of presences to produce a model, regardless of the number of assigned presences in the simulation (Figure 13).

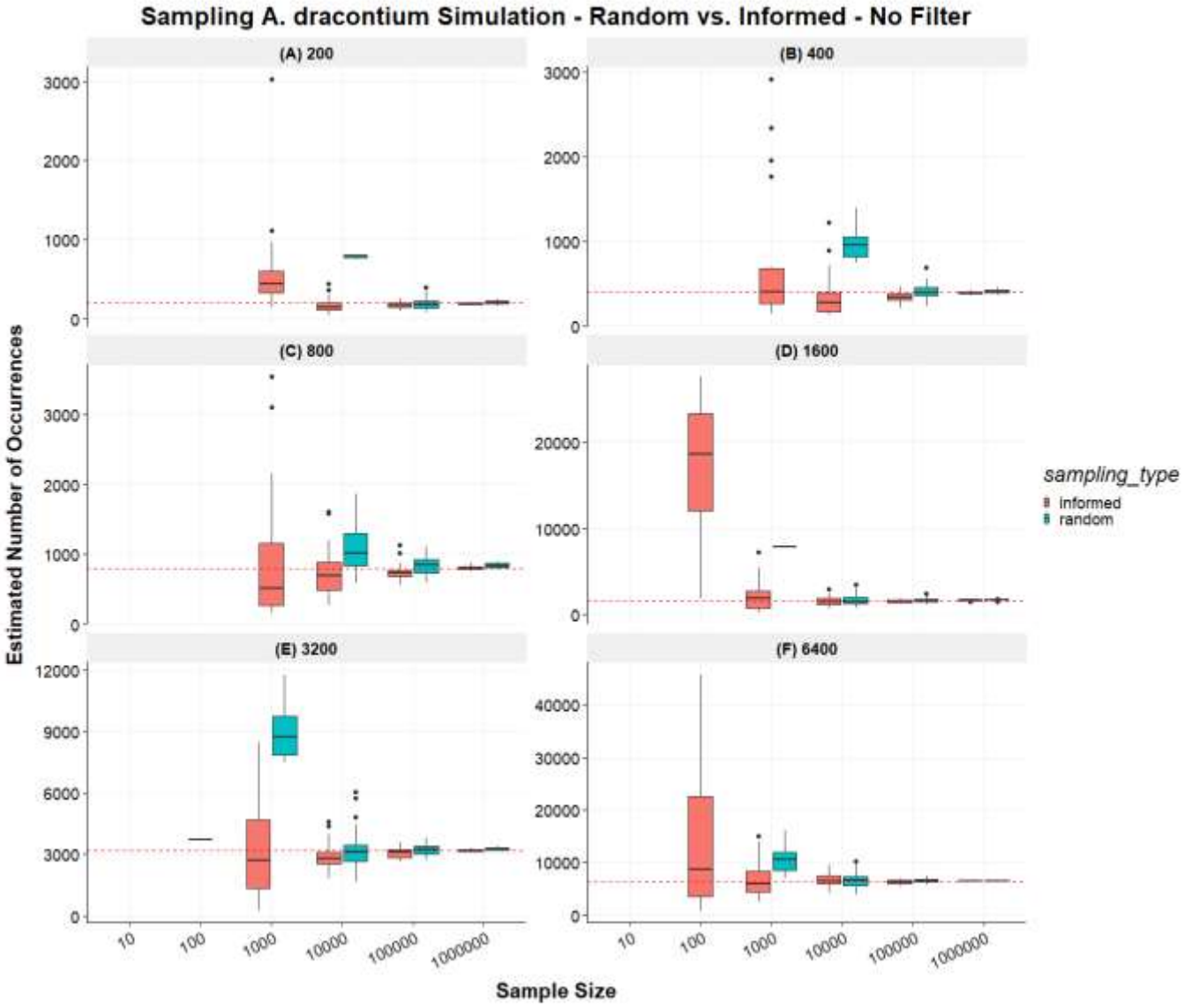


Figure 11: Random (blue) and informed (red) sampling of simulated presences (200 – 6400) of *A. dracontium, w*ith 50 sampling iterations at each sample size (10 – 1,000,000). No filter was used to obtain these occurrence estimates. The horizontal dashed red line is the known number of presences assigned to the landscape.

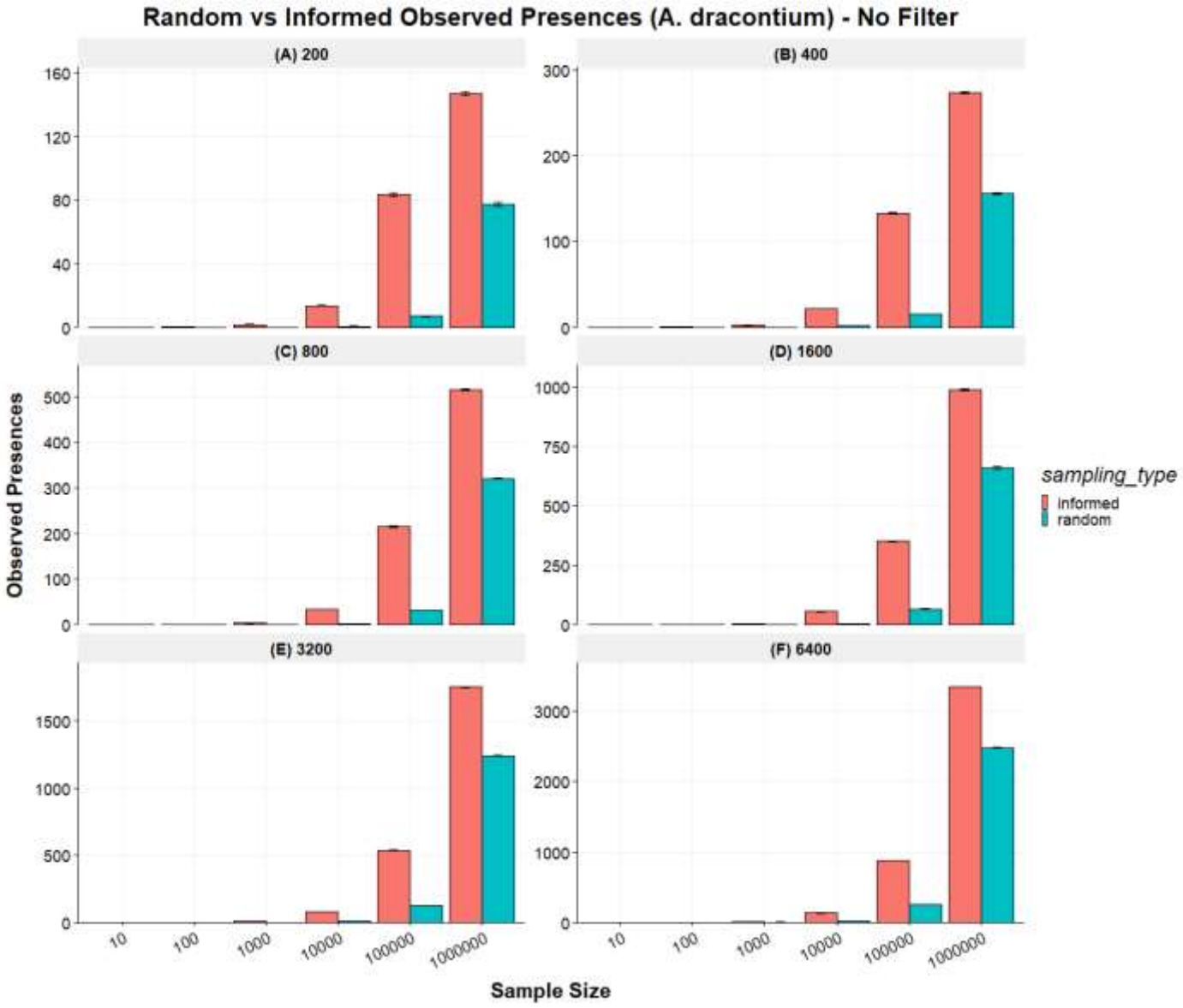


Figure 12: Panels A - F delineate the simulations by the number of presences assigned the landscape shown next to each letter. The mean number of observed presences for both informed (blue) and random (red) sampling techniques are depicted along the y-axis. No filter was used to obtain these occurrence estimates. The horizontal dashed red line is the known number of presences assigned to the landscape.

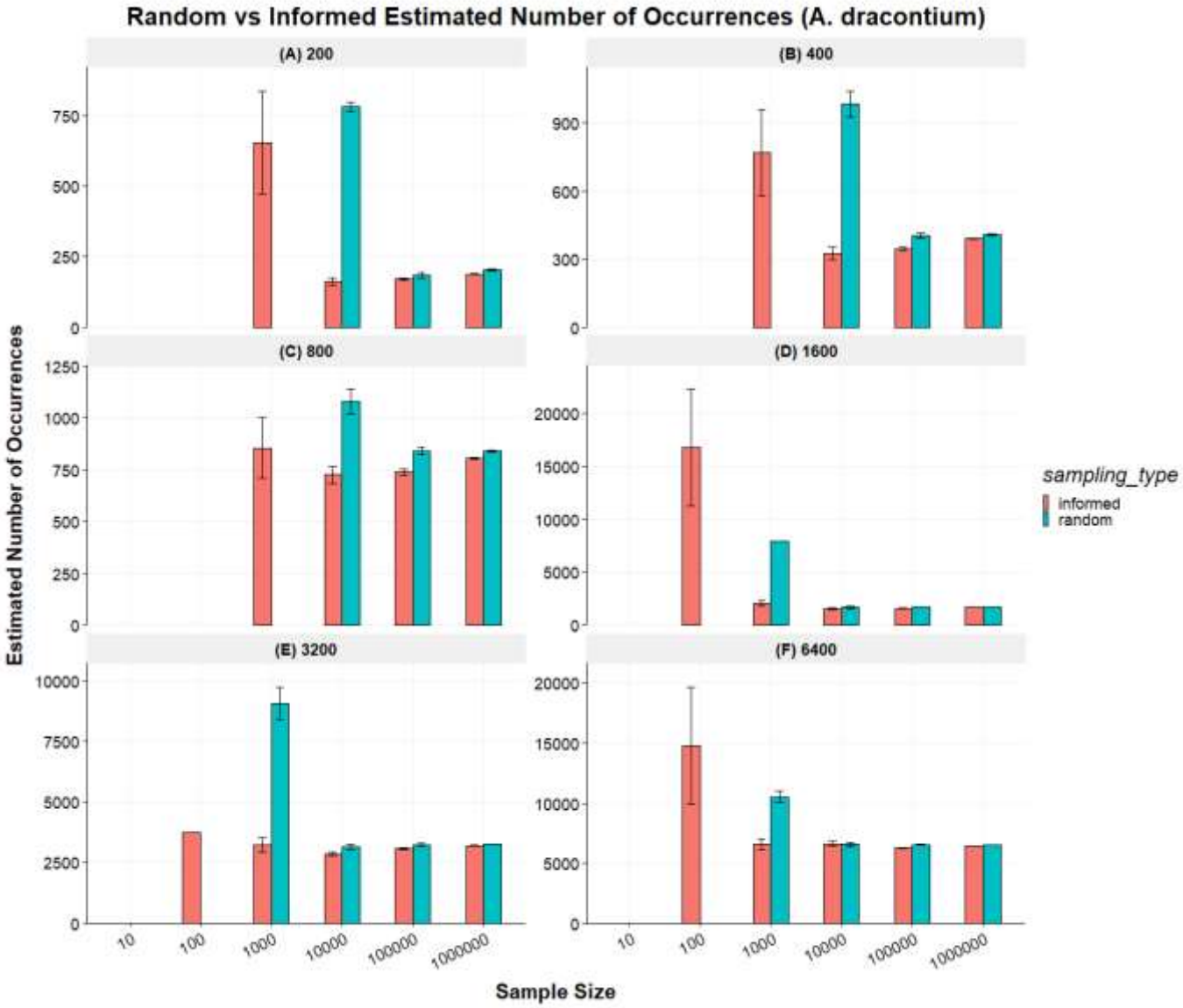


Figure 13: Panels A - F show the mean estimated number of occurrences at each number of presences, informed vs random sampling. No filter was used to obtain these estimates of occurrence.

DISCUSSION

Accurately assessing the conservation status of rare species relies in part on our knowledge of the true number of occurrences, through inventories and ongoing monitoring (Bickerton & Thompson-Black, 2010; Faber-Langendoen et al., 2012; Donley et al., 2013). Using SDMs to estimate habitat suitability and target field surveys for rare plants allows us to link SDM output to probability of occurrence, and theoretically estimate how many undiscovered populations may exist on a landscape. In this study we explored this novel application of SDMs using a case study of rare plant species in Southern Ontario. Specifically, the purposes of this study were: 1) to estimate the number of select rare plant species occurrences that remain undiscovered, and 2) to examine the confidence of these estimates using a simulation process. Our study confirmed that there are undiscovered occurrences of our focal species on the landscape. Even after applying two logical filters to the focal species estimation process, which decreased the number of presences by eliminating cells that had a very low probability of presences, the estimated number of occurrences of each focal species in Southern Ontario appears considerably greater than the current record of known occurrences (Table 3). However, the use of sampling and estimation simulations suggest that estimates of the total number of undiscovered populations are highly uncertain, especially when based on small sample sizes.

When this process of sampling and estimating is simulated on the *A. dracontium* SDM landscape, a pattern emerges where the predicted occurrences are overestimated at low sample sizes and settle around the 'true' number of presences once a random sample of at least 10,000 cells is taken. Notably, when no filter was applied the simulation produced

accurate estimates at larger sample sizes (>100,000). The simulations using informed sampling produced reasonably accurate estimates below this, at around 1,000 samples. When we applied the same filters to the simulated data as we did to the model results using real data, overall estimates of occurrences were lower than those assigned to the simulations.

Conservation implications

Given the apparent relative accuracy of our results using no filters, our method may have a practical application in the listing process (whether species are decidedly at some level of risk) by saving time and money before species are listed and re-assessed. If we can effectively estimate how many occurrences are on the landscape before a species is listed then we might be more efficient at ranking, monitoring, and effectively recovering species. With appropriate caution, the methods we have employed may be used to inform scientists on how to best proceed with species monitoring and abundance surveys/sampling programs. If occurrence estimates are high for a given species, then targeted surveys are more likely to be successful. If estimates are low, then resources should be allocated to the preservation and conservation of existing known habitat for these species.

SDMs can be used effectively to increase the efficiency of field surveys for undiscovered populations of rare plants (McCune, 2016; Rosner-Katz et al., accepted). However, without a large survey effort (on the order of at least 1,000 survey sites using an informed sampling method), we cannot accurately estimate the 'true' number of occurrences on the landscape. Therefore, although we predict that there are likely more populations of *A. dracontium* (for example) than are currently known, we cannot say with certainty how many remain

undiscovered. We recognize that there are practical limitations to surveying 1,000 cells for a rare species. Nevertheless, we still must make conservation decisions based on the best information we have (i.e. the number of known occurrences) until further survey efforts result in a total number of known populations that is greater than the threshold for species-at-risk status.

Even though our analysis suggested many undiscovered occurrences for the focal species (Table 3), we caution against using such information to downgrade the conservation status of these species. Undoubtedly, undiscovered occurrences exist, otherwise the surveyors would not have found them. For example, in 282 plots surveyed over the course of four field seasons, 13 new occurrences of *C. florida* were discovered when surveying only a small portion of remaining habitat (Appendix Table 1). Though these new discoveries are crucial to the process of understanding the status of these at-risk species, it is important to note that the number of populations and their abundance is not the sole deciding factor as to whether a species should be listed. There are other criteria that are considered, including range extent, population size, distribution and current and future threats.

Main caveats and limitations

There are many limitations and assumptions to this study including the use of SDMs for modeling rare species distributions, the uncertainty in our probability estimations, and the use of the logical filters to reduce over-estimation. Limited occurrence data regarding rare species results in difficulty applying the common approaches to modeling species distributions (Engler et al., 2004; Williams et al., 2009). The SDMs also incorporate the sampling biases inherent in

the occurrence data by modeling a species' realized niche rather than its fundamental niche (Williams et al., 2009). We acknowledge that the SDMs used as the foundation of the simulations are incomplete depictions of the true species distributions and habitat requirements and that we are assuming these models to be accurate when using them as we have. We recognize that there may be ecological constraints such as dispersal barriers or edaphic constraints that fundamentally limit the species distribution. However, the successful discovery of multiple new occurrences directed by our focal species SDMs which were validated using independent data, gives us a level of confidence in the high-performance metrics of these models.

The six focal species (Table 1) chosen in Southern Ontario are found at the northern edge of their range and have large portions of their ranges which extend south into the United States. Consequently, our study region only captures part of each species range in their corresponding SDMs rather than their entire global ranges of occurrence; this is not likely a concern however when it comes to model accuracy. According to Luoto et al. (2005), distribution models representing species at the margin of their range are in many cases more accurate than for widespread species whose full range is included in the model.

Our focal species occurrence estimates have wide margins of uncertainty (i.e. based on our field surveys, we obtained an estimate of between 2,056 and 9,624 occurrences of *A. dracontium* in our study area, Table 3), which supports our rationale for exploring the simulation of species occurrence estimation. The simulation results reveal a trend that larger sample sizes allow for more accurate estimates to be calculated. However, the simulations also allowed us to recognize potential systematic errors in GLMs leading to biases in the estimates

of remaining occurrences. There is consistent overestimation of occurrences at low sample sizes when we simulate presences on a landscape and then simulate the sampling of those presences. This is likely due to the bias in only creating a GLM in a sample where a sufficient number of occurrences were detected in our simulated samples (i.e. we removed all resulting samples that observed/detected less than 3 presences). The average number of presences observed at each sample size is shown in Figure 12. At small sample sizes, the simulated samples that could be used in models represent potential outliers: most samples did not contain any occurrences. Thus, they may lead to biased models overestimating habitat suitability and therefore number of occurrences. In the real world, there is risk of a similar phenomenon occurring: when biologists survey for rare species and don't find many, even just by chance, they will not have sufficient data to build a model. However, if by chance they do find many detections they could build a model. The potential implications of this phenomenon for modeling occurrences has, to our knowledge, not been explored.

Applying the filters to our simulated occurrence results also reveals a challenge with using the GLMs. Figure 6 shows how removing the non-forested cells underestimates the number of occurrences. This may be due to an inherent property of binomial GLMs. Even a well-fit binomial GLM may underestimate the highest predicted probability values (i.e., cells where a species is actually present will have predicted probability <1 in the model), and overestimate the lowest predicted probability values (i.e. assigning small but non-zero component of probability space to areas where occurrence is impossible). The filters effectively cut off the low values where the curve is slightly overfit, which across the many removed low-probability cells, results in an overall underestimation of the summed probability of the species being present in

remaining cells. Filtering out the many values with very low probabilities thus resulted in estimated occurrence numbers which were less than the true number of occurrences that were assigned to the landscape. Removing non-forested cells will also tend to remove low-predicted-probability cells, resulting in the same underestimate as when filtering out cells below the lowest probability cell that contains a known occurrence.

Filters such as these may help to direct surveys as a logical technique to refine a search, however, they may not always be useful to assist estimates of occurrence numbers. Thus, we caution the use of filters of this type because of the high uncertainty in our range of estimates (Table 3).

Future work

This study cannot, on its own, be used to suggest population trends over time. Rather, it takes known presence and absence records from a 4-year window in time and attempts to estimate the number of current occurrences. We recognize that the discovery of previously unknown occurrences on the landscape may not balance out the concurrent losses from unreported extinctions (Kaye et al., 2019). We suggest future work to use these occurrence estimation methods (including the building of SDMs) along with updated land use and land cover data to show how the species distributions may shift along with the estimates of how many occurrences are out there, and further exploration as to whether this would work. We know that land use change resulting in habitat loss is a major cause of biodiversity loss (Kaye et al., 2019), therefore in a region like Southern Ontario we can expect the amount of suitable habitat for these species to continue to decline.

We also suggest a companion study looking at population densities be conducted in Southern Ontario, following methods used by Tôrres et al. (2012) to examine jaguar distributions and abundance in the Neotropics. Tôrres et al. (2012) found that low jaguar densities occurred in areas with both low and high habitat suitability, while high values were restricted to areas where the habitat was highly suitable. A similar study conducted in Southern Ontario could test the same methods on rare plant species distributions and abundance. Perhaps this could be another way to limit the cost of surveys while still being able to determine the level of threat that these species face in their remnant habitat patches. Results from Tôrres et al. (2012) implied that the more highly suitable the area is, the more that jaguars are able to persist in those areas, suggesting the importance of conserving the highest quality habitat for species at risk of extinction.

Conclusion

There are numerous challenges in the field of conservation when it comes to understanding and protecting species at risk. Due to their rarity, some species can be difficult to study, and their occurrences can be difficult to predict. We have developed methods for predictive modelling of select species probabilities of occurrence in order to estimate how many undiscovered populations remain on the landscape. When using high performing SDMs, we can predict sites with a high likelihood of occurrence. This higher likelihood results in an overall higher estimate of the number of occurrences than known records would suggest. Our technique appears to provide accurate estimates; however, we advise caution and further exploration of the effect of low sample sizes typical in rare species studies.

REFERENCES

Arponen, A. (2012). Prioritizing species for conservation planning. *Biodiversity and Conservation*, *21*(4), 875–893. https://doi.org/10.1007/s10531-012-0242-1

Balmford, A., Fisher, B., Green, R. E., Naidoo, R., Strassburg, B., Turner, R. K., & Rodrigues, A. S. (2011). Bringing ecosystem services into the real world: an operational framework for assessing the economic consequences of losing wild nature. *Environmental and Resource Economics*, *48*(2), 161-175. https://doi.org/10.1007/s10640-010-9413-2

Bayley, P. B., & Peterson, J. T. (2001). An approach to estimate probability of presence and richness of fish species. *Transactions of the American Fisheries Society*, *130*(4), 620–633. https://doi.org/10.1577/1548-8659(2001)130<0620:aatepo>2.0.co;2

Bennett, J. R., & Arcese, P. (2013). Human influence and classical biogeographic predictors of rare species occurrence. *Conservation Biology*, *27*(2), 417–421. https://doi.org/10.1111/cobi.12015

Bennett, J. R., Maloney, R., & Possingham, H. P. (2015). Biodiversity gains from efficient use of private sponsorship for flagship species conservation. *Proceedings of the Royal Society B:Biological Sciences, 282*(1805), 20142693. https://doi.org/10.1098/rspb.2014.2693

Bickerton, H. and M. Thompson-Black. (2010). Recovery Strategy for the Eastern Flowering Dogwood (*Cornus florida*) in Ontario. Ontario Recovery Strategy Series. Prepared for the Ontario Ministry of Natural Resources, Peterborough, Ontario. vi+ 21 pp. Retrieved from: https://www.ontario.ca/page/eastern-flowering-dogwood-recovery-strategy

Boetsch, J. R., van Manen, F. K., & Clark, J. D. (2003). Predicting rare plant occurrence in Great Smoky Mountains National Park, USA. *Natural Areas Journal 23*(3) 229-237.

Boland, G.J., J. Ambrose, B. Husband, K.A. Elliott and M.S. Melzer. (2012). Recovery Strategy for the American Chestnut (*Castanea dentata*) in Ontario. Ontario Recovery Strategy Series. Prepared for the Ontario Ministry of Natural Resources, Peterborough, Ontario. vi + 43 pp. Retrieved from: https://www.ontario.ca/page/american-chestnut-recovery strategy

Boles, R. L., Lovett-Doust, J., & Lovett-Doust, L. (2000). Population genetic structure in green dragon (Arisaema dracontium, Araceae). *Canadian Journal of Botany*, *77*(10), 1401–1410.https://doi.org/10.1139/b99-089

Burnham, K. P. and Anderson, D. R. (2002). Model selection and inference: a practical information - theoretic approach, 2nd ed. - Springer. https://doi.org/10.1007/978-0-387-224565_3

Cardillo, M., Purvis, A., Sechrest, W., Gittleman, J. L., Bielby, J., & Mace, G. M. (2004). Human population density and extinction risk in the world's carnivores. *PLoS Biology*, *2*(7), 909–914. https://doi.org/10.1371/journal.pbio.0020197

Coristine, L. E., Jacob, A. L., Schuster, R., Otto, S. P., Baron, N. E., Bennett, N. J., Bittick, S. J., Dey, C., Favaro, B., Ford, A., Nowlan, L., Orihel, D., Palen, W. J., Plofus, J. L., Shiffman, D. S., Venter, O., & Woodley, S.

(2018). Informing Canada's commitment to biodiversity conservation: A science-based framework to help guide protected areas designation through Target 1 and beyond. *FACETS*. https://doi.org/10.1139/facets-2017-0102

COSEWIC. (2014). COSEWIC assessment and status report on the Blue Ash Fraxinus quadrangulata in Canada. Committee on the Status of Endangered Wildlife in Canada. Ottawa. xiii + 58 pp. Retrieved from: https://www.canada.ca/en/environment-climate change/services/species-risk-public-registry/cosewic-assessments-status-reports/blue-ash2014.html

COSEWIC wildlife species assessment: candidate wildlife species. (2017). Priority setting for new wildlife species to be assessed. Retrieved from: https://www.canada.ca/en/environmentclimate-change/services/committee-statusendangered-wildlife/wildlife-species-assessment-process-categoriesguidelines/candidate.html

Di Fonzo, M. M. I., Possingham, H. P., Probert, W. J. M., Bennett, J. R., Joseph, L. N., Tulloch, A. I. T., O'Connor, S., Densem, J., & Maloney, R. F. (2016). Evaluating trade-offs between target persistence levels and numbers of species conserved. *Conservation Letters 9*(1) 51-57. doi:10.1111/conl.12179

Donley, R., J.V. Jalava and J. van Overbeeke. (2013). Management Plan for the Green Dragon (*Arisaema dracontium*) in Ontario. Ontario Management Plan Series. Prepared for the Ontario Ministry of Natural Resources, Peterborough, Ontario. vi + 43 pp. Retrieved from: https://www.ontario.ca/page/green-dragon-management-plan

Dunk, J. R., Zielinksi, W. J., & Preisler, H. K. (2004). Predicting the occurrence of rare mollusks in northern California Forests. *Ecological Applications, 14*(3), 713-729. https://doi.org/10.1890/025322

ECCC. (2018). Management Plan for the Crooked-stem Aster (Symphyotrichum prenanthoides) in Canada [Proposed]. Species at Risk Act Management Plan Series. Environment and Climate Change Canada, Ottawa. v + 31 pp. Retrieved from: https://www.canada.ca/en/environment-climate-change/services/species-risk-public-registry/management-plans/crooked-stem-aster-2018-proposed.html

Elith, J., Graham, C. H., Anderson, R. P., Dudı´k, M., Ferrier, S., Guisan, A., Hijmans, R. J., Huettmann, F., Leathwick, J. R., Lehmann, A., Li, J., Lohmann, L. G., Loiselle, B. A., Manion, G., Moritz, C., Nakamura, M., Nakazawa, Y., Overton, J. McC., Peterson, A. T., Phillips, S. J., Richardson, K. S., Scachetti-Pereira, R., Schapire, R. E., Sobero´n, J., Williams, S., Wisz, M. S. and Zimmermann, N. E. (2006). Novel methods improve prediction of species' distributions from occurrence data. *Ecography, 29(2)*, 129-151. https://doi.org/10.1111/j.2006.09067590.04596.x

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. https://doi.org/10.1111/j.1600-0587.2008.05505.x

Elith, J., Phillips, S. J., Hastie, T., Dudík, M., Chee, Y. E., & Yates, C. J. (2011). A statistical explanation of MaxEnt for ecologists. *Diversity and distributions*, *17*(1), 43-57. https://doi.org/10.1111/j.1472-4642.2010.00725.x

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. https://doi.org/10.1111/j.0021-8901.2004.00881.x

Faber-Langendoen, D., J. N., Master, L., Snow, K., Tomaino, A., Bittman, R., Hammerson, G., Heidel, B., Ramsay, L., Teucher, A., & Young, B. (2012). NatureServe Conservation Status Assessments: Methodology for Assigning Ranks, (June), 50. Retrieved from http://www.natureserve.org/biodiversity-science/publications/natureserve-conservation status-assessments-methodology-assigning

Favaro, B., D. C. Claar, C. H. Fox, C. Freshwater, J. J. Holden, A. Roberts, & Derby, U. R. (2014). Trends in extinction risk for imperiled species in Canada. *PLoS ONE, 9*(11). https://doi.org/10.1371/journal.pone.0113118

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*(1), 38–49. https://doi.org/10.1017/s0376892997000088

Fois, M., Cuena-Lombraña, A., Fenu, G., & Bacchetta, G. (2018). Using species distribution models at local scale to guide the search of poorly known species: Review, methodological issues and future directions. *Ecological Modelling*, *385*, 124-132. https://doi.org/10.1016/j.ecolmodel.2018.07.018

Goefroid, S. & Koedam, N. (2003). Identifying indicator plant species of habitat quality and invasibility as a guide for peri-urban forest management. *Biodiversity and Conservation, 12*(8)*,*1699-1713. https://doi.org/10.1023/a:1023606300039

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. https://doi.org/10.1890/09-1190.1

Government of Canada (2019). Schedule 1: List of Wildlife Species at Risk, Justice Laws Website. https://laws.justice.gc.ca/eng/acts/S-15.3/page-17.html#h-435647

Guisan, A., Broennimann, O., Engler, R., Vust, M., Yoccoz, N.G., Lehmann, A. & Zimmerman, N.E. (2006). Using niche-based models to improve the sampling of rare species. *Conservation Biology, 20*(2), 501–511. https://doi.org/10.1111/j.1523-1739.2006.00354.x

Guisan, A., Tingley, R., Baumgartner, J. B., Naujokaitis-Lewis, I., Sutcliffe, Tulloch, A. I. T., Regan, T. J., Brotons, L., McDonald-Madden, E., Mantyka-Pringle, C., Martin, T. G., Rhodes, J. R., Maggini, R., Setterfield, S. A., Elith, J., Schwartz, M. W., Wintle, B. A., Broennimann, O., Austin, M., Ferrier, S., Kearney, M. R., Possingham, H. P., & Buckley, Y. M. (2013). Predicting species distributions for conservation decisions. *Ecology Letters, 16*(12), 1424- 1435. https://doi.org/10.1111/ele.12189

Havens, K., Kramer, A. T., & Guerrant, E. O. (2014). Getting plant conservation right (or not): the case of the United States. *International Journal of Plant Sciences*, *175*(1), 3–10.https://doi.org/10.1086/674103

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*(5), 773-785. https://doi.org/10.1111/j.0906-7590.2006.04700.x

Heywood, V. H. (2017). Plant conservation in the Anthropocene – challenges and future prospects. *Plant Diversity*, *39*(6), 314–330. https://doi.org/10.1016/j.pld.2017.10.004

Heywood, V. H. (2019). Conserving plants within and beyond protected areas-still problematic and future uncertain. *Plant Diversity*, *41*(2), 36–49. https://doi.org/10.1016/j.pld.2018.10.001

Hilty, J., & Merenlender, A. M. (2003). Studying biodiversity on private lands. *Conservation Biology*, *17*(1), 132–137.

Kaye, T. N., Bahm, M. A., Thorpe, A. S., Gray, E. C., Pfingsten, I., & Waddell, C. (2019). Population extinctions driven by climate change, population size, and time since observation may make rare species databases inaccurate. *PloS one*, *14*(10). https://doi.org/10.1371/journal.pone.0210378

Kerr, J. T., & Cihlar, J. (2004). Patterns and causes of species endangerment in Canada. *Ecological Applications*, *14*(3), 743-753. https://doi.org/10.1890/02-5117

Kerr, J. T., Currie, D. J. (1995). Effects of human activity on global extinction risk. *Conservation Biology*, *9*(6), 1528–1538. https://doi.org/10.1046/j.15231739.1995.09061528.x

Kerr, J. T., & Deguise, I. (2004). Habitat loss and the limits to endangered species recovery. *Ecology Letters*, *7*(12), 1163–1169. https://doi.org/10.1111/j.1461-0248.2004.00676.x

Langford, W. T., Gordon, A., Bastin, L., Bekessy, S. A., White, M. D., & Newell, G. (2011). Raising the bar for systematic conservation planning. *Trends in Ecology and Evolution*, *26*(12), 634–640. https://doi.org/10.1016/j.tree.2011.08.001

Leopold, A. (1949). *A Sand County almanac and sketches here and there*. Oxford University Press, New York, USA.

Luoto, M., Poyry, J., Heikkinen, R. K., & Saarinen, K. (2005). Uncertainty of bioclimate envelope models based on the geographcial distribution of species. *Global Ecology and Biogeography, 14*(6)*,* 575-584. https://doi.org/10.1111/j.1466-822x.2005.00186.x

Mackenzie, D. K., Nichols, J. D., Hines, J. E., Knutson, M. G., & Franklin, A. B. (2003). Estimating site occupancy, colonization, and local extinction when a species is detected imperfectly. *Ecology, 84*(8), 2200-2207. https://doi.org/10.1890/02-3090

Martinez-Meyer, E., Peterson, A. T., Servín, J. I., & Kiff, L. F. (2006). Ecological niche modelling and prioritizing areas for species reintroductions. *Oryx*, *40*(4), 411-418. https://doi.org/10.1017/s0030605306001360

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. https://doi.org/10.1111/1365-2664.12702

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. https://doi.org/10.1007/s10980-017-0491-1

McCune, J. L. (2019). A new record of Stylophorum diphyllum (Michx.) Nutt. in Canada: A case study of the value and limitations of building species distribution models for very rare plants. *The Journal of the Torrey Botanical Society*, *146*(2), 119-127. https://doi.org/10.3159/torrey-d-18-00026.1

McDonald, L. L. (2004). Sampling rare populations in W. L. Thompson (Ed.), *Sampling rare or elusive species: concepts, designs, and techniques for estimating population parameters* (pp. 11-42). Washington: Island Press.

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. https://doi.org/10.1111/j.1600-0587.2013.07872.x

Merow, C., Smith, M. J., Edwards, T. C., Guisan, A., Mcmahon, S. M., Normand, S., Thuiller, W., Wüest, R. O., Zimmermann, N. E., & Elith, J. (2014). What do we gain from simplicity versus complexity in species distribution models? *Ecography, 37*(12), 1267-1281. https://doi.org/10.1111/ecog.00845

Myers, N., Mittermeier, R. A., Mittermeier, C. G., Fonseca, G. A. B., & Kent, J. (2000). Biodiversity hotspots for conservation priorities. *Nature, 403*(6772), 853–858. https://doi.org/10.1038/35002501

NatureServe, 2002. Element Occurrence Data Standard, in cooperation with the Network of Natural Heritage Programs and Conservation Data Centers. Retrieved from: http://help.natureserve.org/biotics/Content/Methodology/EO_DataStandard.pdf

NatureServe, 2019. NatureServe Explorer: An online encyclopedia of life [web application] Version 7.1. NatureServe, Arlington, Virginian. Retrieved from: http://explorer.natureserve.org/servlet/NatureServe?searchName=Lithospermum+latifolium.

Oldham, M. (2017). *List of the Vascular Plants of Ontario ' s Carolinian Zone (Ecoregion 7E)*.https://doi.org/10.13140/RG.2.2.34637.33764

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(1), 102–117. https://doi.org/10.1111/j.13652699.2006.01594.x

Phillips, S. J., Anderson, R. P., & Schapire, R. E. (2006). Maximum entropy modeling of species geographic distributions. *Ecological Modelling, 190*(3-4), 231–259. https://doi.org/10.1016/j.ecolmodel.2005.03.026

Pimm, S. L. (2014). The biodiversity of species and their rates of extinction, distribution, and protection. *Science, 344*(6187), 1246752. https://doi.org/10.1126/science.1246752

Possingham, H. P., Andelman, S. J., Burgman, M. A., Medellín, R. A., Master, L. L., Keith, D. A. (2002). Limits to the use of threatened species lists. *Trends in Ecology and Evolution, 17*(11), 503–507. https://doi.org/10.1016/S0169-5347(02)02614-9

Rayner, L., Lindenmayer, D. B., Wood, J. T., Gibbons, P., & Manning, A. D. (2014). Are protected areas maintaining bird diversity? *Ecography*, *37*(1), 43–53. https://doi.org/10.1111/j.1600-0587.2013.00388.

Reznicek, A. A., Voss, E. G., and Walters, B. S. (2011). University of Michigan. Retrieved from: https://michiganflora.net/species.aspx?id=580.

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. https://doi.org/10.7717/peerj.3632

Richart, M., & Hewitt, N. (2008). Forest remnants in the long point region, Southern Ontario: Tree species diversity and size structure. *Landscape and Urban Planning, 86*(1) 25-37. https://doi.org/10.1016/j.landurbplan.2007.12.005

Rosner-Katz, H., McCune, J. L., and Bennett, J. R. (accepted) Using stacked SDMs with accuracy and threat weighting to optimize surveys for threatened plant species. *Biodiversity and Conservation.*

Ross, S. 1994. A first course in probability, 4th ed. Prentice-Hall, Englewood Cliffs, NJ.

Rothfels, C. & T. Smith. 2003. Update COSEWIC Status Report on Green Dragon, Arisaema dracontium (L.) Schott. [REVIEW DRAFT]. Manuscript. 48 pp

Sanderson, E. W., Jaiteh, M., Levy, M. A., Redford, K. H., Wannebo, A. V., & W, G. (2002). The Human Footprint and the Last of the Wild. *BioScience*, *52*(10), 891. https://doi.org/10.1641/0006-3568(2002)052[0891:thfatl]2.0.co;2

Schemske, D. W., Husband, B. C., Ruckelshaus, M. H., Goodwillie, C., Parker, I. M., & Bishop, J. G. (1994). Evaluating approaches to the conservation of rare and endangered plants. *Ecology*, *75*(3), 584-606. https://doi.org/10.2307/1941718

Schuster, R., Wilson, S., Rodewald, A. D., Arcese, P., Fink, D., Auer, T., & Bennett, J. R. (2019). Optimizing the conservation of migratory species over their full annual cycle. *Nature communications*, *10*(1), 1-8. https://doi.org/10.1038/s41467-019-09723-8

Tôrres, N. M., De Marco, P., Santos, T., Silveira, L., de Almeida Jácomo, A. T., & Diniz-Filho, J. A. F. (2012). Can species distribution modelling provide estimates of population densities? A case study with jaguars in the Neotropics. *Diversity and Distributions*, *18*(6), 615–627. https://doi.org/10.1111/j.1472-4642.2012.00892.x

Tulloch, A. I. T., Maloney, R. F., Joseph, L. N., Bennett, J. R., Di Fonzo, M. M. I., Probert, W. J. M., O'Connor, S. M., Densem, J. P., & Possingham, H. P. (2015). Effect of risk aversion on prioritizing conservation projects. *Conservation Biology*, *29*(2), 513- 524.https://doi.org/10.1111/cobi.12386

Tulloch, A. I. T., Barnes, M. D., Ringma, J., Fuller, R. A., & Watson, J. E. M. (2016). Understanding the importance of small patches of habitat for conservation. *Journal of Applied Ecology*, *53*(2), 418–429. https://doi.org/10.1111/1365-2664.12547

USDA (2019). Lithospermum latifolium Michx. American stoneseed, United States Department of Agriculture, Natural Resources Conservation Service. Retrieved from https://plants.usda.gov/core/profile?symbol=LILA2.

Van Proosdij, A. J., Sosef, M. S. M., Wieringa, J. J., & Raes, N. (2016). Minimum required number of specimen records to develop accurate species distribution models. *Ecography, 39*(6), 542-552. https://doi.org/10.1111/ecog.01509

Wilcove, D. S. (2004). The private side of conservation. *Frontiers in Ecology and the Environment, 2*(6), 326-327. https://doi.org/10.1890/1540-9295(2004)002[0326:TPSOC]2.0.CO;2

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. https://doi.org/10.1111/j.1472-4642.2009.00567.x

APPENDICES

Appendix Table 1: Environmental predictor variables included in each of the eight models built for each focal species, and model evaluations (including AUC and TPR values). Bolded/highlighted text indicates the best model (see main text for criteria).

species	model	environmental predictor variables	# indep presences (surveys only)	# independent presences (survey + NHIC)	# independent absences	AUC_indep (surveys only)	TPR_indep (surveys only)	AUC_indep (surveys + NHIC central data)	TPR_indep (surveys + NHIC central data)
Arisaema dracontium	arisdra_03	original(1)	19	113	261	0.8332	0.9473	0.7062	0.8230
Arisaema dracontium	arisdra_06	original(0.5)	19	113	261	0.8260	0.8947	0.7096	0.6549
Arisaema dracontium	arisdra_10	orig+SOLRIS(1)	19	113	250	0.8149	0.9474	0.637	0.7965
Arisaema dracontium	arisdra_11	orig+SOLRIS(0.5)	19	113	250	0.8038	0.9474	0.637	0.6814
Arisaema dracontium	arisdra_12	orig+cont(1)	19	113	261	0.8500	0.8947	0.71	0.7257
Arisaema dracontium	arisdra_13	orig+cont(0.5)	19	113	261	0.8532	0.8947	0.7601	0.7257
Arisaema dracontium	**arisdra_14**	orig+SOLRIS+cont (1)	**19**	**113**	**250**	**0.8722**	**0.9474**	**0.7185**	**0.7788**
Arisaema dracontium	arisdra_15	orig+SOLRIS+cont (0.5)	19	113	250	0.8606	0.9474	0.7357	0.7080
Cornus florida	**cornflo_01**	original(1)	**13**	**86**	**245**	**0.8421**	**0.8462**	**0.7637**	**0.6744**
Cornus florida	cornflo_04	original(0.5)	13	86	245	0.8200	0.7692	0.8025	0.6744
Cornus florida	cornflo_08	orig+SOLRIS(1)	13	86	234	0.8360	0.8462	0.7297	0.6977
Cornus florida	cornflo_09	orig+SOLRIS(0.5)	13	86	234	0.8248	0.7692	0.7841	0.7209
Cornus florida	cornflo_10	orig+cont(1)	13	86	245	0.8245	0.6154	0.7087	0.3837
Cornus florida	cornflo_11	orig+cont(0.5)	13	86	245	0.7997	0.6154	0.7872	0.5698
Cornus florida	cornflo_12	orig+SOLRIS+cont (1)	13	86	234	0.7906	0.7692	0.6803	0.4602
Cornus florida	cornflo_13	orig+SOLRIS+cont (0.5)	13	86	234	0.8198	0.6923	0.7326	0.4884
Symphyotrichum prenanthoides	**symppre_01**	original(1)	**10**	**30**	**226**	**0.9119**	**0.5000**	**0.8861**	**0.3667**
Symphyotrichum prenanthoides	symppre_02	original(0.5)	10	30	226	0.9119	0.3000	0.8951	0.2000
Symphyotrichum prenanthoides	symppre_03	orig+SOLRIS(1)	10	30	215	0.8972	0.1000	0.8730	0.2000
Symphyotrichum prenanthoides	symppre_04	orig+SOLRIS(0.5)	10	30	215	0.9093	0.4000	0.8884	0.2667
Symphyotrichum prenanthoides	symppre_05	orig+cont(1)	10	30	226	0.9049	0.2000	0.8873	0.2333
Symphyotrichum prenanthoides	symppre_06	orig+cont(0.5)	10	30	226	0.9097	0.1000	0.8982	0.1333
Symphyotrichum prenanthoides	symppre_07	orig+SOLRIS+cont (1)	10	30	215	0.9033	0.4000	0.8826	0.2000
Symphyotrichum prenanthoides	symppre_08	orig+SOLRIS+cont (0.5)	10	30	215	0.9047	0.3000	0.8786	0.2667
Fraxinus quadrifolia	fraxqua_1	original(1)	5	69	275	0.8938	1.0000	0.8246	0.7536
Fraxinus quadrifolia	fraxqua_2	original(0.5)	5	111	275	0.8800	0.8000	0.8536	0.6957
Fraxinus quadrifolia	fraxqua_3	orig+SOLRIS(1)	5	111	264	0.8833	1.0000	0.7873	0.7101
Fraxinus quadrifolia	fraxqua_4	orig+SOLRIS(0.5)	5	111	264	0.8962	0.0000	0.7885	0.0000
Fraxinus quadrifolia	fraxqua_5	orig+cont(1)	5	111	275	0.8822	1.0000	0.7803	0.6811
Fraxinus quadrifolia	**fraxqua_6**	orig+cont(0.5)	**5**	**111**	**275**	**0.9033**	**1.0000**	**0.8203**	**0.6377**
Fraxinus quadrifolia	fraxqua_7	orig+SOLRIS+cont (1)	5	111	264	0.8561	1.0000	0.7751	0.7246
Fraxinus quadrifolia	fraxqua_8	orig+SOLRIS+cont (0.5)	5	111	264	0.8795	1.0000	0.8092	0.5942
Lithospermum latifolium	LITHLAT_1	original(1)	8	29	243	0.8631	1.0000	0.6044	0.6552
Lithospermum latifolium	LITHLAT_2	original(0.5)	8	29	243	0.8879	1.0000	0.6093	0.6207
Lithospermum latifolium	LITHLAT_4	orig+SOLRIS(1)	8	29	243	0.8049	0.6250	0.5515	0.4138
Lithospermum latifolium	LITHLAT_5	orig+SOLRIS(0.5)	8	29	243	0.7268	0.6250	0.5187	0.4483
Lithospermum latifolium	LITHLAT_6	orig+cont(1)	8	29	243	0.9192	1.0000	0.6330	0.6552
Lithospermum latifolium	**LITHLAT_7**	orig+cont(0.5)	**8**	**29**	**243**	**0.9203**	**1.0000**	**0.6374**	**0.6207**
Lithospermum latifolium	LITHLAT_8	orig+SOLRIS+cont (1)	8	29	243	0.8179	0.8750	0.5658	0.4827
Lithospermum latifolium	LITHLAT_9	orig+SOLRIS+cont (0.5)	8	29	243	0.8400	0.6350	0.5658	0.5862
Castanea dentata	castden_01	original(1)	7	276	250	0.8309	0.8571	0.8408	0.8913
Castanea dentata	castden_02	original(0.5)	7	276	250	0.8451	0.8571	0.8641	0.7681
Castanea dentata	castden_05	orig+SOLRIS(1)	7	276	239	0.8398	0.8571	0.8193	0.9058
Castanea dentata	castden_06	orig+SOLRIS(0.5)	7	276	239	0.8219	0.7143	0.8387	0.8080
Castanea dentata	castden_07	orig+cont(1)	7	276	250	0.8497	0.8571	0.8545	0.8986
Castanea dentata	castden_08	orig+cont(0.5)	7	276	250	0.8526	0.8571	0.8665	0.8986
Castanea dentata	castden_09	orig+SOLRIS+cont (1)	7	276	250	0.8417	0.8571	0.8488	0.9203
Castanea dentata	**castden_10**	orig+SOLRIS+cont (0.5)	**7**	**276**	**250**	**0.8651**	**0.8571**	**0.8457**	**0.8406**

Legend for environmental predictor variables column:

original(1) = original 14 variables only, regularization set to 1
original(0.5) = as above, regularization set to 0.5
orig+SOLRIS(1) = original 14 variables plus SOLRIS land use/land cover, regularization set to 1
orig+SOLRIS(0.5) = as above, regularization set to 0.5
orig+cont(1) = original 14 variables plus forest contiguity, regularization set to 1
orig+cont(0.5) = as above, regularization set to 0.5
orig+SOLRIS+cont (1) = original 14 variables plus SOLRIS land use/land cover + forest contiguity, regularization set to 1
orig+SOLRIS+cont (0.5) = as above, regularization set to 0.5

Appendix Table 2: Environmental predictors used in SDMs, and their source.

type	variable	description	source/reference	web
'original' set: topography, soil, geology, climate	elevation	elevation in meters	Canadian Digital Elevation Model	http://geogratis.gc.ca/
	slope	slope in degrees	Canadian Digital Elevation Model	http://geogratis.gc.ca/
	aspect	aspect converted to a linear variable using formula in Williams et al. 2009	Canadian Digital Elevation Model	http://geogratis.gc.ca/
	soil texture	texture of the majority soil type, e.g. "clay loam", categorical with 24 categories	Soil Survey Complex, Ontario Ministry of Agriculture	https://www.ontario.ca/page/land-information-ontario
	soil drainage	drainage of the majority soil type, e.g. "well-drained", categorical with 9 categories	Soil Survey Complex, Ontario Ministry of Agriculture	https://www.ontario.ca/page/land-information-ontario
	surficial geology	main type of surficial deposit, categorical with 40 categories	Surficial Geology of Southern Ontario, Ontario Geological Survey 2010	https://www.ontario.ca/page/land-information-ontario
	annual mean temperature	the average of the average monthly temperature (°C)	Canada Forest Service (McKenney et al. 2011)	https://cfs.nrcan.gc.ca/projects/3/2
	mean temperature of the growing season	average temperature during the growing season (°C)	Canada Forest Service (McKenney et al. 2011)	https://cfs.nrcan.gc.ca/projects/3/2
	isothermality	measure of how large the day-to-night temperature oscillation is in comparison to the summer-to-winter oscillation	Canada Forest Service (McKenney et al. 2011)	https://cfs.nrcan.gc.ca/projects/3/2
	mean temperature of the wettest quarter	the average temperature for the three months with the highest cumulative precipitation (°C)	Canada Forest Service (McKenney et al. 2011)	https://cfs.nrcan.gc.ca/projects/3/2
	annual precipitation	sum of all totally monthly precipitation (mm)	Canada Forest Service (McKenney et al. 2011)	https://cfs.nrcan.gc.ca/projects/3/2
	total precipitation for the growing season	sum of precipitation recorded during growing season (mm)	Canada Forest Service (McKenney et al. 2011)	https://cfs.nrcan.gc.ca/projects/3/2
	precipitation seasonality	variation in monthly precipitation over one year (%)	Canada Forest Service (McKenney et al. 2011)	https://cfs.nrcan.gc.ca/projects/3/2

	precipitation of the warmest quarter	total precipitation for the three warmest months (mm)	Canada Forest Service (McKenney et al. 2011)	https://cfs.nrcan.gc.ca/projects/3/2
land cover type	land use/land cover	land cover type, e.g. “deciduous forest”, categorical with 25 categories	Southern Ontario Land Resource Information System (MNRF)	https://www.ontario.ca/data/southern-ontario-land-resource-information-system-solris-20
landscape context	forest contiguity	calculated the number of cells within an 81 cell neighborhood including the focal cell that are >50% forested	Southern Ontario Land Resource Information System (MNRF)	https://www.ontario.ca/data/southern-ontario-land-resource-information-system-solris-20

Appendix Table 3: Mean number of presences detected in simulation using both random and informed sampling, with no filters for low-probability cells, and the true number of occurrences assigned to the landscape (*A. dracontium*).

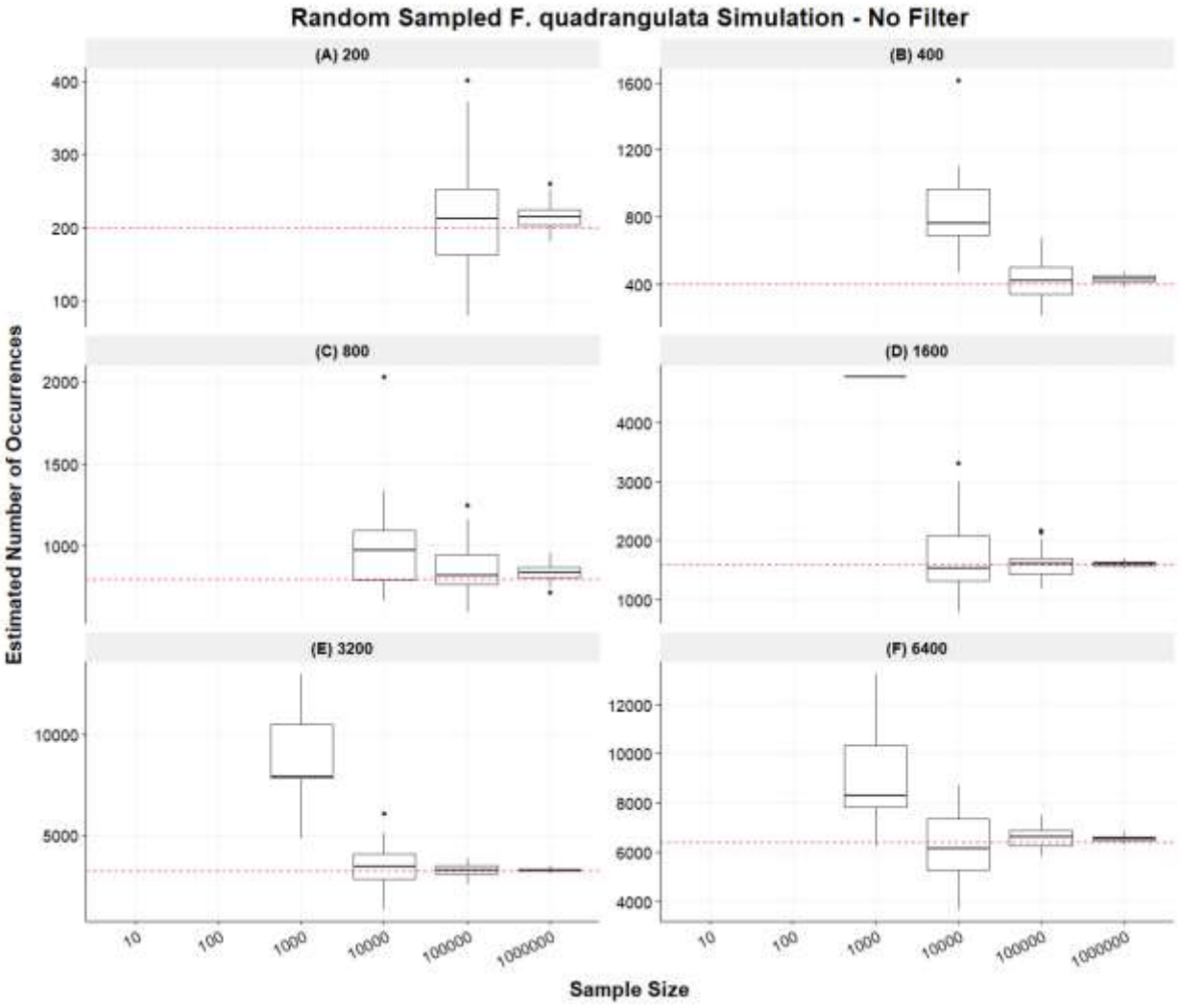
Calculation = ((# of true occurrences – mean # of presences detected)/# of true occurrences *100)

# of true occurrences	Sample size	Random sampling		Informed sampling	
		Mean # presences detected	Difference between true and detected (**% decrease** between detected and true)	Mean # presences detected	Difference between true and detected (**% decrease** between detected and true)
200	10	0	-100.00	0.02	-99.99
	100	0	-100.00	0.00	-100.00
	1,000	0	-100.00	0.02	-99.99
	10,000	0	-100.00	0.04	-99.98
	100,000	0.04	-99.98	0.08	-99.96
	1,000,000	0.04	-99.98	1.00	-99.50
400	10	0	-100.00	0.22	-99.95
	100	0.04	-99.99	0.34	-99.92
	1,000	0.1	-99.98	0.44	-99.89
	10,000	0.02	-100.00	0.72	-99.82
	100,000	0.02	-100.00	0.80	-99.80
	1,000,000	0.16	-99.96	1.54	-99.62
800	10	0.08	-99.99	1.76	-99.78
	100	0.18	-99.98	2.48	-99.69
	1,000	0.34	-99.96	3.46	-99.57
	10,000	0.54	-99.93	5.42	-99.32
	100,000	1.08	-99.87	8.52	-98.94
	1,000,000	2.4	-99.70	14.10	-98.24
1600	10	0.66	-99.96	13.72	-99.14
	100	1.6	-99.90	21.78	-98.64
	1,000	2.9	-99.82	34.04	-97.87
	10,000	6.2	-99.61	56.08	-96.50
	100,000	12	-99.25	79.50	-95.03
	1,000,000	25.06	-98.43	130.80	-91.83
3200	10	6.98	-99.78	83.52	-94.78
	100	15.5	-99.52	133.14	-91.68
	1,000	32.3	-98.99	214.84	-86.57
	10,000	65.9	-97.94	351.62	-78.02
	100,000	123.7	-96.13	534.90	-66.57
	1,000,000	249.3	-92.21	880.38	-44.98
6400	10	77.1	-98.80	147.06	-97.70
	100	156.14	-97.56	273.66	-95.72
	1,000	320.8	-94.99	514.58	-91.96
	10,000	662	-89.66	988.96	-84.55
	100,000	1239	-80.64	1753.26	-72.61
	1,000,000	2482	-61.22	3350.02	-47.66

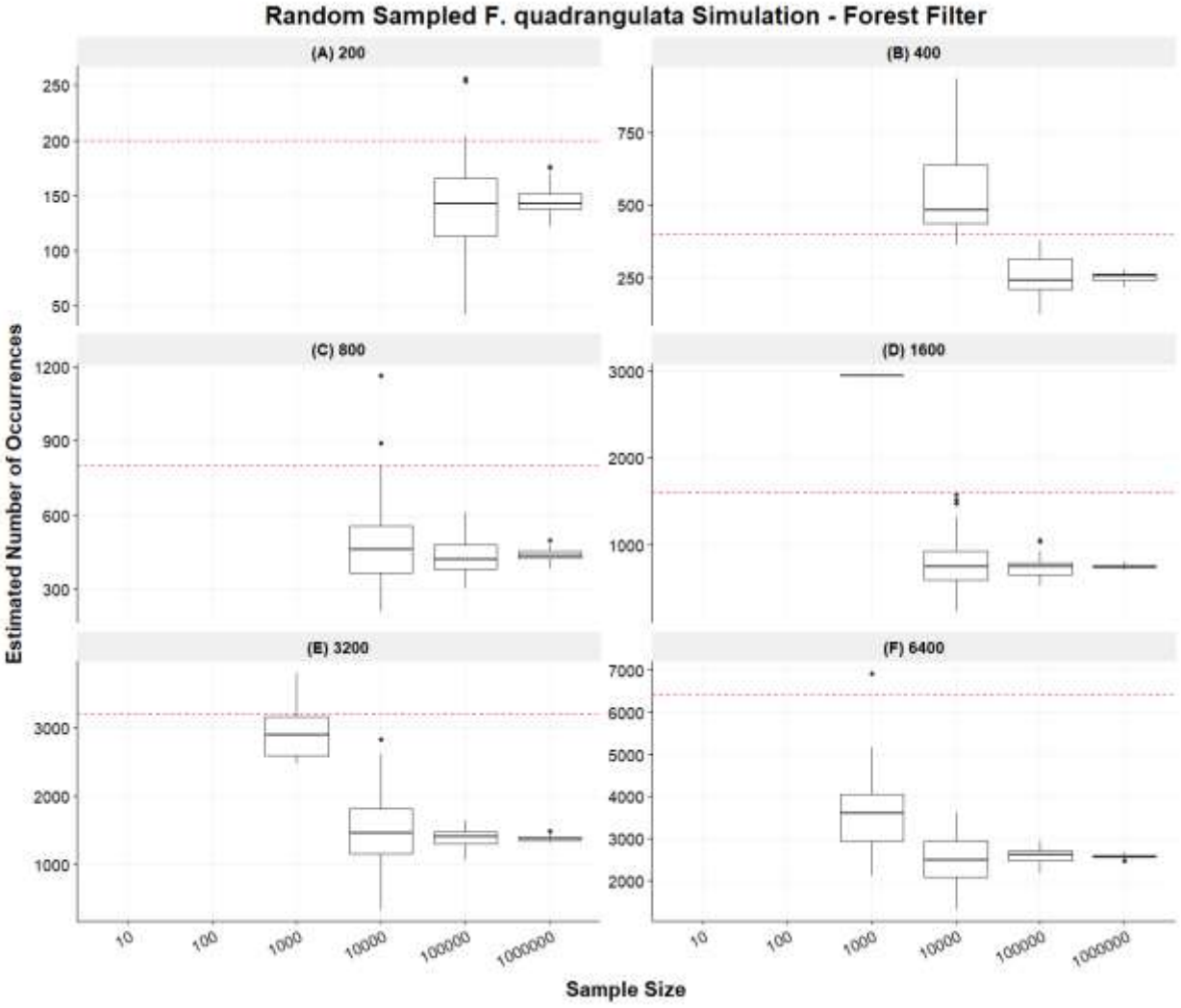
Appendix Table 4: Mean number of estimated occurrences in simulations using both random and informed sampling, with no filters for low-probability cells, and the true number of occurrences assigned to the landscape (*A. dracontium*).

Calculation = ((mean estimate of occurrence - # of true occurrences)/mean estimate of occurrence *100)

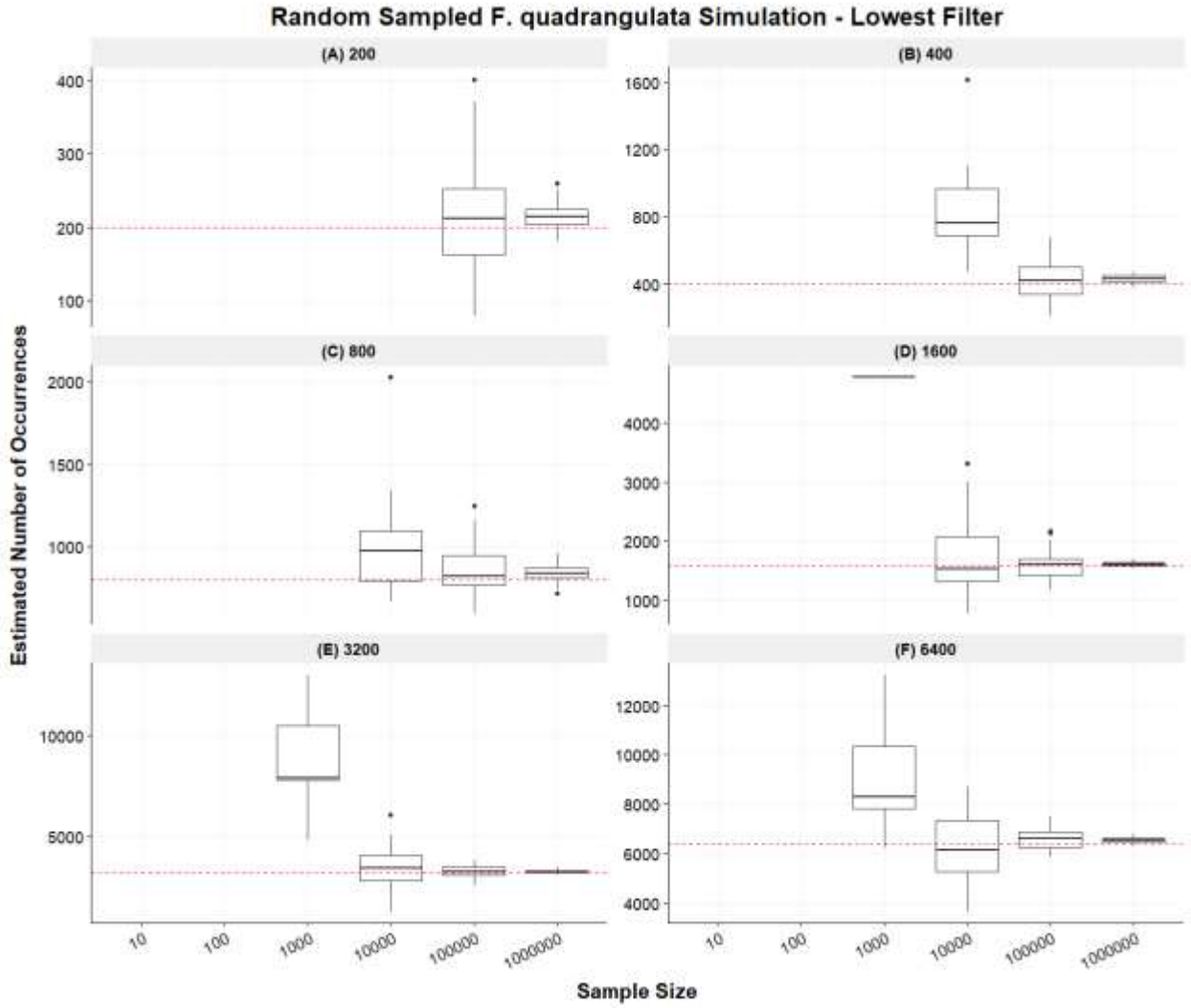
# of true occurrences	Sample size	Random sampling		Informed sampling	
		Mean estimate of occurrence	Difference between true and estimate **(% increase "+" or % decrease "-")**	Mean estimate of occurrence	Difference between true and estimate **(% increase "+" or % decrease "-")**
200	10	NA	NA	NA	NA
	100	NA	NA	NA	NA
	1000	NA	NA	654.00	69.42
	10,000	780.46	74.37	162.31	-18.85
	100,000	185.69	-7.16	170.35	-14.83
	1,000,000	202.73	1.35	188.00	-6.00
400	10	NA	NA	NA	NA
	100	NA	NA	NA	NA
	1000	NA	NA	769.04	47.99
	10,000	983.06	59.31	324.45	-18.89
	100,000	404.26	1.05	345.31	-13.67
	1,000,000	409.39	2.29	391.05	-2.24
800	10	NA	NA	NA	NA
	100	NA	NA	NA	NA
	1000	NA	NA	855.72	6.51
	10,000	1079.17	25.87	725.97	-9.25
	100,000	842.27	5.02	739.07	-7.62
	1,000,000	841.94	4.98	805.89	0.74
1600	10	NA	NA	NA	NA
	100	NA	NA	16745.45	90.45
	1000	7905.42	79.76	2075.94	22.93
	10,000	1677.72	4.63	1567.97	-2.00
	100,000	1723.48	7.16	2845.00	43.76
	1,000,000	1735.45	7.80	1688.98	5.27
3200	10	NA	NA	NA	NA
	100	NA	NA	3739.93	14.44
	1,000	9066.28	64.70	3220.78	0.65
	10,000	3142.47	-1.80	2845.32	-11.08
	100,000	3239.73	1.23	3073.53	-3.95
	1,000,000	3251.25	1.58	3202.82	0.09
6400	10	NA	NA	NA	NA
	100	NA	NA	14782.88	56.71
	1,000	10529.36	39.22	6601.01	3.05
	10,000	6559.32	2.43	6605.59	3.11
	100,000	6537.18	2.10	6254.13	-2.28
	1,000,000	6512.31	1.72	6442.34	0.66



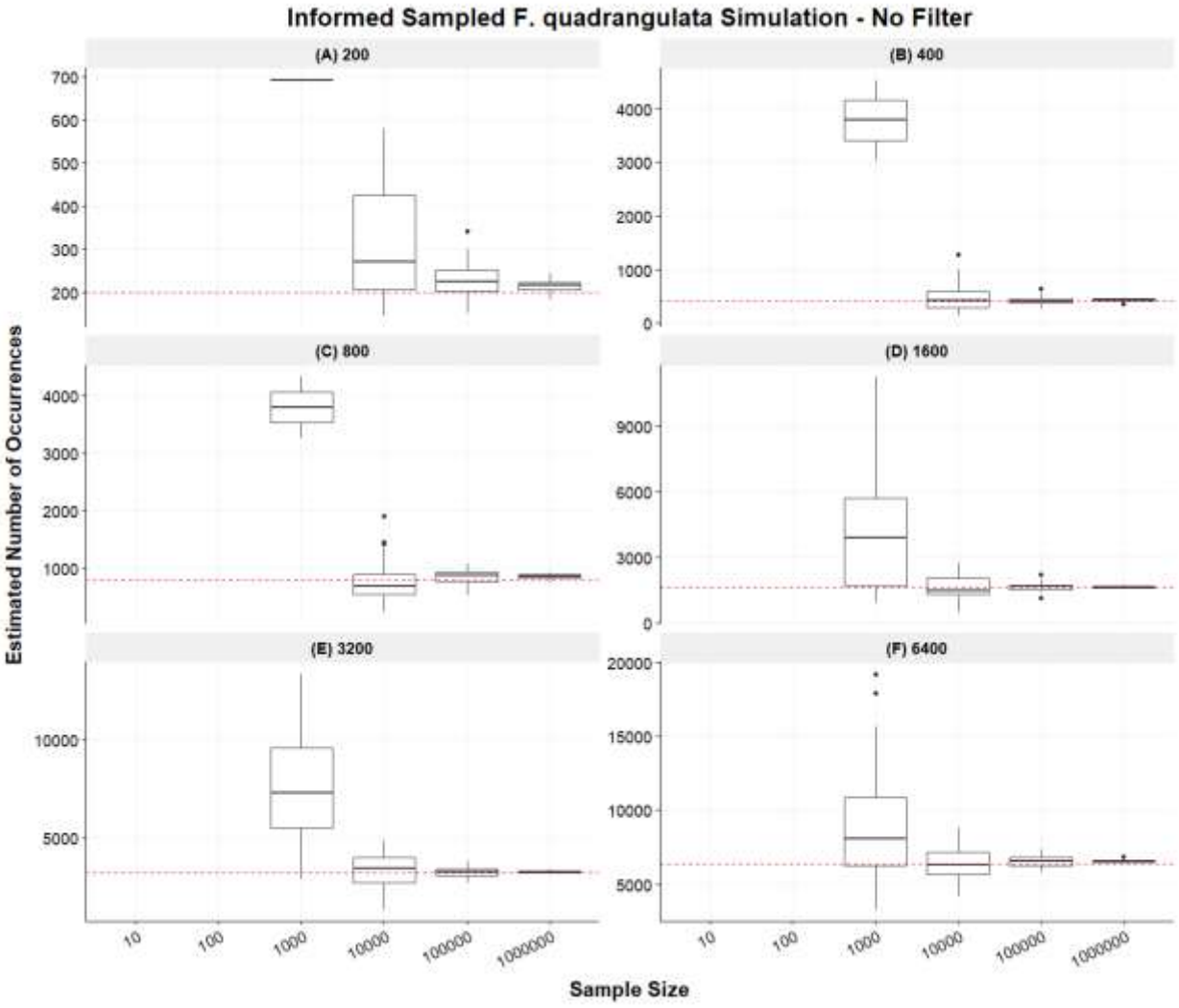
Appendix Figure 14: Random sampling of simulated presences (200-6400) assigned to the *F. quadrangulata* SDM (landscape). 50 iterations at each sample size (10 – 1,000,000). No filter was used to obtain these occurrence estimates. The horizontal dashed red is the known number of presences assigned to the landscape.



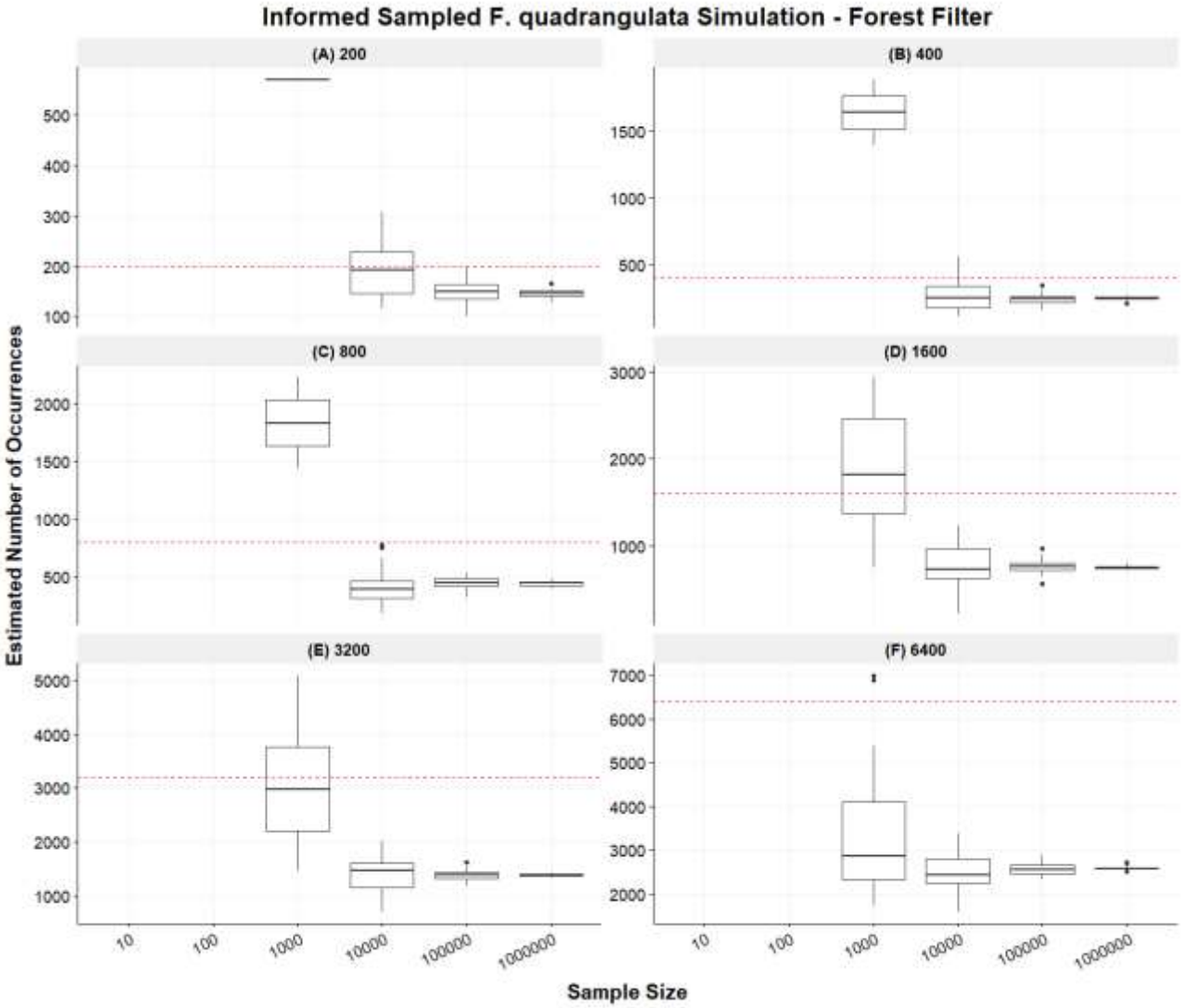
Appendix Figure 15: Random sampling of simulated presences (200-6400) assigned to the *F. quadrangulata* SDM (landscape). 50 iterations at each sample size (10 – 1,000,000). Non-forested cells were removed from occurrence estimates. The horizontal dashed red is the known number of presences assigned to the landscape.



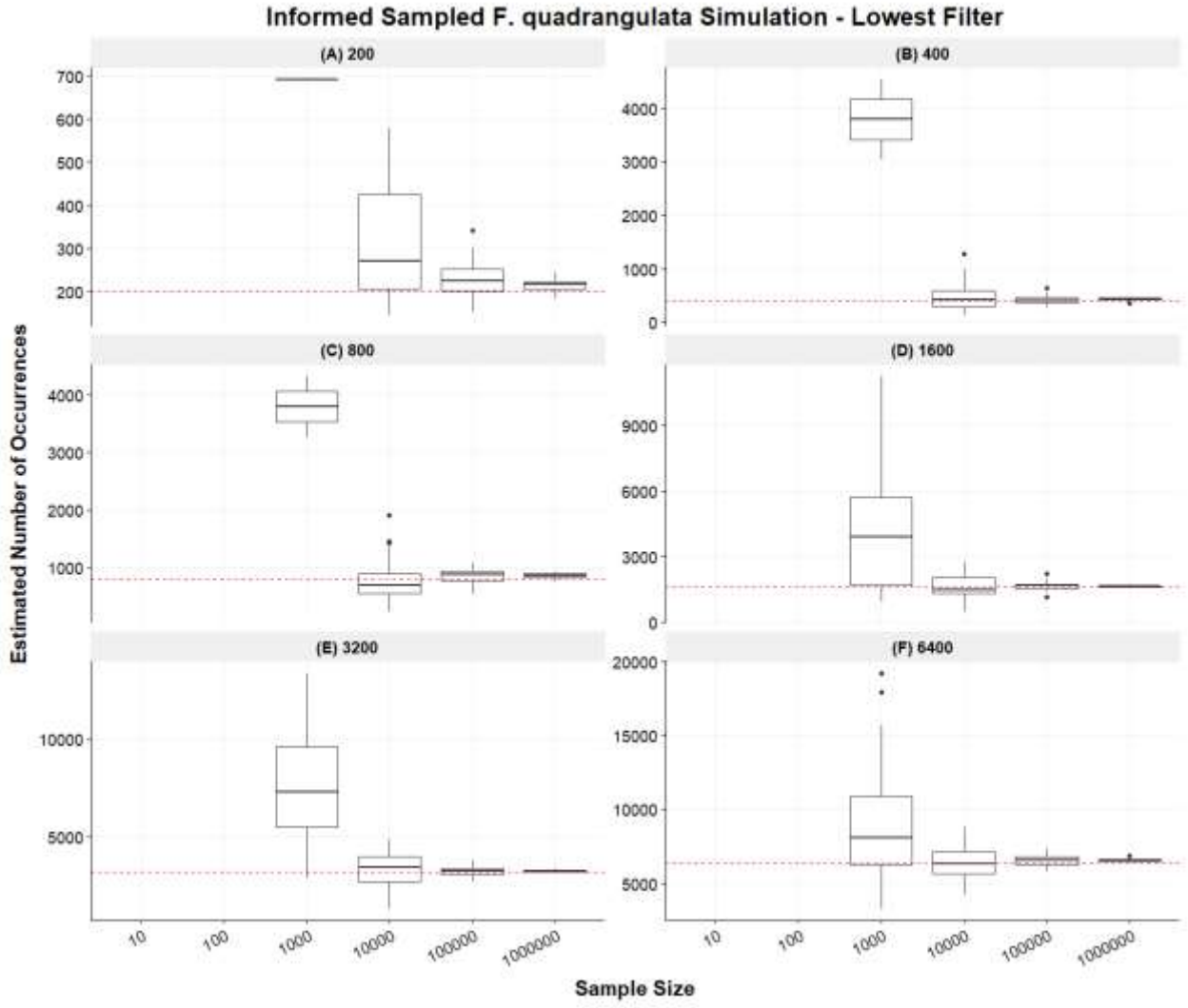
Appendix Figure 16: Random sampling of simulated presences (200-6400) of *F. quadrangulata* SDM, with 50 iterations at each sample size (10 – 1,000,000). The cells with a probability of containing a presence lower than 0.02 were removed from occurrence estimates. The horizontal dashed red line is the known number of presences assigned to the landscape.



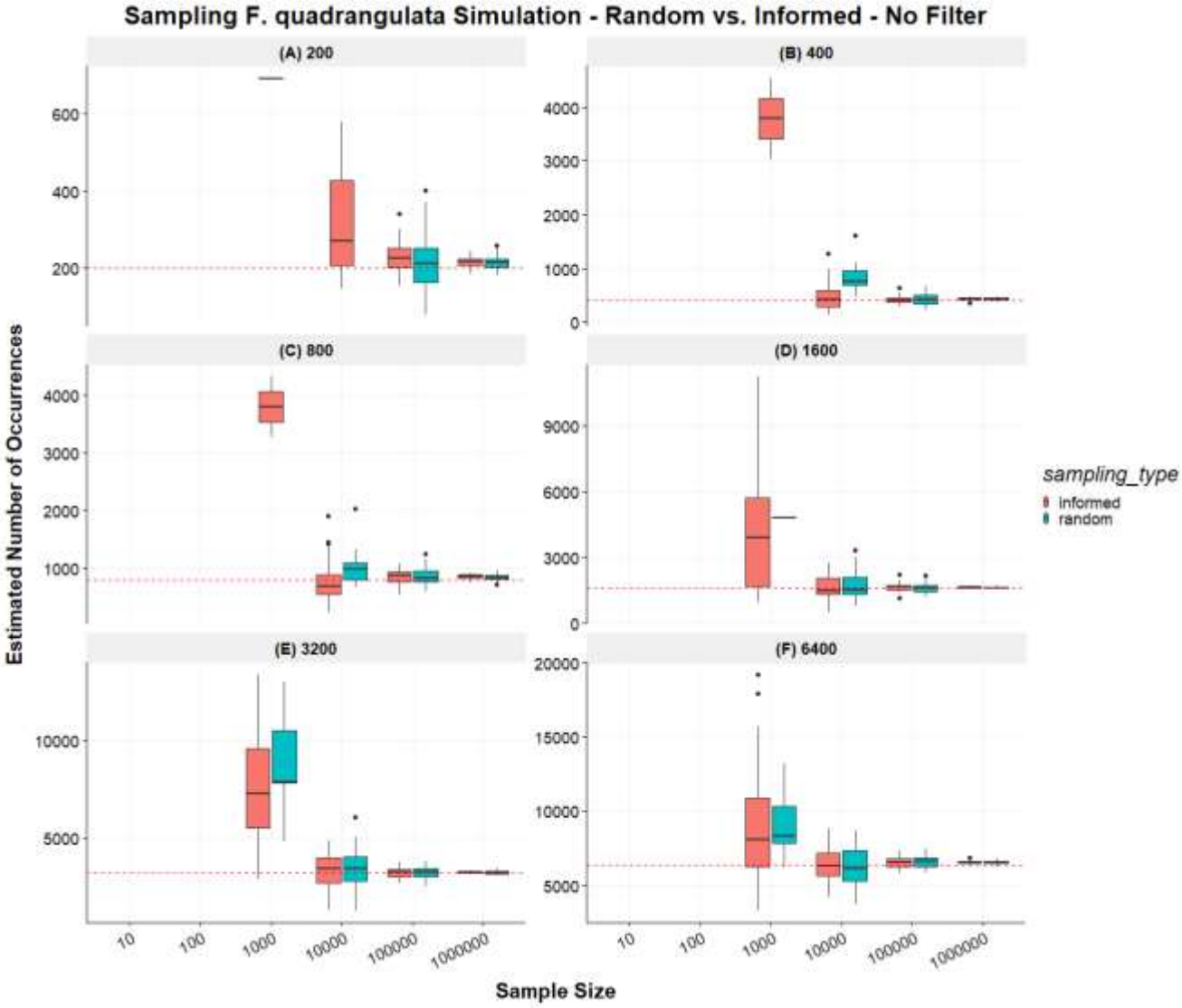
Appendix Figure 17: Informed sampling of simulated presences (200-6400) assigned to the *F. quadrangulata* SDM (landscape). 50 iterations at each sample size (10 – 1,000,000). No filter was used to obtain these occurrence estimates. The horizontal dashed red is the known number of presences assigned to the landscape.



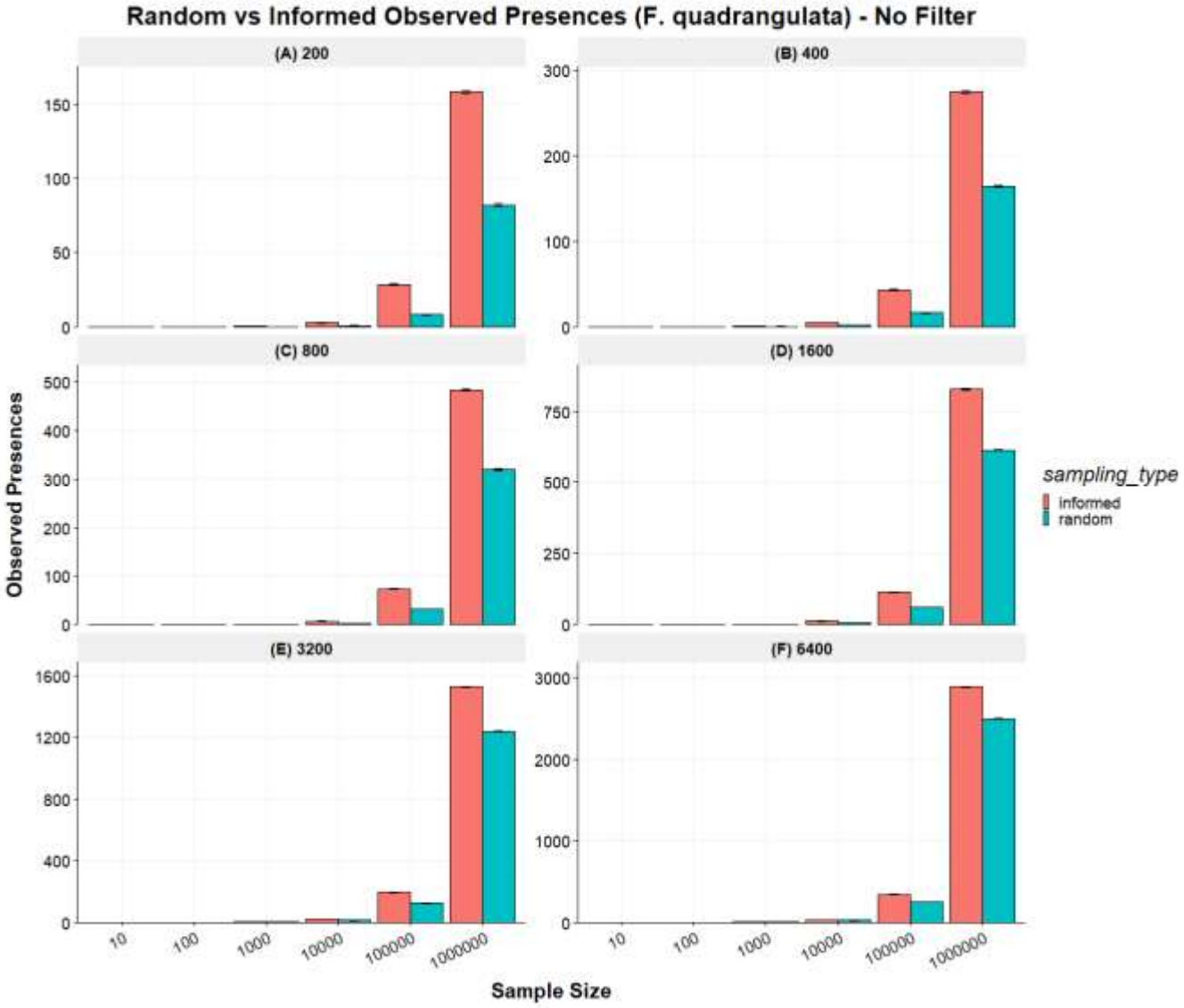
Appendix Figure 18: Informed sampling of simulated presences (200-6400) assigned to the *F. quadrangulata* SDM (landscape). 50 iterations at each sample size (10 – 1,000,000). Non-forested cells were removed from occurrence estimates. The horizontal dashed red is the known number of presences assigned to the landscape.



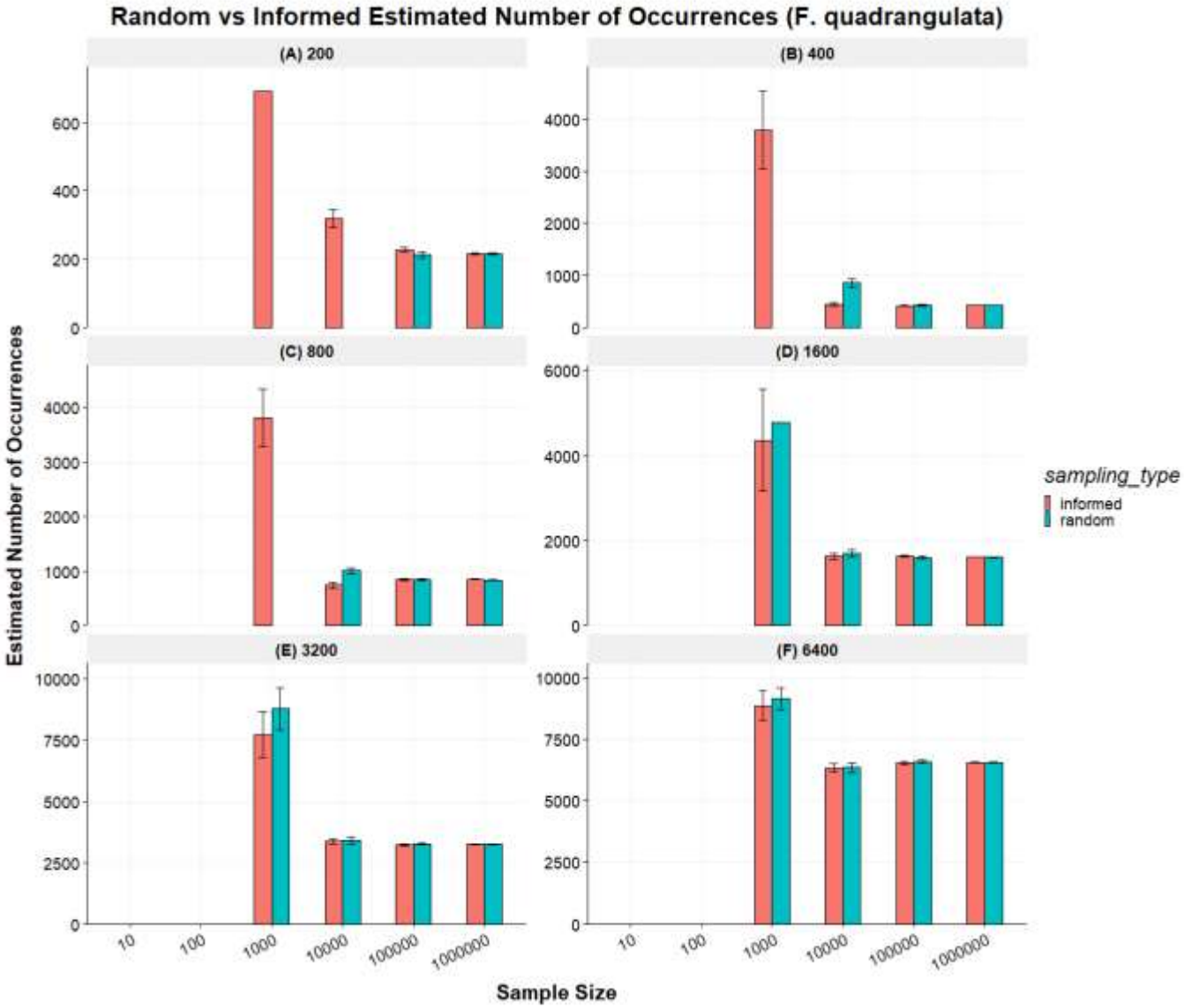
Appendix Figure 19: Informed sampling of simulated presences (200-6400) of *F. quadrangulata*, with 50 iterations at each sample size (10 – 1,000,000). The cells with a probability of containing a presence lower than 0.02 were removed from occurrence estimates. The horizontal dashed red line is the known number of presences assigned to the landscape.



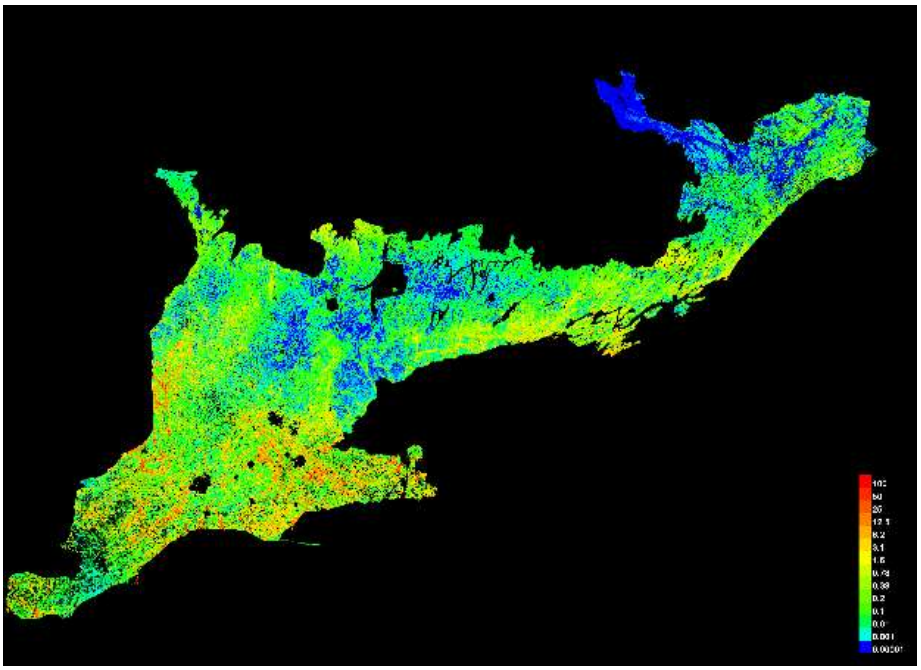
Appendix Figure 20: Random (blue) and informed (red) sampling of simulated presences (200 – 6400) of *F. quadrangulata, w*ith 50 sampling iterations at each sample size (10 – 1,000,000). No filter was used to obtain these occurrence estimates. The horizontal dashed red line is the known number of presences assigned to the landscape.



Appendix Figure 21: Panels A – F delineate the simulations by the number of presences assigned the landscape shown next to each letter. The mean number of observed presences for both informed (blue) and random (red) sampling techniques are depicted along the y-axis. No filter was used to obtain these occurrence estimates. The horizontal dashed red line is the known number of presences assigned to the landscape.



Appendix Figure 22: Panels A - F show the mean estimated number of occurrences at each number of presences, informed vs random sampling. No filter was used to obtain these estimates of occurrence.



Appendix Figure 23: *F. quadrangulata* SDM used as landscape for simulation. Cropped using the same dimensions as for *A. dracontium* to contain 2.6 million cells. The SDM is depicted as a heat map showing suitable sites in red/orange and less suitable sites in blue.

Appendix 24: R code used for simulations

```
rm(list=ls())


library(cowplot)
library(dplyr)
library(ggplot2)
library(reshape)
library(scales)
library(viridis)
library(visreg)
library(parallel)
library(doParallel)
library(foreach)
```

```
library(iterators)
library(sp)
library(raster)

# load in SDM raster
SDM<-raster("~/R/projects/landscape_simulation/Arisdra_simulation/ARISDRA_14.asc")
str(SDM)
names(SDM)

dim(SDM$layer)
# Crop the raster to make it a more manageable size
SDM = crop(SDM, extent(300000, 700000, 4600000, 4850000))
plot(SDM)
# need to work out what a sensible row and column to crop at are. you could use dim(SDM$layer),
# make sure you use plot(SDM) to view the results to make sure the crop is where you want it.
# However you crop the SDM you will have to crop forest in the same way so they match up

# load in pres/abs records
survey<-
read.csv("~/R/projects/landscape_simulation/Arisdra_simulation/ARISDRA_v14_sim_records.csv")
head(survey)

names(survey)[names(survey)=="arisdra14"]<-"ARISDRA_14"
head(survey)
names(SDM)[names(SDM)=="layer"]<-"ARISDRA_14"
names(SDM)
dim(SDM$ARISDRA_14)

# load in binary forest raster
```

```
forest<-raster("~/R/projects/landscape_simulation/Arisdra_simulation/forestnow.asc")

dim(forest)
forest = crop(forest, extent(300000, 700000, 4600000, 4850000))
dim(forest)
plot(forest)

# Create a glm to relate the presabs to the habitat suitability (ARISDRA_14) as the explanatory factor
glm1<-glm(presabs ~ ARISDRA_14, family="binomial",
        contrasts = c("contr.sum", "contr.poly"),
        data = survey)
summary(glm1)



glm_plot <- ggplot(data = survey, aes(y = presabs, x = ARISDRA_14)) +
  geom_smooth(method = "glm",
            method.args = list(family = binomial(link = "logit"))) +
  geom_point(size = 2.5)

glm_plot

# now create a new raster by predicting the probability of occurrence in each cell
# based on glm1:
# predict from raster isn't properly predicting values from the glm to the raster ----------
# need to run example in help for predict from raster package

prob_map_rast<- predict(SDM, glm1, fun=predict, na.rm=TRUE, progress="text"
                , type = "response"
                )
```

```
str(prob_map_rast)
min(prob_map_rast)
max(prob_map_rast)

names(prob_map_rast)
names(prob_map_rast)<-"ARISDRA_14" ### ### ###
dim(prob_map_rast)

# Convert prob_map would have to a matrix in order to work
# as an argument in functions later on in the code
# convert raster to matrix and view
### ### ### Also remove NAs - they were causing trouble
prob_map <- as.matrix(prob_map_rast)
image(prob_map)
## Removed these, not needed and pretty sure this is causing the non-conformable array problem
tt<-as.data.frame(prob_map_rast)
ttt<-na.omit(tt)
prob_map<-as.matrix(ttt)
dim(prob_map) # same as SDM
# melt into vector
prob_map_melt <- melt(prob_map)
dim(prob_map_melt)
names(prob_map_melt)[names(prob_map_melt)=="value"]<- "ARISDRA_14"
names(prob_map_melt)
# convert forest to matrix then multiply by prob_map to get forest filter
forest <- as.matrix(forest)
```

```
# Now make the forest filter, for this to work the forest matrix has to be the same dimensions as the
#prob_map
forest_melt = melt(forest) # should be a vector the same length as melt(prob_map)
dim(forest_melt)
names(forest_melt)

#convert SDM from raster to matrix first then to vector using melt()
SDM<-as.matrix(SDM)
SDM_melt<-melt(SDM)       ### ### ###
names(SDM_melt)[names(SDM_melt)=="value"]<-"maxent_output"  ### ### ###
names(SDM_melt)

# remove the NA from SDM and prob_map and all filters
NA_element = is.na(SDM_melt$maxent_output)
SDM_NAomit = filter(SDM_melt, !NA_element)
forest_filt = filter(forest_melt, !NA_element)
prob_map = filter(prob_map_melt, !NA_element)
  mutate(prob_map, ARISDRA_14 = ARISDRA_14 * forest_filt$value)
samp_cutoff = 0.01580523
low_filt = mutate(prob_map, filt = ifelse(ARISDRA_14 <samp_cutoff, 0, 1))
#forest_filt is now a single binary string of 1s and 0s
dim(SDM_NAomit)
dim(forest_filt)
dim(prob_map)

# quick check the X1 and X2 (lat and long) values line up,
# also confirm no NA
summary(SDM_NAomit)
summary(prob_map)
```

```
summary(forest_filt)

### ### ### Remove NAs from SDM as well at this point
SDM <- SDM_NAomit

# Assign presences to landscape (prob_map)

# Assuming that we arrived at the current number of presences (n_pres) through
# a process of random extinctions

make_spp_dist_from_data <- function(SDM, prob_map, n_pres){
 #replaced hab_suit_scale with prob_map

 # make a draw from a bernolli distribution for each cell
 pres_map <- apply(prob_map, MARGIN = c(1, 2), FUN = function(p){
  return(rbinom(1, size = 1, prob = p))
 })

 # if the random draw has generated too many presences. The pres_map will generally have around
 # 0.1 - 0.25 occupied, which will work well for our purposes.
 while(sum(pres_map, na.rm = TRUE) > (n_pres * 1.1)){

  # extinction process, assume the probability of extinction is the inverse of probability of
  # occurrence in cell, conditional on the cell being occupied, need to be the case in reality,
  # but a plausible model
  ext_map <- apply((1 - prob_map) * pres_map, MARGIN = c(1, 2), FUN = function(p){
   return(rbinom(1, size = 1, prob = p))
  })
```

```
    # randomly choose sum(pres_map) - n_pres extinctions to actually occur
    ext_inds <- base::sample(which(ext_map == 1),
                             size = sum(pres_map) - n_pres, replace = TRUE)

    pres_map[ext_inds] <- 0

  }

  return(list(maxent_output = SDM,
              prob_map = prob_map,
              pres_ab_map = pres_map))

}


## Populate Landscape simulation ----

n_samp = 10 # set back to c(100) when error fixed
max_samp = 100000 #10000 # now that the landscape is bigger we may want to sample more?
samp_mult = 10 ### ### ### also we want to have an integer so get an integer sample size
max_samp_reached = FALSE
count = 1
while(!max_samp_reached){
  n_samp[count + 1] = n_samp[count]  * samp_mult
  count = count + 1
  if(n_samp[count] > max_samp) max_samp_reached = TRUE
}

n_sim <- 50 # number of 'samples' at each individual samples size (set to 5 until code tested)
```

```
n_pres <- c(200,400,800,1600,3200,6400) #  of true presences

# This is a more robust way to build the dataframe that will not break when you
# change the loop structure so long as count is incremented in the inner most loop
# This will also make it a bit easier to do in parallel.
# I also added in some print out markers so you can see which bits are running and how long each bit
# takes. I cleaned up the indenting, so the code is more readable and it is more obvious the loop or if
# statement each line of code belongs to.

# do this multi-threaded to seed up, there are a few ways to do this in R I this one makes the
# fewest changes to the code

# This sets up the parallel
cl = makeCluster(15) # will use 4 cores, change number to use more
registerDoParallel(cl)

t1 = Sys.time()

# do the simulations in parallel. replace %dopar% with %do% to run on single thread
# for testing. The foreach parallel interface requires that you tell it all the
# external libraies used in the {} that are run in parallel in this case 'scales' and 'reshape' and 'dplyr'

# clears log file that tracks progress
writeLines(c(""), "log.txt")

res_list = foreach(i = 1:n_sim, .packages = c("scales", "reshape", "dplyr")) %dopar% {#this is the loop
where we take samples a bunch of times

  sink("log.txt", append=TRUE)
```

```
 cat(paste("Starting sim: ", i, "\n"))
 sink()


 count = 1
 res_df = list()


 for(np in n_pres){


   # make a new landscape for each sim for each number of presences
   spp_dist_ob <- make_spp_dist_from_data(SDM_NAomit$maxent_output,
as.matrix(prob_map$ARISDRA_14) ,n_pres = np) # use the coloumn not the whole dataframe, then
convert to matrix


   allvars <- cbind(melt(spp_dist_ob$maxent_output), melt(spp_dist_ob$prob_map),
                    melt(spp_dist_ob$pres_ab_map))


   names(allvars)[1] <- "maxent_output"
   names(allvars)[4] <- "prob_map"
   names(allvars)[7] <- "pres"



   for(ns in n_samp) {
    # To save the results in a data.frame
    res_df[[count]] = data.frame(sample.size = ns,
                                 sim_id = i,
                                 n_pres = np,
                                 obs.pres  = NA,
                                 pred.pres = NA,
                                 pred.pres.forest.filter = NA,
```

```
                    pred.pres.lowest.filter = NA
  )

  #print(paste0(count, ': Sampling landscape'))

# RANDOM SAMPLE
  sampled.hab <- allvars[sample.int(nrow(allvars), ns), ]
# run code to the end using the above sample function

# INFORMED SAMPLE
# run entire code using the line below to replace the random sample line above. This line uses the
# prob_map to preferentially sample cells with a higher probability of presence for the species in
# question.
  # sampled.hab <- allvars[sample.int(nrow(allvars), ns, prob = allvars$prob_map), ]


  obs.pres <- sum(sampled.hab$pres) ### this counts up our observed presences in our sampled # cells

  res_df[[count]]$obs.pres <- sum(sampled.hab$pres)
      ###this counts up our observed presences in our sampled # cells

  if(obs.pres < 3) {res_df[[count]]$pred.pres = NA} else {
   #I added this loop so that if fewer than 3 presences were counted,
   #we could not do a GLM - 3 is arbitrary and probably low

   ## Build GLM and predict() ----

   #print(paste0(count, ':  fit and predict from GLM'))
```

```
      test1.sample <- glm(pres ~ maxent_output, data = sampled.hab, family = "binomial")
      #the GLM

      test2.sample<-predict(test1.sample, allvars, type = "response")   ### ### ### There was an error here -  no object called maxent_output so this is the fix
      ### note here we predict across whole dataset with sampled GLM,
      ### each cell gets a probability

      res_df[[count]]$pred.pres <- sum(test2.sample)
      res_df[[count]]$pred.pres.forest.filter <- sum(test2.sample[forest_filt$value == 1])
      res_df[[count]]$pred.pres.lowest.filter <- sum(test2.sample[low_filt$filt == 1])
      ###this is where we sum up the predicted # presences using the GLM


    }  #end of if-else statement

    #print(paste0(count, ':   Finished sim'))

    count = count + 1

  }  #end n_samp loop
 } # end n_pres loop

 # return the consolidated data frame
 bind_rows(res_df)

} # end sim loop
```

```
# free the workers
stopCluster(cl)

t2 = Sys.time()

#time taken
t2 - t1

result_df <- bind_rows(res_list)

Summary_out <- result_df %>%
  group_by(sample.size, n_pres) %>%
  summarise(n_obs = n(),
            num_na = sum(is.na(pred.pres)))

# Write output files to csv
setwd("~/R/projects/landscape_simulation/R_outputs_csv_files")
write.csv(result_df, file = "result_df_arisdra_filtered_Nov5B.csv")
write.csv(Summary_out, file = "Summary_out_arisdra_filtered_Nov5B.csv")

# To test and confirm that the simulation did not assign presences to cells with
# no forest

prob_map <- filter(prob_map, ARISDRA_14 = ARISRDA_14 *forestfilt$value > 0)
non_forest <- filter(prob_map, ARISDRA_14 = ARISDRA_14 * forestfilt$value == 0)

forest_test = filter(allvars[,c('prob_map', 'pred')])
non_forest_test = filter(allvars, prob_map == 0)
```