Using species distribution models to inform IUCN Red List assessments

Mindy M. Syfert[a,b,c], Lucas Joppa[c], Matthew J. Smith[c], David A. Coomes[a], Steven P. Bachman[d], Neil A. Brummitt[b]

[a]Forest Ecology and Conservation Group, Department of Plant Sciences, University of Cambridge, Cambridge, CB2 3EA, UK
[b] Department of Life Sciences, The Natural History Museum, London, SW7 5BD, UK
[c] Computational Science Laboratory, Microsoft Research, CB1 2FB, UK
[d] Royal Botanic Gardens, Kew, Richmond, Surrey TW9 3AB, UK

Corresponding author and present address: mmsyfert@gmail.com; The Natural History Museum, London, SW7 5BD, UK

Email addresses: s.bachman@kew.org; n.brummitt@nhm.ac.uk; dac18@cam.ac.uk, lujoppa@microsoft.com; matthew.smith@microsoft.com

ABSTRACT

Characterising a species' geographical extent is central to many conservation assessments, including those of the IUCN Red List of Threatened Species. The IUCN recommends that extent of occurrence (EOO) to be quantified by drawing a minimum convex polygon (MCP) around known or inferred presence localities. EOO calculated from verified specimens is commonly used in Red List assessments when other data are scarce, as is the case for many threatened plant species. Yet rarely do these estimates incorporate inferred localities from species distribution models (SDMs). A key impediment stems from uncertainty about how SDM predictions relate to EOO. Here we address this issue by comparing the EOOs estimated from specimen localities with EOOs derived from SDMs for plant species occurring in Costa Rica and Panama. We first analyse 20 plant species, with well-known and well-sampled distributions, and train SDMs to subsamples of the data and assess how well the SDM-derived MCPs predict both the MCPs of the subsamples and the MCPs of the complete dataset. We find that when sample sizes are small (5 or 10 samples) the SDM-derived MCPs are actually closer to the complete dataset than to the MCPs of the subsamples, both in terms of EOO and geographically. This occurs when using a probability threshold based on maximum geographical similarity between the SDM-derived MCP and the subsample MCP; other threshold methods performed less well. For the species with less well-known distributions, the SDM-derived EOOs correlate strongly with, but tend to be larger than, EOOs estimated by point data. This implies that a SDM-derived EOO may be more representative of the full EOO than that drawn around known localities. Our findings reveal situations in which SDMs provide useful information that complements the IUCN Red Listing process.

1 INTRODUCTION

The International Union for Conservation of Nature (IUCN) Red List Categories are internationally recognised as the standard for assessments of species extinction risk (Butchart et al. 2005; Mace et al. 2008) and are instrumental in analyses of biodiversity change (Baillie et al. 2008; Butchart et al. 2004; Rodrigues et al. 2006). The IUCN sets formal criteria for Red List assessments, based on a considerable body of population theory (Mace et al. 2008), to standardise them across diverse taxa (criteria A-E; IUCN 2013).

The extent to which risks from threatening factors are spread geographically is a central component of assessing extinction risk (Purvis et al. 2000) and, thus, of Red List assessments - with 'extent of occurrence' (EOO) being one widely accepted measure (Gaston 1991). The EOO is defined by IUCN (2013) as the area that lies within the outermost limits of known or inferred locations. Importantly, EOO is not intended to be an estimate of the amount of occupied or potential habitat nor a general measure of a taxon's range (IUCN 2013); instead it measures the overall geographic spread of the localities at which the species is found (Gaston and Fuller 2009). EOO is most often measured as a minimum convex polygon (MCP) around the known species locations (MCP; IUCN 2013). Although MCP is the standard method for estimating EOO, and the one used in this study, it is one of several possible methods, each of which present their own particular strengths and biases. For example, the alpha-hull has been suggested as a more appropiate measure when a species has a disjunct or concave distribution, or when estimating trends in species ranges (Burgman and Fox 2003). Interpreting the EOO, however quantified, is relatively straightforward when there is confidence that the known locations of a species represent its full geographical spread. This is what we term as a "well sampled" species in this paper. Most species, however, are not well sampled, meaning that the MCP drawn around the known locations may represent only a proportion of its geographic range. For these, it seems reasonable to attempt to infer the species range to estimate the EOO. The use of inferred ranges for calculating EOO is explicitly stated in the IUCN Red List Guidelines: "sites can be inferred from presence of known appropriate habitat, but where the species has not yet been searched for." (IUCN 2013, pg 35). Inferring extinction risk in Red List assessments is commonly conducted using Population Viability Analysis (PVA), but usually relies on posessing sufficient information about the species in question in order to make accurate estimates (Brook et al. 2000; Coulson et al. 2001). When species are poorly sampled there is usually insufficient information to conduct such analyses. This raises the question whether other inference methods can be usefully applied to predict other components of Red List assessments, such as EOO, for poorly sampled species.

Species distribution models (SDMs; also known as bioclimatic models or ecological niche models) have become the most popular method for inferring distributions from observational data. Recent studies have used SDMs to estimate species ranges and occupied areas for the purposes of informing IUCN Red List assessments (e.g. Cardoso et al. 2011; Jiménez-Alfaro et al. 2012; Papes and Gaubert 2007; Pena et al. 2014; Sergio et al. 2007). These focus on small numbers of plant species within national boundaries. However, to date it has been unclear whether SDMs can make informative estimates of the EOO of a species when it is poorly sampled. Analyses of a larger number of species across a wider geographic area are needed in order to address how informative SDM predictions might be for the Red Listing process (Brummitt et al. 2008).

How could SDM predictions inform EOO estimates for poorly sampled species? A traditional approach to predicting a species' distribution would involve training SDMs by correlating species presences and absences with environmental variables and then using those correlations (the model) to predict the probability of a species presence across a landscape. For many species, like all of those considered in this paper, there only exists data on presences, with data on recorded absences not available. For such species SDMs can only predict the relative probability of occurrence across a landscape (see Merow et al 2013 for details). A number of important considerations should be taken before such an approach can be considered suitable for predicting the EOO. Correlative models are typically applied assuming that the species-environment relationships are not likely to dramatically change over the time frame of interest (i.e. the species distribution has reached some form of equilibrium), and this is will not be valid in all cases. Environmental variables should also be carefully chosen such that the predicted distribution relates to likely occupied areas rather than areas potentially suitable. It is widely recognised that characteristics of presence/absence data, such as sampling bias, the specific set of environmental variables chosen, and the modelling technique used all influence the accuracy of the resulting SDM predictions (Elith and Leathwick 2009).

After a fitted SDM has been judged to be of sufficient quality, there remains the decision about how to determine an appropriate threshold probability of occurrence to discriminate between predicted presence and absence locations and so estimate the size of the geographic area occupied by the species. Traditionally, threshold probabilities are chosen in a way that maximizes the discrimination of sites with known presences from the rest of the landscape, or from locations with known absences (Franklin 2009; Liu et al. 2005; Liu et al. 2013). The studies of Liu et al. (2005) and Liu et al. (2013) performed systematic assessments of the consequences and value of different methods for choosing threshold probabilities. For example, in their study of thresholding methods for SDMs built using presence-only data, Liu et al. (2013) found that the commonly applied approach of maximising the sum of the sensitivity and specificity of the SDM resulted in predicted distributions that were more accurate than alternative methods when assessed using a number of different performance metrics. However, a consequence of using such methods to predict species distributions is that they typically result in a species being predicted as absent in some locations where it is known to occur. An approach to remedy this is to opt for the minimum training presence threshold (MTP; also termed lowest presence threshold; Pearson et al. (2007) that predicts all observations of a species as present (Pearson et al. 2007; Thorn et al. 2009). However, this can lead to predicted presences occurring over a much wider geographic range of localities than occur in reality (Bean et al. 2012) – an unhelpful over-prediction when IUCN Red List assessments should follow the precautionary principle and assign the most threatened category plausible (pg 35, IUCN 2013), although this over-prediction proved helpful in the study of Pearson et al. (2007), who used them to identify unknown populations or unknown species.

In this study, we take a different approach. Instead of choosing a threshold based on a model's discriminatory ability (the standard approach), we determine it by maximizing – across all probability thresholds – the geographical similarity between the MCP drawn

around the known presence localities and an MCP drawn around presences predicted by the SDM (SDM-derived MCPs). Initial investigations using small subsamples from a well-sampled species revealed that the EOOs of the resulting SDM-derived MCPs were consistently close to that drawn around all of the samples (rather than just the subsamples). We therefore tested our method on 20 plant species for which the EOO is considered well known, due to extensive sampling. This allows us to implement the approach on subsamples of the known occurrence data for each species (as low as just five known presence localities, a situation common in threatened species) and assess the extent to which the SDM-derived MCPs matched the MCPs of the complete dataset and the extent to which the EOOs of those MCPs were similar. Remarkably, we find that, for the majority of our species, the SDM-derived MCPs are actually closer to the MCPs measured from the complete dataset (the "true" range – noting that it is almost never known perfectly) than they are to the MCPs measured from the subsets of data used for model training, and this consequently results in better estimates of the EOO, especially when sample sizes are small. This indicates that, for these taxa in this geographical area, the SDM-derived MCPs are often more informative for estimating EOO for a Red List assessment than the available data alone are when sample sizes are small. This is the main novel finding from our work. We then extend these findings to 30 less well-known species on the IUCN Red List, and assess how the inferred EOOs could influence existing Red List assessments for these species.

2 DATA AND METHODS

2.1 Occurrence data

Our occurrence datasets comprise presence-only specimen data from the Royal Botanic Gardens, Kew (RBG Kew) and the Natural History Museum, London (NHM), supplemented by additional online specimen data, collated over the period 2006-2011. All occurrence data were for plant species from Central America. We identified 20 species for which their distributions can be said to be both well-known and well represented by a relatively large number (between 54 and 246) of recorded presence localities, from a variety of taxonomic families and life forms. We refer to these as the "well-known species" (see Appendix 1 for details). Distributions of species in this area are known to change their distribution extents over interglacial cycles (Graham 2010) although we assume that their present natural distribution is close to being in equilibrium with their environment. More recently anthropogenic disturbances have greatly altered tropical landscapes (Hansen et al. 2008), hence affecting species' distributions. However, for the purposes of this study we assume that such effects have not yet dramatically influenced the EOO of the species (we discuss this assumption below).

We also selected 30 species for which full IUCN Red List assessments have recently been conducted by staff at RBG Kew and NHM (Appendix 1); six were assessed as being of conservation concern. We refer to these as "SRLI species" because the occurrence data for these came from the plant component of the Sampled Red List Index (SRLI), an indicator to measure the current rate of loss of biodiversity by tracking trends in the conservation status of

a randomly-selected sample of species (see details in: Baillie et al. 2008; Brummitt et al. 2008). We focused on monocotyledonous (monocot) and pteridophyte (fern) SRLI species occurring in (but not necessarily endemic to) Costa Rica and Panama, which are within the Mesoamerica biodiversity hotspot (Myers et al. 2000). Many species in this region are poorly represented in the world's herbaria, so limited knowledge of their true distribution exists; nonetheless, conservation assessments are urgently needed for these and thousands of species like them. The problems in this region are thus typical of conservation assessments more widely.

2.2 Environmental Variables

Different sets of environmental data layers were selected for each of the major groups of plant species: monocots, dicots and ferns. All variables had a 30 arc second (~1km at the Equator) spatial resolution (see Appendix 2 for details). These were selected from the Worldclim database, version 1.4 (http://www.worldclim.org; Hijmans et al. 2005) and from the Consultative Group for International Agricultural Research Consortium for Spatial Information (CGIAR-CSI; http://www.cgiar-csi.org). Their selection was based on a combination of correlation, principal components and cluster analyses (Syfert et al. 2013) to minimise the correlation between the layers, and expert judgement based on the ecology of the taxa. This selection led to the initial pool of 24 different candidate environmental variables being reduced to between 5 and 6 environmental variables, depending on the taxonomic group being modelled (Appendix 2). We note that none of the chosen environmental variables explicitly account for biotic interactions (e.g. dispersal or competition) or historical constraints, which are known to be important in determining where some species occur (Gaston and Fuller 2009), and the lack of these factors could lead to an overestimation in the SDM EOO (Marcer et al. 2013). However, we do not have enough biological knowledge about any of our chosen species to account for these factors, a typical situation when assessing the conservation status of many plant species.

2.3 Species Distribution Modelling

We used MaxEnt software to build species distribution models (Version 3.3.3; (Phillips et al. 2006; see also Elith et al. 2011 for details; Phillips and Dudik 2008) because it performs well with presence-only data, even with low numbers of known presence localities (Elith et al. 2006; Hernandez et al. 2006; Pearson et al. 2007). We adopted the default regularisation parameters but restricted MaxEnt to using only linear and quadratic functional forms, constraining it to producing relatively simple models that do not over-fit the training data (Merow et al. 2013; Syfert et al. 2013). In a previous study (Syfert *et al*., 2013) and in preliminary investigations to this study, we found that this approach generates useful predictive models for a range of species, even when they have very different numbers of data samples. Sampling bias was controlled for by including a sampling bias dataset (Phillips et al. 2009) constructed from all georeferenced plant occurrence data from the GBIF data portal and from the SRLI project. The spatial extent considered for each species was the area containing the presence data plus a 200 km buffer, following VanDerWal et al. (2009). Experimentation with species of contrasting known range sizes led us to conclude that this buffer struck a good balance between allowing for a wide range of background localities

outside the known species range, without excessively compromising the ability of the models to predict finer scale spatial variation in the species' probability of occurrence (VanDerWal et al. 2009). This also meant that for most species the spatial extent over which SDMs were applied extended from the Pacific coast to the Atlantic coast for most of Central America and only limited the predicted extent in the northern and southern limits of the species range.

2.4 Analytical methods

Minimum convex polygons (MCPs) were calculated using the adehabitat package (Calenge 2006) within the statistical software R (version 2.11; R Development Core Team 2010). Calculating MCPs for SDM predictions requires choosing a threshold value at which to discriminate predicted presences from absences. MCPs were calculated for each species using the Minimum Training Presence (MTP) threshold: the largest probability of occurrence threshold that includes all of the training presence data (Phillips et al. 2006). In addition, MCPs were calculated for each species at all logistic-converted threshold probabilities from 0.05 to 1 in steps of 0.05, to compare MCPs for two methods: similarity in area, and similarity in geographical space (see Figure 1 for details). The latter was assessed as a measure of geographical overlap using the Jaccard Similarity Index (JSI; Araújo et al. 2005; Sangermano and Eastman 2012):

$$JSI = C/(A+B-C), \qquad \text{(Equation 1)}$$

where A is the area of one MCP, B the area of the other MCP, and C is their area of overlap. JSI values of 1 indicate complete congruence in range and 0 indicates no overlap between the MCP from known presences and the SDM-derived MCP (Figure 1c). For all comparisons between MCPs of known presences and of SDM predictions we searched for the threshold probability of presence that gave the most similarity in area (Figure 1b) and in geographical space (maxJSI; Figure 1d). This approach is different from previous studies in which the area of the pixels predicted as present were simply added together to estimate the EOO (e.g. Gauto et al. 2011; Pena et al. 2014; Sergio et al. 2007). Instead, here we draw an MCP around the predicted presences, which additionally contains predicted absences within its range, and use that to estimate the EOO (Figure 1).

Random subsampling was used to assess SDM performance against randomly withheld data, except when models were trained to subsamples of the well-known species. When species had greater than 10 presence localities we fitted SDMs 10 times, each time with a random 20% reserved for model evaluation (Franklin 2009). A leave-one-out method was performed for species with sample sizes below 10 (Pearson et al. 2007). The area under the curve (AUC) in a receiver operating characteristic (ROC) plot was used to assess the model's ability to discriminate between presence localities and other localities in the environment (Franklin 2009), where an AUC value of 1.0 indicates perfect discrimination ability and a value of 0.5 or less indicates a prediction no better than random (although the maximum discrimination ability for presence only data is less than 1.0; see Phillips et al. 2006).

2.5 Assessments using data on well-known species

We randomly generated subsamples of 5, 10, 20, and 30 data points for each of the well-known species, with 5 replicates each (in the Discussion we also include examples of where spatially biased subsamples were used instead of random subsamples). These sizes were chosen to be analogous to those of the SRLI species. Our analyses then consisted of making the following comparisons for each species:

1) Comparing MCPs of the SDM predictions (SDMs trained to the subsamples, termed SDM-derived MCPs) to the MCPs of the subsampled localities (termed subsample MCPs);

2) Comparing MCPs of the SDM predictions (as identified above) to the MCPs of all localities (termed full-sample MCPs);

3) Comparing the MCPs of the subsampled localities (subsample MCPs) to the MCPs of all localities (full-sample MCPs).

3 RESULTS

3.1 Well-known species: model accuracy and thresholds

All of the SDMs trained to the complete dataset for each well-known species discriminate well between localities with and without presences (mean test AUC = 0.872, values from 0.7 to 0.9 indicate moderate predictive performance (Appendix 2; Franklin 2009). The EOOs of the SDM-derived MCPs at different threshold methods are highly correlated with the EOOs of the full-sample MCPs (Appendix 3a). However, when the MTP cut-off is applied, SDMs greatly over-predicted the EOO of the full-sample MCPs. Specifically, the EOO derived using the MTP threshold is on average more than 5 times larger and can be up to 10 times larger than the full-sample MCP (Appendix 3a; see Appendix 4 for examples). In comparison, the maximum geographical similarity threshold leads to predicted EOOs of the SDM-derived MCPs that are closer to the EOOs of the full-sample MCPs (Appendix 3a). Although the similarity in area threshold also leads to the EOO of the SDM-derived MCPs being close to the to the full-sample MCPs (as one would expect; Appendix 3a), it occasionally predicted larger areas than were estimated by the full-sample MCP (e.g. two model replicates for *Cyathea fulva*, one model replicate for *Cyathea schiedeana* and *Hymenophyllum consanguineum*). This is because the similarity in area method relies on the two areas becoming equal at some threshold probability whereas in some cases the highest probabilities of occurrence in the models included pixels that occur over a larger extent than the full-sample MCP.

By definition, the SDM-derived MCPs based on the MTP cut-off include all presence data. The maximum geographical similarity threshold generally includes a high percentage of presences (>80% in every case), while the similarity in area threshold is more variable (Appendix 3b). Therefore, applying a geographical similarity threshold provides an estimate of the full-sample MCP that is more consistently reliable in capturing a higher proportion of presence data. For this reason, we focus the rest of our results below on the SDM-derived

MCPs using the geographical similarity threshold. The threshold at maximum geographical similarity between the SDM-derived MCP and that of the full-sample MCP varies widely between the 20 species, from 0.4 (*Elaphoglossum furfuraceum)* to 0.8 (*C. fulva;* Figure 2e).

3.2 Inference from small sample sizes: subsampling data for well-known species

At subsample sizes of 5 and 10, the SDM-derived MCPs at the maximum geographical similarity threshold are geographically closer to the full-sample MCPs (Figure 2a & b, blue) than the subsample MCPs are to the full-sample MCPs (Figure 2a & b, black), and are also closer to the full-sample MCPs (Figure 2f & g, blue) than they are to the subsample MCPs (Figure 2f & g, red). SDM-derived MCPs from larger subsample sizes tend to have higher geographical similarity than do those from smaller subsample sizes when compared to either the subsample or the full-sample MCPs (Figure 2). This was expected on the basis that more samples provide a more representative training dataset of the environments associated with species presences, leading to better models.

The low geographical similarity between the SDM-derived MCPs and the subsample MCPs at subsample sizes of 5 and 10 (5 samples, median maxJSI= 0.15; 10 samples, median maxJSI= 0.27) occurs because larger MCPs are predicted than the subsample MCPs. For most species, those larger predicted areas lie predominantly within the full-sample MCPs (Appendix 5): subsample sizes of 5 and 10 have over 65% of the models with a proportion of 0.75 or more of SDM-derived MCPs within the full-sample MCP. Geographical similarity increased as subsample size increased and the proportion of the SDM-derived MCPs lying within the full-sample MCPs also increased slightly (20 samples, median maxJSI= 0.47, 0.72; 30 samples, median maxJSI= 0.43, 0.74).

The EOOs of the SDM-derived MCPs at the maximum geographical similarity threshold are significantly closer to the EOOs of the full-sample MCPs than the EOOs of the subsample MCPs at all subsample sizes (Figure 3; paired sample t-test, $p < 0.001$), although the improvement in the estimated EOO is notably larger at sample sizes of 5 and 10 (Figure 3a,b). In addition the EOOs of the SDM-derived MCPs are relatively insensitive on average to the sample sizes used to train the SDMs (Figure 3 and Appendix 6).

3.3 SRLI species

The EOOs of the SDM-derived MCPs correlate significantly with the EOOs of the SRLI MCPs at the maximum geographical similarity ($R^2= 0.925$, $p < 0.001$; Figure 4) but data for the well-known species in Figure 4 tend to lie closer to the one-to-one line than those for the SRLI species. This is similar to what is found with the analysis of small sample sizes from the well-known species (Appendix 7a-d) where the EOOs of the SDM-derived MCPs tend to over predict the subsample MCPs when using small subsets of the data, even though the EOOs of those MCPs were actually closer to the EOOs of the full-sample MCPs (Figure 4). The slope of this relationship is not different from the one-to-one slope (standardized major axis analysis (SMA; Warton et al. 2012); $R^2=0.027$, slope = 1.007, upper slope CI= 0.906 and lower slope CI= 1.120), indicating that the tendency for the SDM-derived MCPs to be larger than the SRLI sample MCPs does not vary with the area size being predicted.

SDM-derived MCPs for the SRLI species show a range of maximum geographic similarity (i.e. maxJSI) with the SRLI MCPs (median maxJSI = 0.43, minimum = 0.12 maximum = 0.733) similar to that found for the well-known species, and no relationship with sample size and AUC values. However, unlike the well-known species, the SDMs trained to the data for SRLI species vary considerably in AUC values (Appendix 1).

3.4 Case studies: species with small sample sizes assessed for the Sampled Red List Index for Plants

We focus on two SRLI species with small sample sizes: *Ctenitis chiriquiana* (5 occurrences) and *Brachionidium dressleri* (10 occurrences). These are assessed as IUCN threatened categories *Endangered* and *Vulnerable*, respectively, through assessments carried out for the SRLI for Plants project (Figure 5; Brummitt et al. 2008). Model discrimination ability varied between the two species, while the maximum geographical similarity (maxJSI) was low: *Brachionidium dressleri* (mean test AUC= 0.936, maxJSI= 0.269); *Ctenitis chiriquiana* (mean test AUC=0.771, maxJSI=0.240). The low geographical similarity occurs because areas larger than the MCP derived from the data are predicted (Figure 5); this is similar to what is observed when using subsamples of data for species with well-known distributions, where the SDM-derived MCPs more accurately reflect the true range of the species (Figure 5, Appendix 4b,d). For example, compare these results with those for subsamples of two well-known species: *Polystichum concinnum* (subsample size 5) and *Anthurium watermaliense* (subsample size 10). For both species, although more distinctly for *A. watermaliense*, larger EOOs were predicted by the SDM than the area estimated from the subsample MCP; the maxJSI was higher when the SDM-derived MCP was compared to the full-sample MCP (Figure 5c & d) than to the subsample MCP (Figure 5e & f). This is representative of our subsampling analysis more generally. For both of the SRLI species the areas of the SDM-derived MCPs lie within the *Vulnerable* category (5,000 km² < EOO < 20,000 km²), although *C. chiriquiana* was assessed in the *Endangered* category (100 km² < EOO < 5,000 km²).

4 DISCUSSION

The IUCN recommends a data-driven assessment of species' conservation status (IUCN 2013), although either the known or inferred sites of occurrence could be used to calculate EOO. Our results here demonstrate that, for plant species in Central America, the EOOs estimated from SDMs can be more representative of known EOOs than those derived solely from a small number of specimens (Figure 3). Moreover, our approach to constrain the SDM predictions to the geographic shape of the point-based EOO appears to provide a conservative approach to identifying potentially suitable environments where a species might occur but has yet to be found. SDMs offer the opportunity to increase the objectivity of these assessments by providing quantitative range estimates based on the relationship between species and their environment (Sangermano and Eastman 2012). Hence, IUCN conservation assessments could benefit from the inclusion of SDMs as objective information without possibly subjective biases from experts (Fourcade et al. 2013). However, it is likely that conservation assessments could benefit the most by employing SDMs in conjunction with expert judgement (Marcer et al. 2013).

4.1 Estimating EOO derived from SDMs

Our conclusions are drawn from plant species from several higher taxonomic groups (dicots, monocots and ferns), occurring over a restricted geographical area (Central and South America). These were chosen because we have particular interest in assessing the conservation status of plant species from this area, rather than aiming to be more general. It is therefore possible that our method works particularly well for these data and here we discuss reasons for why our insights might not extend more generally.

In our case, the areas of environmental conditions similar to those in which the species has been observed appear to be good predictors of where the species is also likely to occur. This may not always be true: species may not occupy the entire niche space revealed by correlative SDMs as a consequence of biotic factors (e.g. dispersal limitation, competition), disturbance effects (e.g. hurricanes) or anthropogenic effects (e.g. deforestation) not included in the model fitting process (Elith and Leathwick 2009). Additionally, the SDM approach we have taken here does not explicitly take into account non-equilibrium species dynamics. While we believe this is a reasonable assumption for the plant species in this study, it could generate misleading predictions for species that are rapidly expanding or contracting in their ranges. Conversely, nor do the Red List Criteria explicitly take into account non-equilibrium dynamics as assessments of conservation status are based on a snapshot of a dynamic process of species' range-formation. We may have been fortunate that our predictions of range extent using limited abiotic data enabled reliable predictions of the EOO for our species in which historical factors and dispersal limitation were not essential for their range extents (Gaston and Fuller 2009). However, for other species this approach may lead to the predicted EOO overestimating the true EOO; this would obviously be undesirable in cases where the species is actually more threatened than the method implies, risking underestimating the Red List Category. When the EOO is overestimated by the SDM, the EOO derived from the point-based MCP should be reported in a conservation assessment, although the EOO overestimate from a carefully constructed SDM could provide an ecologically-based inference of possible important variables missing from the models, such as historical processes (Marcer et al. 2013).

Our predictions were also made assuming that the species environment relationship has reached some form of equilibrium (Austin 2002). The assumption is often made with large-scale distribution modeling in which the underlying biology is poorly known (Austin 2002; Guisan and Zimmermann 2000). This assumption might be particularly problematic for species that have experienced rapid changes in their environmental niche space, such as threatened species that have already experienced dramatic reductions in their geographic range, or mobile species that have only been observed in a restricted part of their environmental space. This again would lead to the EOO being overestimated. Unfortunately it is impossible to use presence only data alone to identify whether this might be the case. It therefore seems possible that the additional environmentally suitable areas predicted by our method may not accurately reflect where the species is also likely to occur, leading to the inferred EOO being overestimated, again an undesirable result in the Red List assessment process.

Surprisingly, our approach provides informative predictions for species known only from very few collections (5-10 specimens), supporting previous studies that have also shown the potential usefulness of models derived from small sample sizes (Hernandez et al. 2006; Pearson et al. 2007; Thorn et al. 2009). This might again be because our chosen species reliably occupy clear regions of environmental space. However, it might also be because EOO is relatively easy to predict accurately: errors made in predicting the precise locations and areas of specific localities are not as important as whether the overall extent of environmentally suitable areas are predicted well.

4.2 Applying SDMs to Red Lists assessments

Many threatened plant species are known from only a few localities (< 15), and their Red List assessments are often based on a small number of collections (Rivers et al. 2011). Our results indicate that under such circumstances, the inferred EOOs can provide useful additional information about how much larger the actual EOO could be, and where it is likely to extend geographically. For example, in the case of our two example SRLI species, *C. chiriquiana* and *B. dressleri*, the EOOs of the SDM-derived MCP would imply that their distributions are likely to be larger than that indicated by the data. These examples show the potential to strengthen our confidence that these species are likely to have restricted ranges occurring along an ecological gradient; in this case the suitable environments coincide with the presence of cloud forests in the Talamanca Mountains. We concur that identifying potentially suitable environmental conditions for the species provides greater value for inferring the ecology of poorly-known species. These predictions could clearly be useful to guide future field expeditions and/or be important information for reassessments (Pearson et al. 2007). The predictions also imply that on the basis of the inferred EOO, the IUCN ratings of these species should be *Vulnerable* rather than *Endangered* (Figure 5). Accepting such a revision on the basis of this study would perhaps contradict the precautionary principle; however, this information may be useful when prioritising species within conservation categories for attention: one might choose to investigate or protect an observed and inferred Critically Endangered species rather than one that is observed to be Critically Endangered but is robustly inferred to be Vulnerable.

Our method might also be used to highlight data points that require further investigation for possible errors or inaccuracies in taxonomy or locality. This could be assessed by further investigating why high habitat suitability is associated with points outside the known range (in the case of discovering potential unknown populations (Guisan et al. 2006) or for reintroduction activities (Thorn et al. 2009)), or why some presence localities have particularly low probabilities associated with them (in the case of checking the accuracy of data points).

4.3 Insights to Methods

Our analysis of the SRLI species produced a number of situations in which the model discrimination ability was low (AUC< 0.7, Appendix 1). If we were attempting to gain insights into the EOO of those species we would normally try to improve the model further, at least to obtain a sufficiently high discriminatory power for the model to be useful (e.g. AUC > 0.7, Franklin 2009). Our models trained with the full-sample data for the well-known species always had consistently high predictive performance, measured as AUC (Appendix 1); however, the maximum geographic similarity varied (Figure 2). This is due to the model predicting presence data well, but not including some known localities in the modelled range when a particular threshold is applied (examples are shown for SRLI species in Appendix 8,

but we obtained similar results for some of the well-known species). The fact that the geographical similarity between the SDM predicted EOOs and that of the well-known species was rarely >80% (Figure 2), and often much less, indicates that many of our models made a >20% commission error in predicting the species geographic extent. It is therefore important to consider, in addition the model's discrimination ability, whether the predicted range at maximum geographical similarity does indeed include a high proportion of known presence localities, or a high proportion of the area bounded by these presence localities.

Occasionally, SDM-derived MCPs predicted for SRLI species had spatial predictions that included localities at the very edges of the study extent (e.g. Appendix 8a). In such situations it is difficult to know whether the SDM is over-predicting the EOO or whether the study extent (a 200km buffer around the point location data) does not include the entire species range: ecological and biological expertise, or further field investigations, is needed to judge which of these is more likely. However, limiting the geographic extent of the background data relative to the species presence data (VanDerWal et al. 2009) appears to have avoided this situation for the majority of our species. Although it is possible that our chosen buffer size may not have been suitable for all species, as we only performed preliminary tests on a selection of species with contrasting known geographic extents, our background selection approach did allow us to conduct essentially the same analysis across a wide range of species with different numbers of collections and different geographic extents. We therefore recommend this approach in future analyses. A potential modification might be to replace the fixed 200km buffer with one that is scaled by the degree of separation between known presence localities, e.g. mean inter-point distance, or greatest distance between any pair of known localities.

We used random subsets of the well-known species data to represent the distributions of less well-known species. However, this does not represent scenarios in which the known localities represent a geographically biased subset of the true distribution (e.g. if localities had only been recorded during an expedition to one particular country even through the species distribution spanned several countries). In Appendix 9 we present the results of a preliminary investigation into the sensitivity of our findings to using geographically-biased subsets of our data (10 samples each). As we found with the random subsets, on average the EOO of the SDM is closer to that of the true EOO than the geographically biased subsample (comparing Appendix 9b to Appendix 9c). Moreover, the SDM EOO tends to provide a conservative estimate of the true EOO (Appendix 9b), although a minority of the SDM estimates are actually larger than the true EOO – an over-prediction that would preferably be avoided when making IUCN assessments. Overall however these preliminary results imply that our findings are also robust to geographic sampling bias.

Although we would hope that a high proportion of known presence localities would be included in the inferred EOO, we advise caution in using a probability threshold that includes all known localities (the MTP threshold) in SDM predictions; in our case this clearly over-predicted the MCPs (Appendix 3a). Our observation of over-prediction when using the MTP threshold corroborates with one recent study (Bean et al. 2012) but contrasts with two others (Pearson et al. 2007; Thorn et al. 2009). Pearson et al. (2007) found over-prediction

was a useful feature in which they were interested in the overall predicted distribution (rather than EOO) to identify potentially new species populations and Thorn et al. (2009) focused on smaller study extents, Indonesian islands, and highlight the over-predictions as potential reintroduction zones for threatened species. For our study, in the majority of cases we find that setting a threshold based on maximum geographical similarity strikes an appropriate balance between including a high proportion of presence localities in the predicted range (a particular concern for predicting rare species, Williams et al. 2009) but not predicting an excessively large EOO relative to the EOO estimated from point data. Based on our results, the EOO of the SDM-predicted MCP using the maximum geographical similarity threshold does not tend to over-predict the EOO of the full-sample MCP. Similarly, Sangermano & Eastman (2012) found maxJSI to be an effective threshold method for refining range maps.

We adopted an approach using MaxEnt for the purposes of this study in which we have found to work well for a range of species, and this approach includes adopting the default regularisation parameters but limiting MaxEnt to using linear and quadratic features only (see Methods). However applying SDMs to inform the Red Listing process should always carefully consider the appropriateness and effects of the chosen SDM methods and their settings (Merow at al. 2013), such as exploring the effects of adopting different regularisation parameters in MaxEnt (e.g. Elith et al. 2010) or employing different SDM methods entirely.

4.4 Conclusion

Estimates of EOO underpin most Red List assessments under IUCN Criterion B, as accurate population estimates are hardly ever available for plant species. In such cases, verifiable and geo-referenced herbarium specimens usually represent the best available data for conservation assessment purposes (Brummitt et al. 2008; Rivers et al. 2011). IUCN guidelines suggest that minimum, maximum and best estimates of EOOs are recorded in cases where there is uncertainty (IUCN 2013); this creates a role for SDMs in species conservation assessments because they can be used to estimate what the EOO might be under different assumptions. In particular, recent studies give support to the role of SDMs in estimating EOO to assist in assessing species' conservation status when data are limited (Pena et al. 2014) or when species are rare (Marcer et al. 2013). In a recent paper investigating the role of SDMs to guide conservation decisions, Guisan et al. (2013) stress the need for SDMs to be developed through practice-oriented case studies. Our case study of plants from the neotropics provides a feasible approach towards applying SDMs to conservation assessments that has the potential to be cost effective for megadiverse regions with high rates of habitat loss. However, while the methods we propose here work for the selected plant species, more extensive taxonomically- and geographically-extensive testing of SDMs, as well as tests using modelled artificial species with prescribed ecological and life history characteristics, are needed before this approach can be recommended for general application to the Red Listing process.

Challenges remain in identifying guidelines so that SDMs are used appropriately in order that they usefully inform conservation assessments. For example, using SDMs as supplementary information could aid in the understanding of whether a species has a

restricted range, whether potential distributions indicate additional suitable localities for the species and can be used to guide future surveys, or perhaps for the reassessment process. In addition it is essential that assessors evaluate the conservation implications of commission and omission errors, especially to avoid underestimating the extinction risk of a species. However, based on the results of this study we conclude with some general findings and possible approaches for incorporating SDMs into Red List assessments:

- Control for sampling bias (Kramer-Schadt et al. 2013; Merow et al. 2013; Syfert et al. 2013) and consider background extent (e.g. apply an appropriate buffer to define a suitable extent; VanDerWal et al. 2009); this is imperative.
    - The SDM might not be informative when the predictions go to the edge of the study extent; it is likely that the full geographic spread has not been fully considered (e.g Appendix 8a)

- Evaluate whether the model has reasonable discriminatory power in order to be useful (e.g. AUC > 0.7, Franklin 2009).
    - The SDM might not be informative when the SDM test or training AUC is below 0.70, which indicates that the model poorly predicts its own data (e.g Appendix 8a)
    - If relevant, investigate why high probabilities might be associated with points outside the known range, or why some known presence localities have particularly low probabilities associated with them. The SDM might not be informative if these discrepancies cannot be explained.

- Evaluate the maximum geographical similarity (maxJSI); lower values indicate that the degree of overlap between point-based and SDM-derived EOOs is small and highlight whether the low value represents an informative or unrealistic degree of overlap.
    - The SDM might not be informative when the geographical overlap extends far beyond the point-based EOO (e.g Appendix 8a).

- Evaluate whether the predicted range at maximum geographical similarity includes a high proportion of known presence localities, or a high proportion of the area bounded by these presence localities.
    - The SDM might not be informative when only a small fraction of the point-based EOO is captured by the SDM-derived EOO (e.g Appendix 8b).

- Evaluate if the results could be used to help target future survey efforts
    - In such cases, the predicted distribution rather than the predicted MCP is more likely to be useful (e.g. Appendix 4a and Pearson et al. 2007).

4.5 Acknowledgements

Justin Moat (RBG Kew) provided comments on the manuscript; Alex Monro (NHM, London) contributed data and advice on Neotropical ecosystems. Resit Akçakaya, Richard Pearson and four anonymous reviewers provided valuable comments to the manuscript. This work is supported by Microsoft Research through its PhD Scholarship Programme.

5 Literature Cited

Araújo, M.B., Thuiller, W., Williams, P.H., Reginster, I., 2005. Downscaling European species atlas distributions to a finer resolution: implications for conservation planning. Global Ecology and Biogeography 14, 17-30.

Austin, M.P., 2002. Spatial prediction of species distribution: an interface between ecological theory and statistical modelling. Ecological Modelling 157, 101-118.

Baillie, J.E.M., Collen, B., Amin, R., Akcakaya, H.R., Butchart, S.H.M., Brummitt, N., Meagher, T.R., Ram, M., Hilton-Taylor, C., Mace, G.M., 2008. Toward monitoring global biodiversity. Conservation Letters 1, 18-26.

Bean, W.T., Stafford, R., Brashares, J.S., 2012. The effects of small sample size and sample bias on threshold selection and accuracy assessment of species distribution models. Ecography 35, 250-258.

Brook, B.W., O'Grady, J.J., Chapman, A.P., Burgman, M.A., Akcakaya, H.R., Frankham, R., 2000. Predictive accuracy of population viability analysis in conservation biology. Nature 404, 385-387.

Brummitt, N.A., Bachman, S.P., Moat, J., 2008. Applications of the IUCN Red List: towards a global barometer for plant diversity. Endangered Species Research 6, 127-135.

Burgman, M.A., Fox, J.C., 2003. Bias in species range estimates from minimum convex polygons: implications for conservation and options for improved planning. Animal Conservation 6, 19-28.

Butchart, S.H.M., Stattersfield, A.J., Baillie, J., Bennun, L.A., Stuart, S.N., Akcakaya, H.R., Hilton-Taylor, C., Mace, G.M., 2005. Using Red List Indices to measure progress towards the 2010 target and beyond. Philosophical Transactions of the Royal Society B-Biological Sciences 360, 255-268.

Butchart, S.H.M., Stattersfield, A.J., Bennun, L.A., Shutes, S.M., Akçakaya, H.R., Baillie, J.E.M., Stuart, S.N., Hilton-Taylor, C., Mace, G.M., 2004. Measuring Global Trends in the Status of Biodiversity: Red List Indices for Birds. Plos Biology 2, e383.

Calenge, C., 2006. The package "adehabitat" for the R software: A tool for the analysis of space and habitat use by animals. Ecological Modelling 197, 516-519.

Cardoso, P., Borges, P.A.V., Triantis, K.A., Ferrandez, M.A., Martin, J.L., 2011. Adapting the IUCN Red List criteria for invertebrates. Biological Conservation 144, 2432-2440.

Coulson, T., Mace, G.M., Hudson, E., Possingham, H., 2001. The use and abuse of population viability analysis. Trends in Ecology & Evolution 16, 219-221.

Elith, J., Graham, C.H., Anderson, R.P., Dudik, 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.M., Peterson, A.T., Phillips, S.J., Richardson, K., Scachetti-Pereira, R., Schapire, R.E., Soberon, J., Williams, S., Wisz, M.S., Zimmermann, N.E., 2006. Novel methods improve prediction of species' distributions from occurrence data. Ecography 29, 129-151.

Elith, J., Kearney, M., Phillips, S., 2010. The art of modelling range-shifting species. Methods in Ecology and Evolution 1, 330-342.
Elith, J., Leathwick, J.R., 2009. Species Distribution Models: Ecological Explanation and Prediction Across Space and Time. Annual Review of Ecology Evolution and Systematics 40, 677-697.
Elith, J., Phillips, S.J., Hastie, T., Dudik, M., Chee, Y.E., Yates, C.J., 2011. A statistical explanation of MaxEnt for ecologists. Diversity and Distributions 17, 43-57.
Fourcade, Y., Engler, J.O., Besnard, A.G., Rödder, D., Secondi, J., 2013. Confronting expert-based and modelled distributions for species with uncertain conservation status: A case study from the corncrake (*Crex crex*). Biological Conservation 167, 161-171.
Franklin, J., 2009. Mapping Species Distributions: Spatial Inference and Prediction, first edn. Cambridge University Press, Cambridge, UK.
Gaston, K.J., 1991. How Large Is a Species' Geographic Range? Oikos 61, 434-438.
Gaston, K.J., Fuller, R.A., 2009. The sizes of species' geographic ranges. Journal of Applied Ecology 46, 1-9.
Gauto, I., Spichiger, R., Stauffer, F., 2011. Diversity, distribution and conservation status assessment of Paraguayan palms (Arecaceae). Biodiversity and Conservation 20.
Graham, A., 2010. A Natural History of the New World. University of Chicago Press, Chicago, IL.
Guisan, A., Broennimann, O., Engler, R., Vust, M., Yoccoz, N.G., Lehmann, A., Zimmermann, N.E., 2006. Using niche-based models to improve the sampling of rare species. Conservation Biology 20, 501-511.
Guisan, A., Tingley, R., Baumgartner, J.B., Naujokaitis-Lewis, I., Sutcliffe, P.R., 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, 1424–1435.
Guisan, A., Zimmermann, N.E., 2000. Predictive habitat distribution models in ecology. Ecological Modelling 135, 147-186.
Hansen, M.C., Stehman, S.V., Potapov, P.V., Loveland, T.R., Townshend, J.R.G., DeFries, R.S., Pittman, K.W., Arunarwati, B., Stolle, F., Steininger, M.K., Carroll, M., DiMiceli, C., 2008. Humid tropical forest clearing from 2000 to 2005 quantified by using multitemporal and multiresolution remotely sensed data. Proceedings of the National Academy of Sciences 105, 9439-9444.
Hernandez, P.A., Graham, C.H., Master, L.L., Albert, D.L., 2006. The effect of sample size and species characteristics on performance of different species distribution modeling methods. Ecography 29, 773-785.
Hijmans, R.J., Cameron, S.E., Parra, J.L., Jones, P.G., Jarvis, A., 2005. Very high resolution interpolated climate surfaces for global land areas. International Journal of Climatology 25, 1965-1978.
IUCN, 2013. Guidelines for Using the IUCN Red List Categories and Criteria. Prepared by the Standards and Petitions Subcommittee.
Jiménez-Alfaro, B., Draper, D., Nogués-Bravo, D., 2012. Modeling the potential area of occupancy at fine resolution may reduce uncertainty in species range estimates. Biological Conservation 147, 190-196.
Kramer-Schadt, S., Niedballa, J., Pilgrim, J.D., Schröder, B., Lindenborn, J., Reinfelder, V., Stillfried, M., Heckmann, I., Scharf, A.K., Augeri, D.M., Cheyne, S.M., Hearn, A.J., Ross, J., Macdonald, D.W., Mathai, J., Eaton, J., Marshall, A.J., Semiadi, G., Rustam, R., Bernard, H., Alfred, R., Samejima, H., Duckworth, J.W., Breitenmoser-Wuersten, C., Belant, J.L., Hofer, H., Wilting, A., 2013. The importance of correcting for

sampling bias in MaxEnt species distribution models. Diversity and Distributions 19, 1366–1379.
Liu, C., Berry, P.M., Dawson, T.P., Pearson, R.G., 2005. Selecting thresholds of occurrence in the prediction of species distributions. Ecography 28, 385-393.
Liu, C., White, M., Newell, G., 2013. Selecting thresholds for the prediction of species occurrence with presence-only data. Journal of Biogeography 40, 778-789.
Mace, G.M., Collar, N.J., Gaston, K.J., Hilton-Taylor, C., Akcakaya, H.R., Leader-Williams, N., Milner-Gulland, E.J., Stuart, S.N., 2008. Quantification of Extinction Risk: IUCN's System for Classifying Threatened Species. Conservation Biology 22, 1424-1442.
Marcer, A., Sáez, L., Molowny-Horas, R., Pons, X., Pino, J., 2013. Using species distribution modelling to disentangle realised versus potential distributions for rare species conservation. Biological Conservation 166, 221-230.
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, 1058-1069.
Myers, N., Mittermeier, R.A., Mittermeier, C.G., da Fonseca, G.A.B., Kent, J., 2000. Biodiversity hotspots for conservation priorities. Nature 403, 853-858.
Papes, M., Gaubert, P., 2007. Modelling ecological niches from low numbers of occurrences: assessment of the conservation status of poorly known viverrids (Mammalia, Carnivora) across two continents. Diversity and Distributions 13, 890-902.
Pearson, R.G., Raxworthy, C.J., Nakamura, M., Peterson, A.T., 2007. Predicting species distributions from small numbers of occurrence records: a test case using cryptic geckos in Madagascar. Journal of Biogeography 34, 102-117.
Pena, J.C.d.C., Kamino, L.H.Y., Rodrigues, M., Mariano-Neto, E., de Siqueira, M.F., 2014. Assessing the conservation status of species with limited available data and disjunct distribution. Biological Conservation 170, 130-136.
Phillips, S.J., Anderson, R.P., Schapire, R.E., 2006. Maximum entropy modeling of species geographic distributions. Ecological Modelling 190, 231-259.
Phillips, S.J., Dudik, M., 2008. Modeling of species distributions with Maxent: new extensions and a comprehensive evaluation. Ecography 31, 161-175.
Phillips, S.J., Dudik, M., Elith, J., Graham, C.H., Lehmann, A., Leathwick, J., Ferrier, S., 2009. Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. Ecological Applications 19, 181-197.
Purvis, A., Gittleman, J.L., Cowlishaw, G., Mace, G.M., 2000. Predicting extinction risk in declining species. Proceedings of the Royal Society B-Biological Sciences 267, 1947-1952.
R Development Core Team, 2010. *R: a Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.r-project.org/.
Rivers, M.C., Taylor, L., Brummitt, N.A., Meagher, T.R., Roberts, D.L., Lughadha, E.N., 2011. How many herbarium specimens are needed to detect threatened species? Biological Conservation 144, 2541-2547.
Rodrigues, A.S.L., Pilgrim, J.D., Lamoreux, J.F., Hoffmann, M., Brooks, T.M., 2006. The value of the IUCN Red List for conservation. Trends in Ecology & Evolution 21, 71-76.
Sangermano, F., Eastman, J.R., 2012. A GIS framework for the refinement of species geographic ranges. International Journal of Geographical Information Science 26, 39-55.

Sergio, C., Figueira, R., Draper, D., Menezes, R., Sousa, A.J., 2007. Modelling bryophyte distribution based on ecological information for extent of occurrence assessment. Biological Conservation 135, 341-351.

Syfert, M.M., Smith, M.J., Coomes, D.A., 2013. The Effects of Sampling Bias and Model Complexity on the Predictive Performance of MaxEnt Species Distribution Models. Plos One 8(2): e55158.

Thorn, J.S., Nijman, V., Smith, D., Nekaris, K.A.I., 2009. Ecological niche modelling as a technique for assessing threats and setting conservation priorities for Asian slow lorises (Primates: Nycticebus). Diversity and Distributions 15, 289-298.

VanDerWal, J., Shoo, L.P., Graham, C., William, S.E., 2009. Selecting pseudo-absence data for presence-only distribution modeling: How far should you stray from what you know? Ecological Modelling 220, 589-594.

Warton, D.I., Duursma, R.A., Falster, D.S., Taskinen, S., 2012. smatr 3– an R package for estimation and inference about allometric lines. Methods in Ecology and Evolution 3, 257-259.

Williams, J.N., Seo, C.W., 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, 565-576.

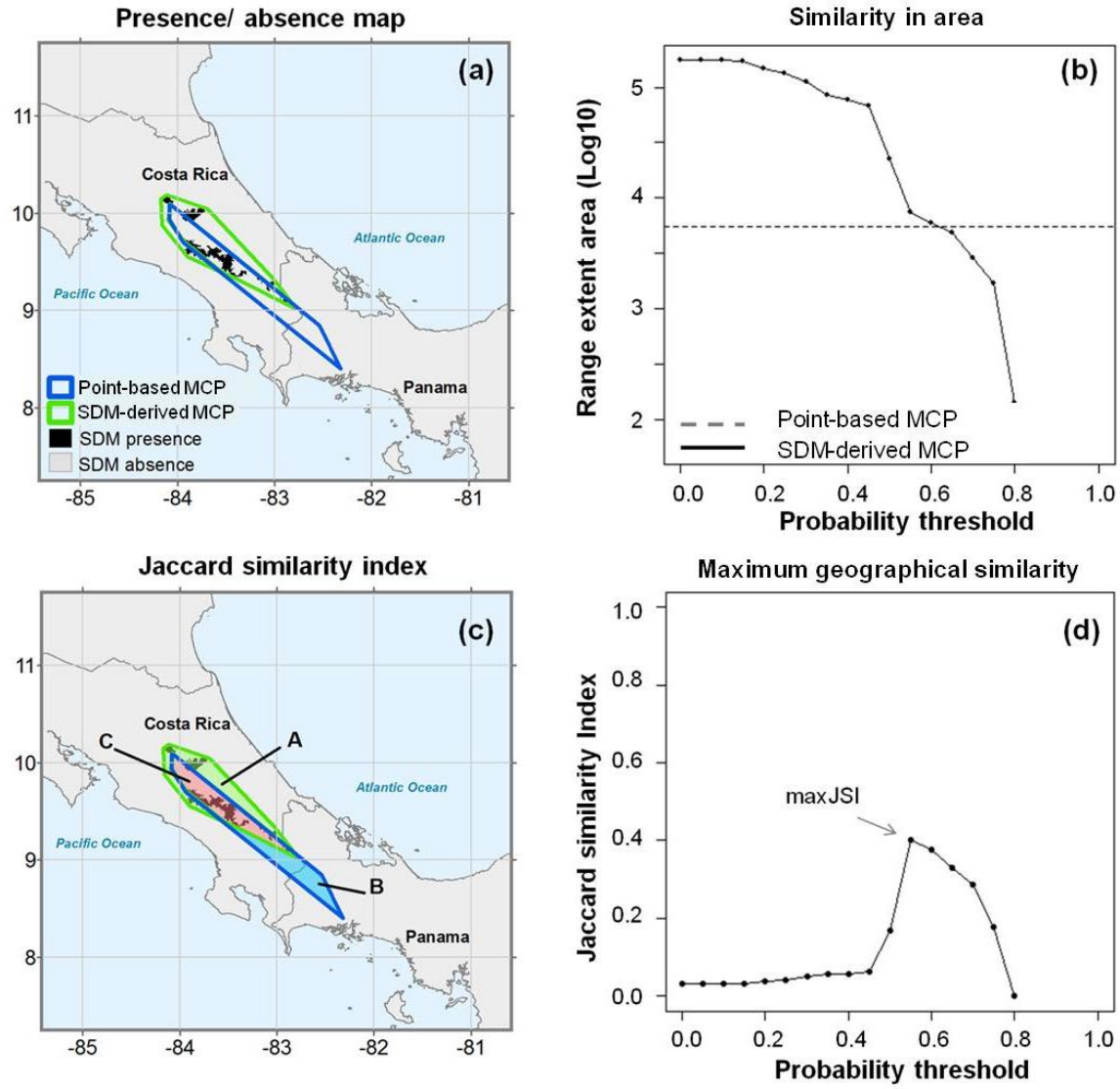


Figure 1. a) Map showing the point data MCP and the SDM-derived MCP overlaid on a presence/absence map as inferred from the SDM with a threshold based on maximum geographical similarity (in this example, 0.65); b) comparing similarity of area (range size) the point data MCP to MCPs predicted by species distribution modelling across a range of probability thresholds (which are used to convert modelled continuous probabilities into predictions of presence and absence); c) Jaccard Similarity Index: C/(A+B-C) in which A= area of SDM-derived MCP, B= area of point data MCP, and C= area of overlap; d) by comparing geographical similarity across multiple probability thresholds, the maximum similarity between the SDM-derived MCP and point data MCP is obtained (maxJSI).

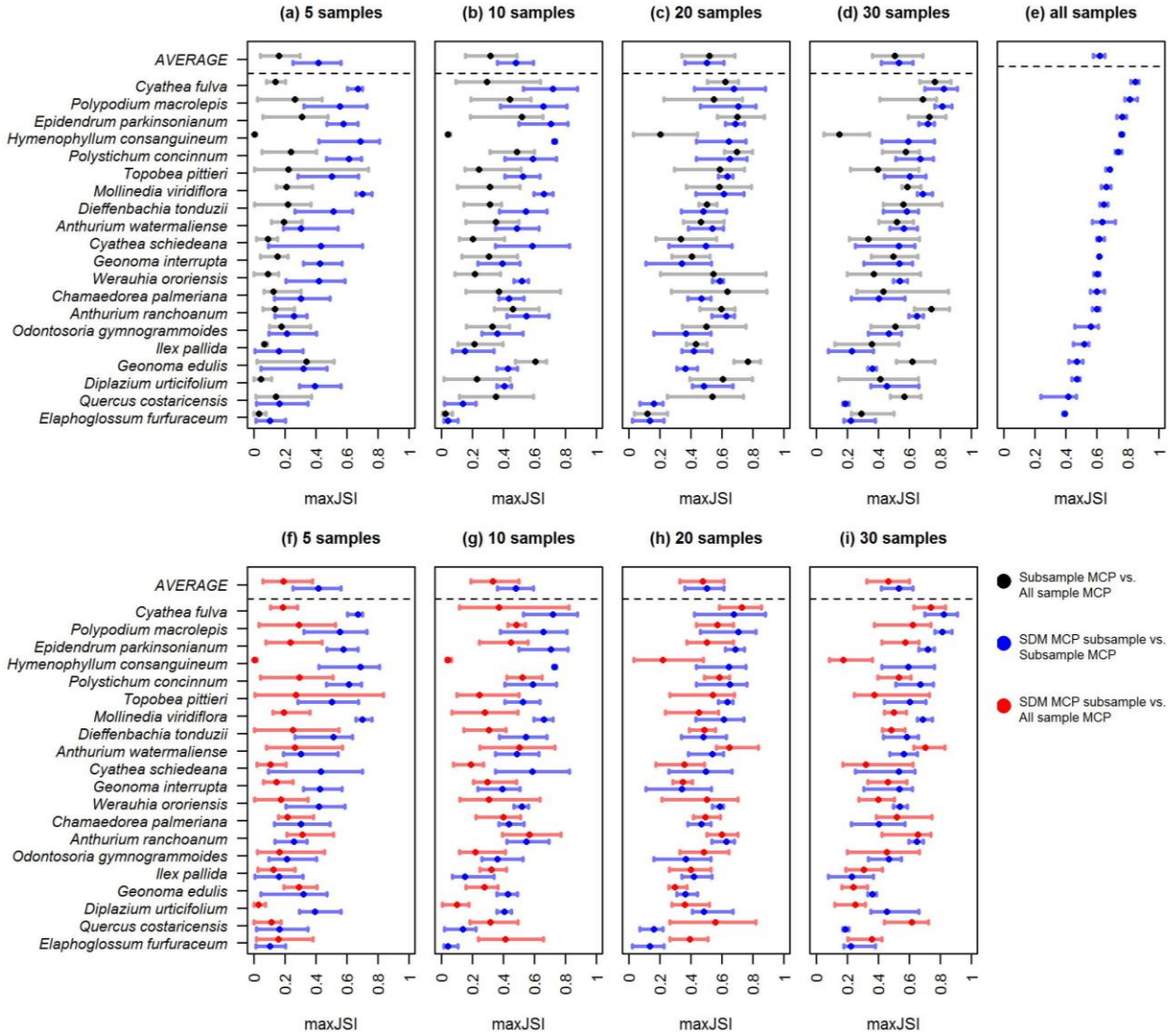


Figure 2. Comparisons of the maximum geographical similarity (maxJSI; see Figure 1 for explanation) between sample MCPs and SDM-derived MCPs for 20 species with well-known distributions using different subsample sizes; (a-d): maxJSI of SDM-derived MCP with full-sample MCP (blue) and maxJSI of subsample MCP with full-sample MCP (black); (f-i): maxJSI of subsample SDM-derived MCP with subsample MCP (red) and subsample SDM-derived MCP with full-sample MCP (blue). Error bars indicate the minimum and maximum maxJSI from 5 replicates. e) corresponds to comparisons for SDMs fitted using all locality data for that species. AVERAGE is the average of the average, maximum and minimum values for the 20 species.

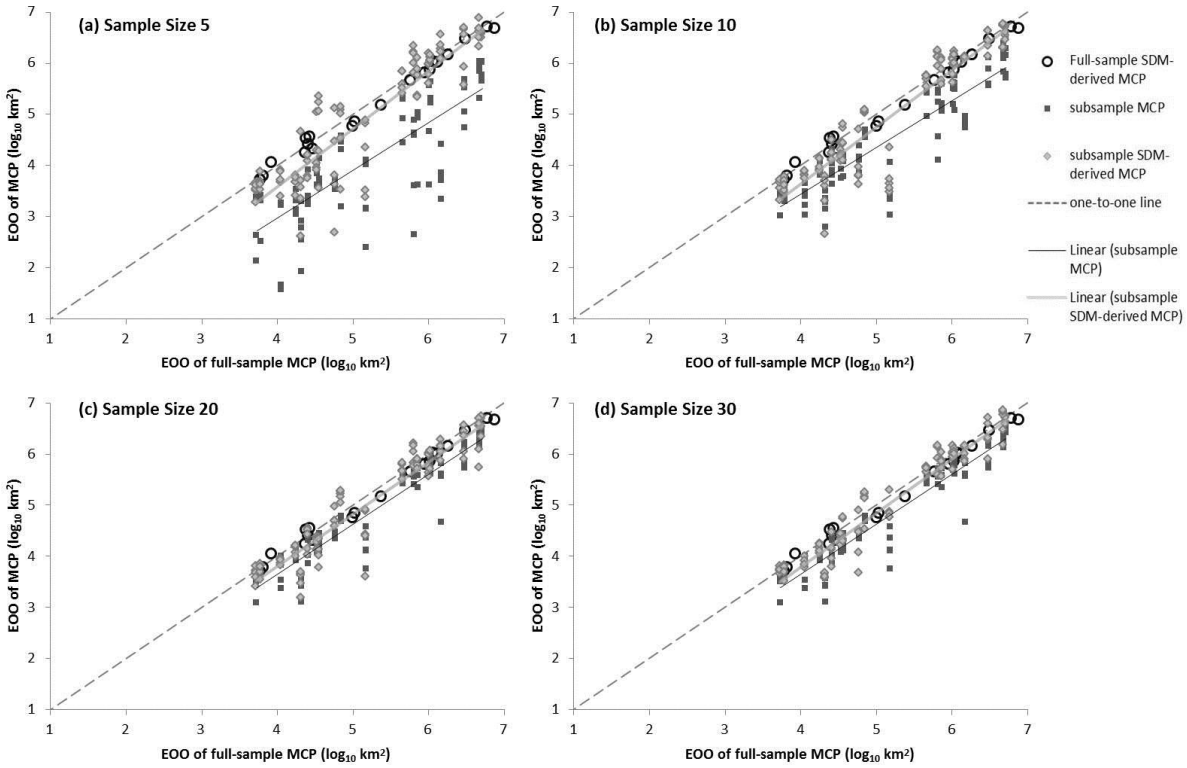


Figure 3. The relationship between the EOO of the full-sample MCP to the EOOs of the SDM-derived MCP (grey diamonds) and subsample MCP (dark grey squares) for each subsampling group. The relationship between the full-sample MCP and SDM-derived MCP is also shown (open black circles).

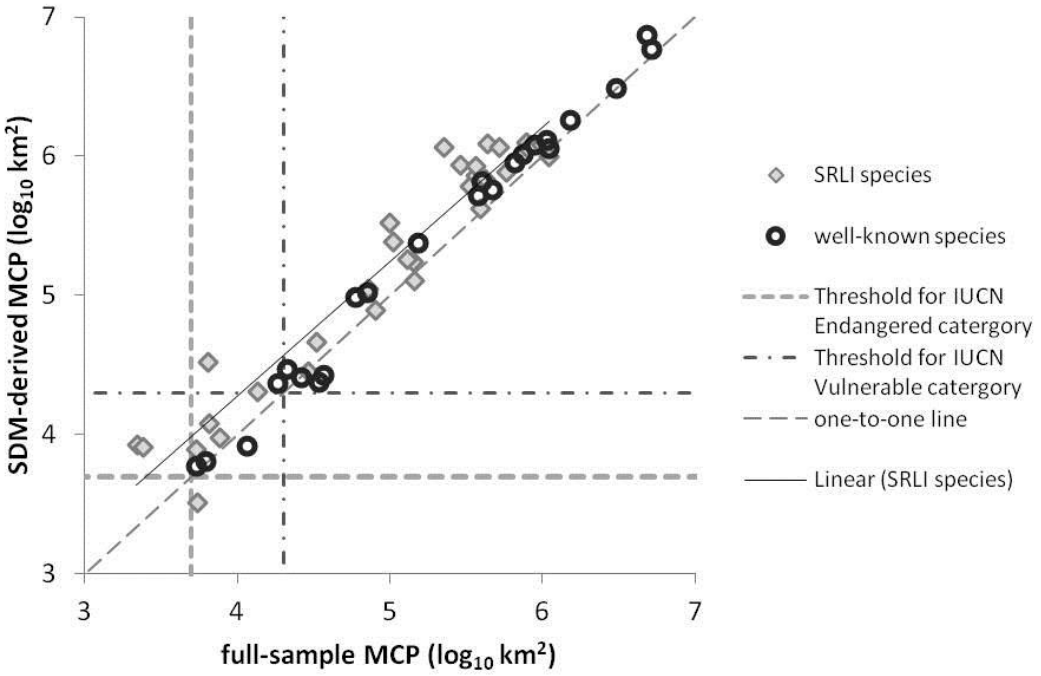


Figure 4. The relationship between the EOO of full-sample MCPs and SDM-derived MCPs, where the SDM was thresholded to maximize geographical similarity for each species. The threshold at which a species is defined as endangered (< 5,000 km²) and vulnerable (< 20,000 km²) according to IUCN criterion B are shown. Area of full-sample MCPs ($\log_{10}$ km²) at the maximum geographical similarity for SRLI species and the SDM-derived MCP ($\log_{10}$ km²) are significantly correlated (p-value < 0.001); the slope of this relationship is not different from the one-to-one slope (SMA; R^2=0.027, slope = 1.007).

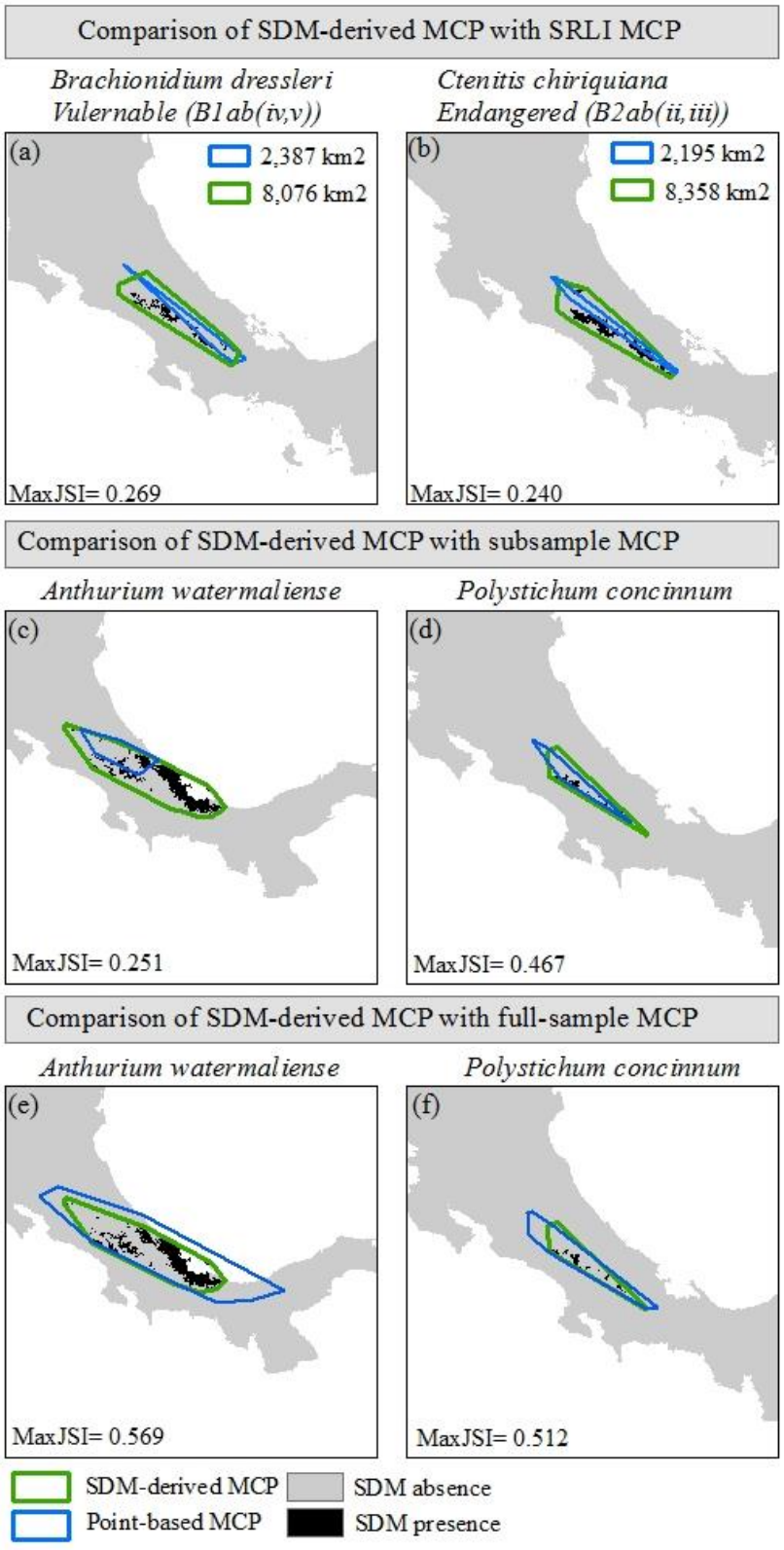


Figure 5. SRLI sample MCPs superimposed on presence/absence maps from which maximum geographical similarity occurs for SRLI case studies with a small number of occurrences (a and b). Selected well-known species with comparative results to the SRLI species in which the SDM-derived MCP larger than the subsample MCP (c and d) but is actually within the full-sample MCP (e-f).

Supplementary Material

Appendix 1

Model performance for species with well-known distributions and species from the Species Red List Index (SRLI) database; training AUC (area under the curve) refers to the AUC assessed on the data used to train the models and test AUC refers to the AUC based on data that were withheld from training (see Methods in the main manuscript for details.

Species (d=dicot; f=fern; m= monocot) Species with well-known distributions	**Sample size**	**Training AUC (±SD)**	**Test AUC (±SD)**
Werauhia ororiensis (m)	54	0.938 ± 0.006	0.954 ± 0.025
Chamaedorea palmeriana (m)	55	0.930 ± 0.006	0.915 ± 0.030
Epidendrum parkinsonianum (m)	61	0.871 ± 0.013	0.847 ± 0.037
Geonoma edulis (m)	68	0.867 ± 0.009	0.853 ± 0.044
Polystichum concinnum (f)	75	0.953 ± 0.003	0.948 ± 0.008
Polypodium macrolepis (f)	78	0.972 ± 0.003	0.969 ± 0.013
Odontosoria gymnogrammoides (f)	79	0.862 ± 0.009	0.838 ± 0.038
Quercus costaricensis (d)	90	0.979 ± 0.001	0.978 ± 0.007
Anthurium watermaliense (m)	99	0.868 ± 0.007	0.857 ± 0.033
Diplazium urticifolium (f)	99	0.890 ± 0.006	0.877 ± 0.034
Cyathea schiedeana (f)	103	0.782 ± 0.013	0.745 ± 0.041
Ilex pallida (d)	103	0.927 ± 0.005	0.925 ± 0.024
Elaphoglossum furfuraceum (f)	111	0.949 ± 0.006	0.951 ± 0.031
Anthurium ranchoanum (m)	112	0.909 ± 0.005	0.893 ± 0.022
Cyathea fulva (f)	120	0.855 ± 0.003	0.853 ± 0.016
Hymenophyllum consanguineum (f)	137	0.867 ± 0.007	0.847 ± 0.023
Dieffenbachia tonduzii (d)	161	0.789 ± 0.005	0.784 ± 0.020
Topobea pittieri (m)	161	0.862 ± 0.005	0.847 ± 0.017
Mollinedia viridiflora (d)	180	0.799 ± 0.010	0.780 ± 0.021
Geonoma interrupta (m)	246	0.788 ± 0.004	0.775 ± 0.029
Red List species			
*Ctenitis chiriquiana (f)**	5	0.889 ± 0.01	0.771 ± 0.218
Olyra standleyi (m)	5	0.78 ± 0.05	0.533 ± 0.323
*Barbosella geminata (m)**	6	0.904 ± 0.02	0.881 ± 0.121

*Acianthera hondurensis (m)**	8	0.82	±	0.03	0.791	± 0.218
*Vriesea camptoclada (m)**	9	0.779	±	0.03	0.666	± 0.268
*Brachionidium dressleri (m)**	10	0.947	±	0	0.936	± 0.026
Guzmania sibundoyorum (m)	12	0.748	±	0.04	0.628	± 0.168
Trichopilia turialbae (m)	17	0.86	±	0.02	0.767	± 0.103
*Telipogon biolleyi (m)**	17	0.835	±	0.03	0.737	± 0.08
Terpsichore alfarii (f)	18	0.861	±	0.02	0.79	± 0.118
Polytaenium chlorosporum (f)	21	0.899	±	0.01	0.875	± 0.029
Platystele minimiflora (m)	21	0.691	±	0.07	0.614	± 0.111
Pleurothallis rowleei (m)	23	0.88	±	0.02	0.871	± 0.095
Anthurium alatipedunculatum (m)	24	0.945	±	0	0.94	± 0.02
Trichopilia marginata (m)	24	0.846	±	0.02	0.683	± 0.059
Marattia interposita (f)	27	0.882	±	0.01	0.862	± 0.034
Palmorchis trilobulata (m)	27	0.798	±	0.03	0.653	± 0.066
Brassia verrucosa (m)	30	0.727	±	0.01	0.687	± 0.029
Pleopeltis fructuosa (f)	32	0.925	±	0.01	0.922	± 0.025
Cyathea williamsii (f)	33	0.894	±	0.01	0.925	± 0.041
Maxillaria hedwigiae (m)	33	0.806	±	0.01	0.696	± 0.061
Cnemidaria cocleana (f)	34	0.857	±	0.03	0.847	± 0.031
Pitcairnia nigra (m)	37	0.839	±	0.02	0.872	± 0.074
Polypodium friedrichsthalianum (f)	42	0.891	±	0	0.89	± 0.026
Terpsichore atroviridis (f)	49	0.905	±	0.01	0.892	± 0.069
Zygophlebia sectifrons (f)	56	0.934	±	0.01	0.919	± 0.025
Elaphoglossum moranii (f)	57	0.887	±	0.01	0.897	± 0.024
Polypodium ursipes (f)	66	0.937	±	0	0.936	± 0.019
Danaea wendlandii (f)	78	0.878	±	0.01	0.87	± 0.021
Elaphoglossum longicrure (f)	86	0.741	±	0.01	0.717	± 0.049

*Species of conservation concern

Appendix 2

Detailed Methods

Environmental Variables

Correlative distribution models for plant species typically incorporate data on nutrients (soil), water availability (precipitation/ evaporation), light (radiation) and temperature (Franklin 2009, Guisan and Zimmermann 2000). Good quality data on soil nutrients and solar radiation were not available for the large study extents we considered (ranging from as far north as Mexico to as far south as central Bolivia) and so we restricted our choice of environmental variables to climate and water availability data.

We aimed to select between 5 and 10 variables that were ecologically relevant for each plant group (Elith and Leathwick 2009). Hierarchical clustering, principal components and Pearson correlation analyses were used to select a subset of environmental variables to reduce the degree of multicollinearity (Franklin 2009), and resulted in the selection of different sets of environmental variables for each plant group (monocotyledonous plants, dicotyledonous plants and ferns). Environmental variables selected for the monocot species were: the annual temperature range (the difference between the warmest and coldest months), the ratio of annual actual evapotranspiration to annual potential evapotranspiration (AET/PET), the minimum temperature of the coldest month, and the precipitation of the coldest, driest and warmest quarters (three month period); fern species: the annual precipitation, AET, water deficit (calculated as PET- AET; Stephenson 1998), the minimum temperature of the coldest month, and the precipitation of the coldest and warmest quarters; dicots: AET, annual temperature range, precipitation of the coldest quarter, water deficit, and precipitation seasonality.

Annual precipitation has long been recognized as a major determinant of species' distributions (Woodward and Williams 1987). The species' tolerance to drought and cold temperatures are characterized as water deficit and minimum temperature, respectively. AET is the amount of water loss given existing evaporative energy in a system and the available water provided by precipitation and storage in the soil (Frank and Inouye 1994), while AET/PET is the index of humidity (Thuiller et al. 2006), which estimates the drought stress as evaporative demand that cannot be satisfied due to limited water supply. Precipitation of the coldest and warmest quarters and precipitation of the driest quarter differentiates the length of the dry season between the Pacific and Atlantic slopes, thus discriminating a species' sensitivity to the duration of minimal precipitation. Similarly, precipitation seasonality can differentiate the length of the dry season between the slopes if many other precipitation variables are highly correlated.

MaxEnt

MaxEnt is among the best-performing of the different presence-only correlative SDM approaches available (Elith et al. 2006, Mateo et al. 2010, Williams et al. 2009). Sampling bias can seriously influence the predictive accuracy of SDMs and several methods have been proposed to deal with the issue. Phillips et al. (2009) proposed generating pseudo-absences from a large dataset that has a bias similar to the occurrence data, and we have shown this approach allows more accurate predictions than sampling pseudo-absences from random locations within the study area (Syfert et al. 2013). We applied this approach here by

obtaining all available georeferenced plant occurrence data for Central and South America available through the GBIF data portal (www.GBIF.org) as well as the data on plant species assessed for the SRLI project (~2.6 million records in total, as of July 2010). The spatial extent of the area modelled also influences model performance (Peterson et al. 2011). In our case, species ranges varied from small ranges contained with Costa Rica and Panama to much larger ranges, for instance, extending from Nicaragua to Venezuela. Hence, we built models with varied extents, fitting to the species ranges to allow for a biologically meaningful fit between a species occurrence and the associated environmental variables. We choose a 200 km buffer around the presence data of each species, which follows a similar approach to Van Derwal *et al*. (2009), in which they found this to be the most favourable distance for generating pseudo-absences from occurrence data in tropical Australia.

References:

Elith, J., C. H. Graham, R. P. Anderson, M. Dudik, S. Ferrier, A. Guisan, R. J. Hijmans, F. Huettmann, J. R. Leathwick, A. Lehmann, J. Li, L. G. Lohmann, B. A. Loiselle, G. Manion, C. Moritz, M. Nakamura, Y. Nakazawa, J. M. Overton, A. T. Peterson, S. J. Phillips, K. Richardson, R. Scachetti-Pereira, R. E. Schapire, J. Soberon, S. Williams, M. S. Wisz, and N. E. Zimmermann. 2006. Novel methods improve prediction of species' distributions from occurrence data. Ecography **29**:129-151.

Elith, J., and J. R. Leathwick. 2009. Species Distribution Models: Ecological Explanation and Prediction Across Space and Time. Annual Review of Ecology Evolution and Systematics **40**:677-697.

Frank, D. A., and R. S. Inouye. 1994. Temporal Variation in Actual Evapotranspiration of Terrestrial Ecosystems - Patterns and Ecological Implications. Journal of Biogeography **21**:401-411.

Franklin, J. 2009. Mapping Species Distributions: Spatial Inference and Prediction. Cambridge University Press, Cambridge, UK.

Guisan, A., and N. E. Zimmermann. 2000. Predictive habitat distribution models in ecology. Ecological Modelling **135**:147-186.

Mateo, R. G., T. B. Croat, Á. M. Felicísimo, and J. Muñoz. 2010. Profile or group discriminative techniques? Generating reliable species distribution models using pseudo-absences and target-group absences from natural history collections. Diversity and Distributions **16**:84-94.

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

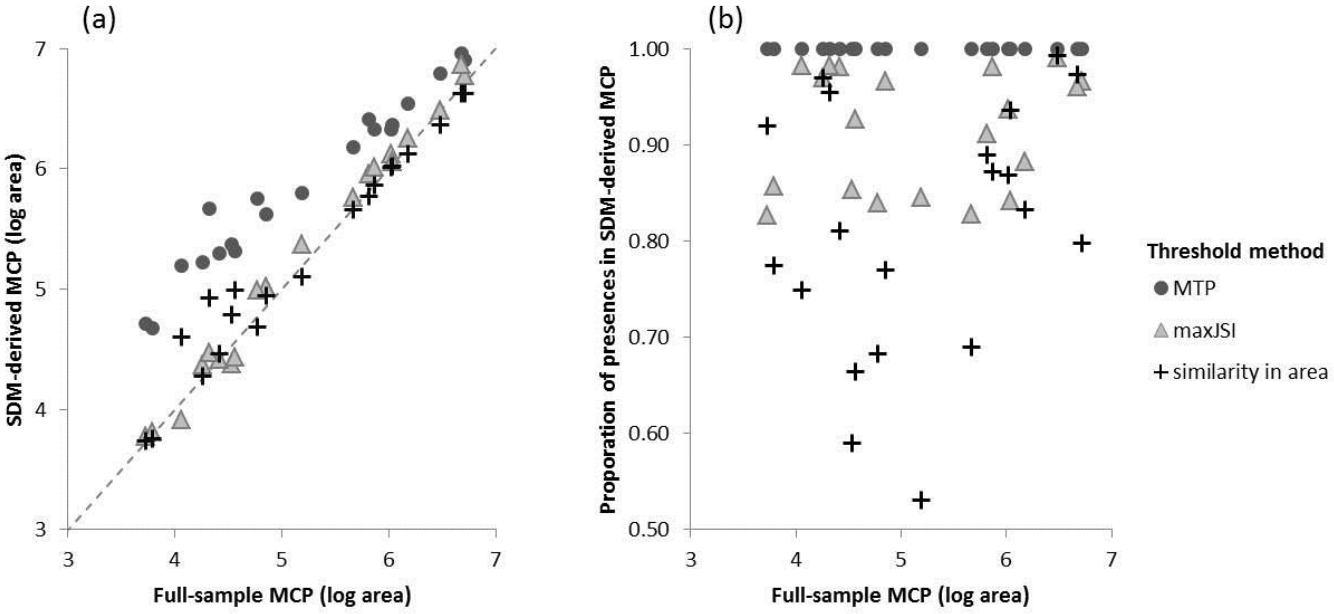
Phillips, S. J., M. Dudik, J. Elith, C. H. Graham, A. Lehmann, J. Leathwick, and S. Ferrier. 2009. Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. Ecological Applications **19**:181-197.

Stephenson, N. L. 1998. Actual evapotranspiration and deficit: biologically meaningful correlates of vegetation distribution across spatial scales. Journal of Biogeography **25**:855-870.

Thuiller, W., G. F. Midgley, M. Rouget, and R. M. Cowling. 2006. Predicting patterns of plant species richness in megadiverse South Africa. Ecography **29**:733-744.

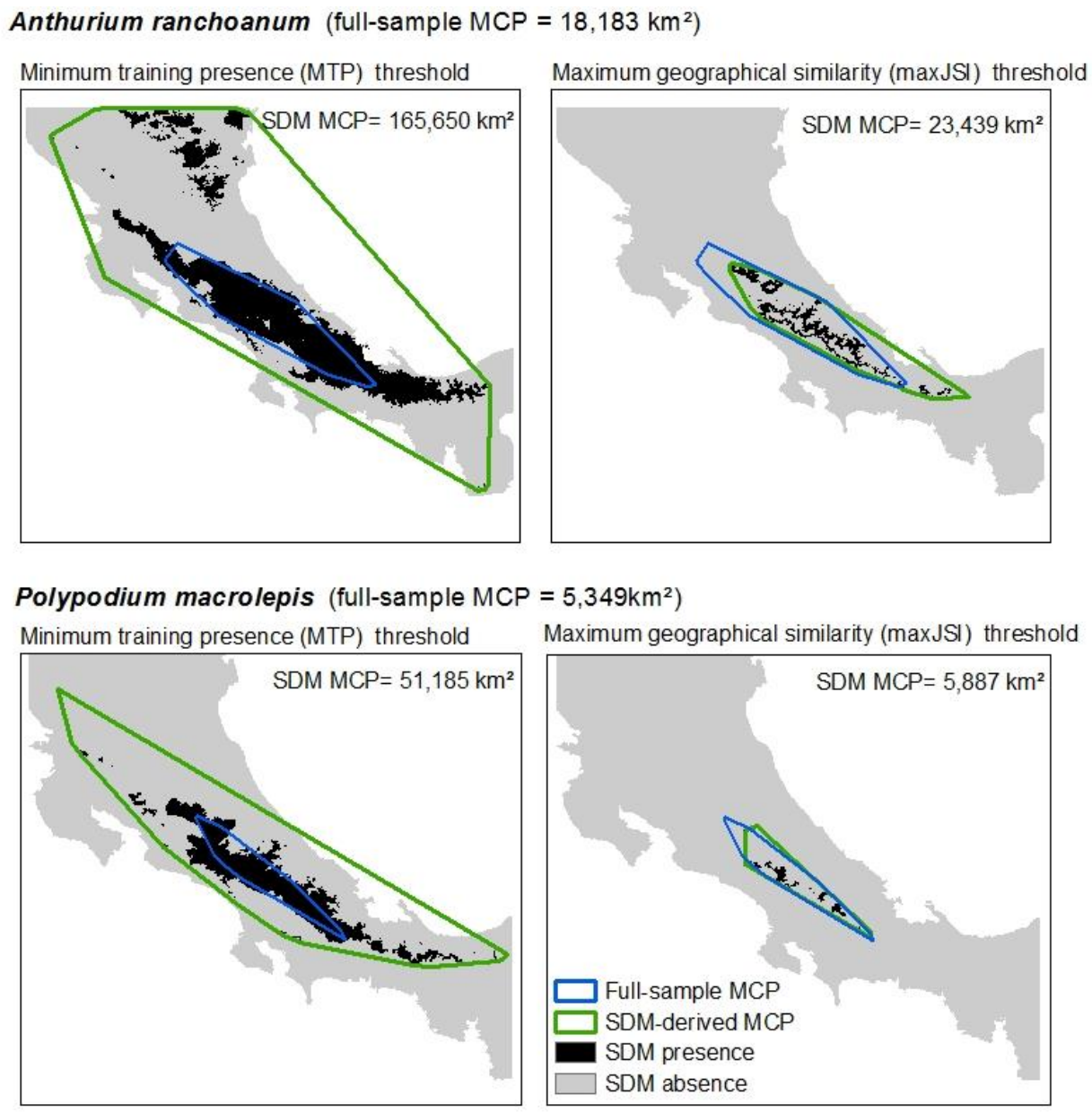
VanDerWal, J., L. P. Shoo, C. Graham, and S. E. William. 2009. Selecting pseudo-absence data for presence-only distribution modeling: How far should you stray from what you know? Ecological Modelling **220**:589-594.
Williams, J. N., C. W. Seo, J. Thorne, J. K. Nelson, S. Erwin, J. M. O'Brien, and M. W. Schwartz. 2009. Using species distribution models to predict new occurrences for rare plants. Diversity and Distributions **15**:565-576.
Woodward, F. I., and B. G. Williams. 1987. Climate and plant distribution at global and local scales. Plant Ecology **69**:189-197.

Appendix 3



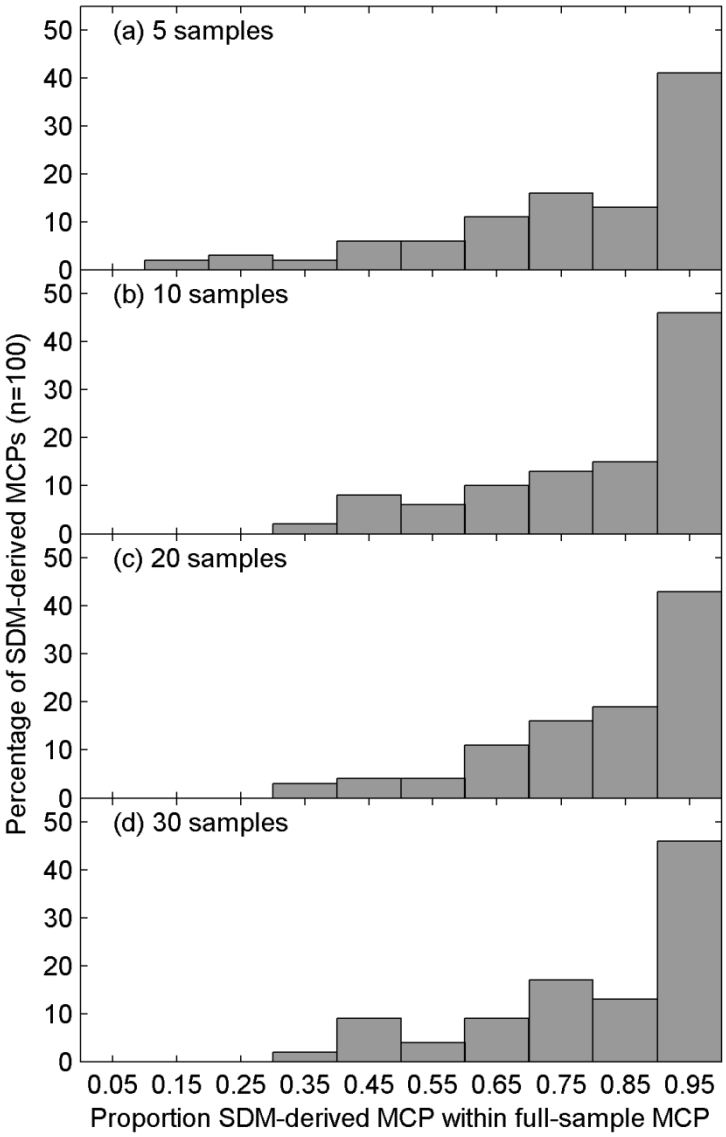
The relationship between the EOO of the SDM-derived MCP and the EOO of the full-sample MCP for three threshold methods: minimum training presence (MTP), maximum geographical similarity (maxJSI) and similarity in area; (b) the proportion of presences predicted present using each threshold method.

Appendix 4



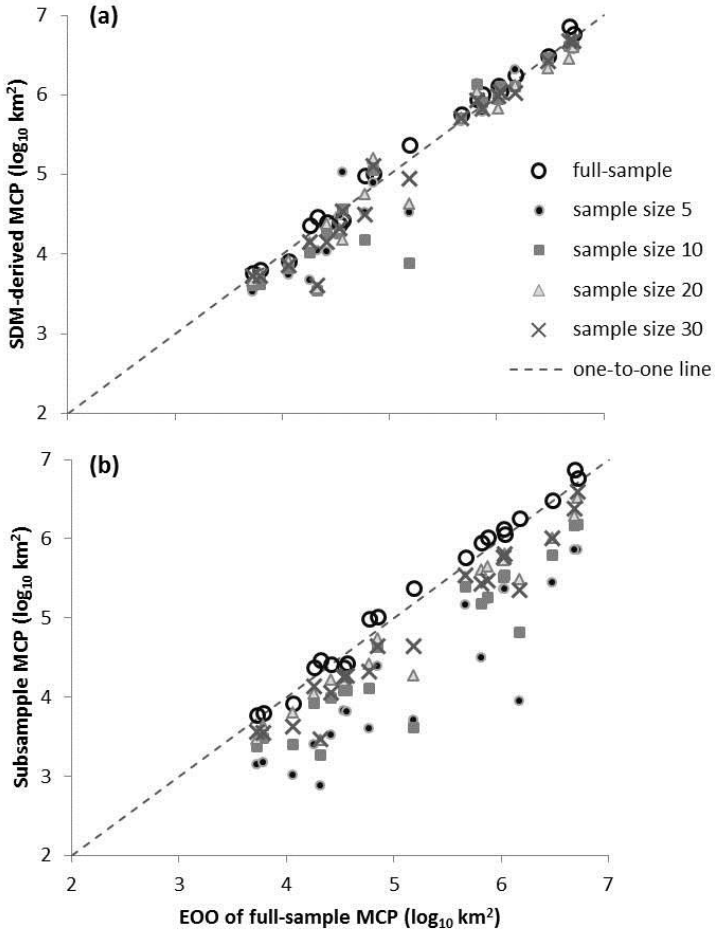
Comparison of threshold methods for two examples of well-known species. Blue polygons are MCPs based on locality data (full-sample), green polygons are SDM-derived MCPs, black areas denote predicted presences and grey areas are predicted absences. In both examples, the SDM-derived MCP based on the MTP threshold is considerably larger than the MCP estimated from the locality data (full-sample). In contrast the SDM-derive MCP from the maxJSI threshold is only marginally larger than the MCP estimated from the locality data.

Appendix 5



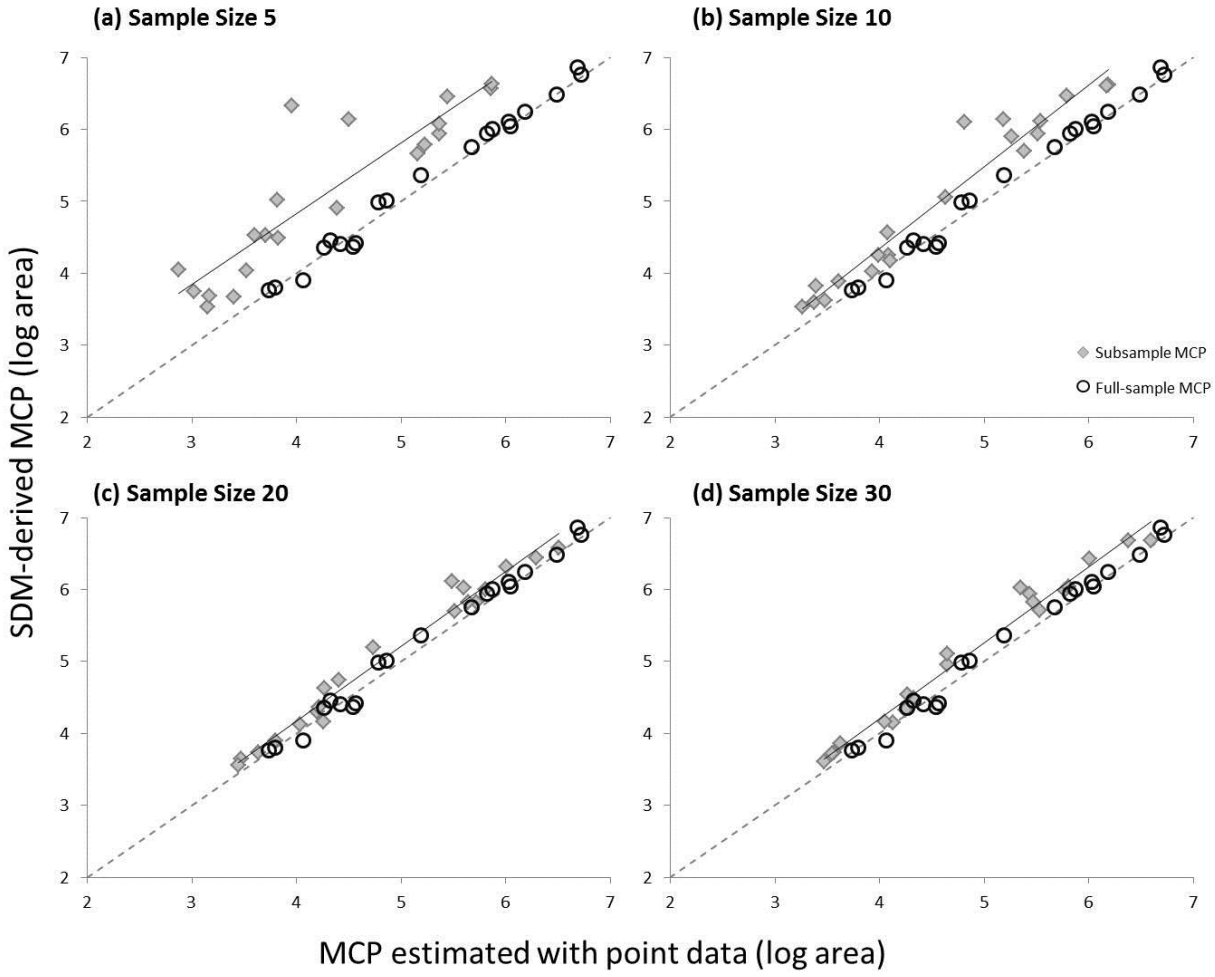
Proportion of SDM-derived MCPs (trained with subsampled data) within full-sample MCP. (a) 66% of the models are within 0.75 proportion of the full-sample MCP; (b) 72% of the models are within 0.75 proportion of the full-sample MCP; (c) 72% of the models are within 0.75 proportion of the full-sample MCP; (d) 74% of the models are within 0.75 proportion of the full-sample MCP.

Appendix 6



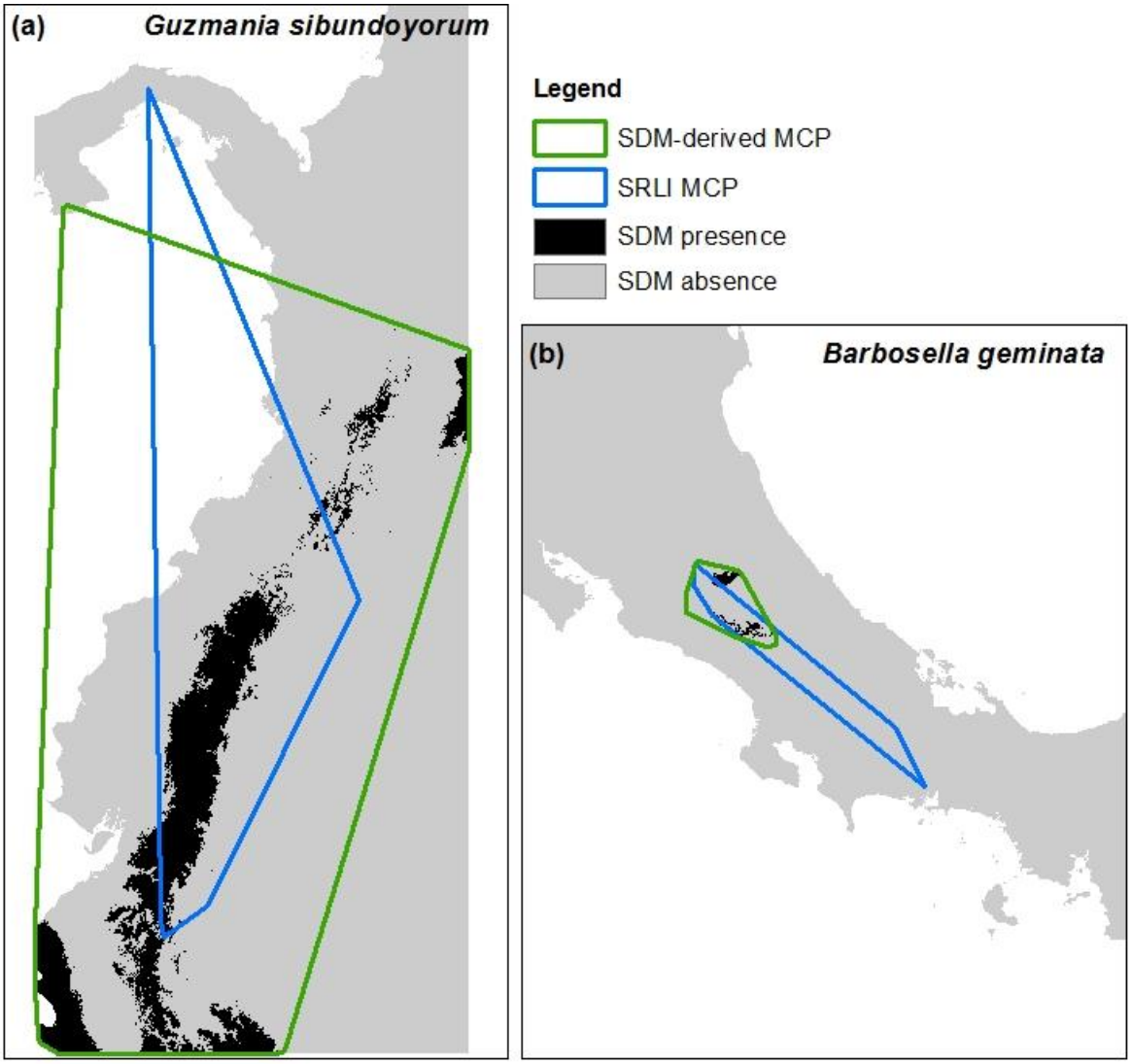
The relationship between the EOO of the full-sample MCP (each species averaged from 5 model replicates) to the EOO of the SDM-derived MCP (a) and subsample MCP (b) for each subsampling group. The relationship between the full-sample MCP and SDM-derived MCP is also shown (open black circles).

Appendix 7



The relationship between the EOO of subsample MCPs (log scale, km²) and SDM-derived MCPs (log scale, km²) based on the threshold at which maximum geographical similarity occurs for each subsampling group of species with well-known distributions (grey diamonds, significantly correlated (p-value <0.001) for all subsampling groups). The relationship between the area of full-sample MCP (i.e. complete dataset) and SDM-derived MCP is also shown (open black circles). The slope of these relationships between the area of subsample MCPs and SDM-derived MCPs is not different from the one-to-one slope for sample sizes, 5, 20 and 30 (SMA; R^2= 0.202, slope = 1.09, R^2= 0.314, slope = 1.05, R^2= 0.413, slope = 1.07, respectively), but the relationship is different from the one-to-one slope for sample size 10 (SMA; R^2= 0.5673, slope = 1.17, p-value < 0.05).

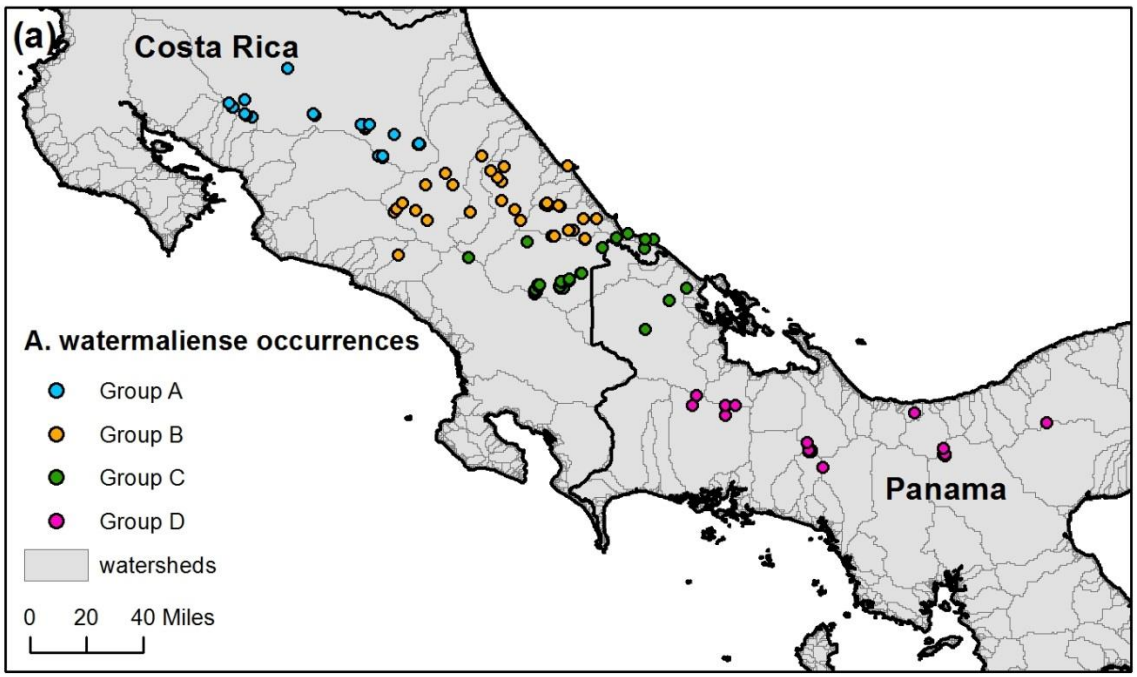
Appendix 8

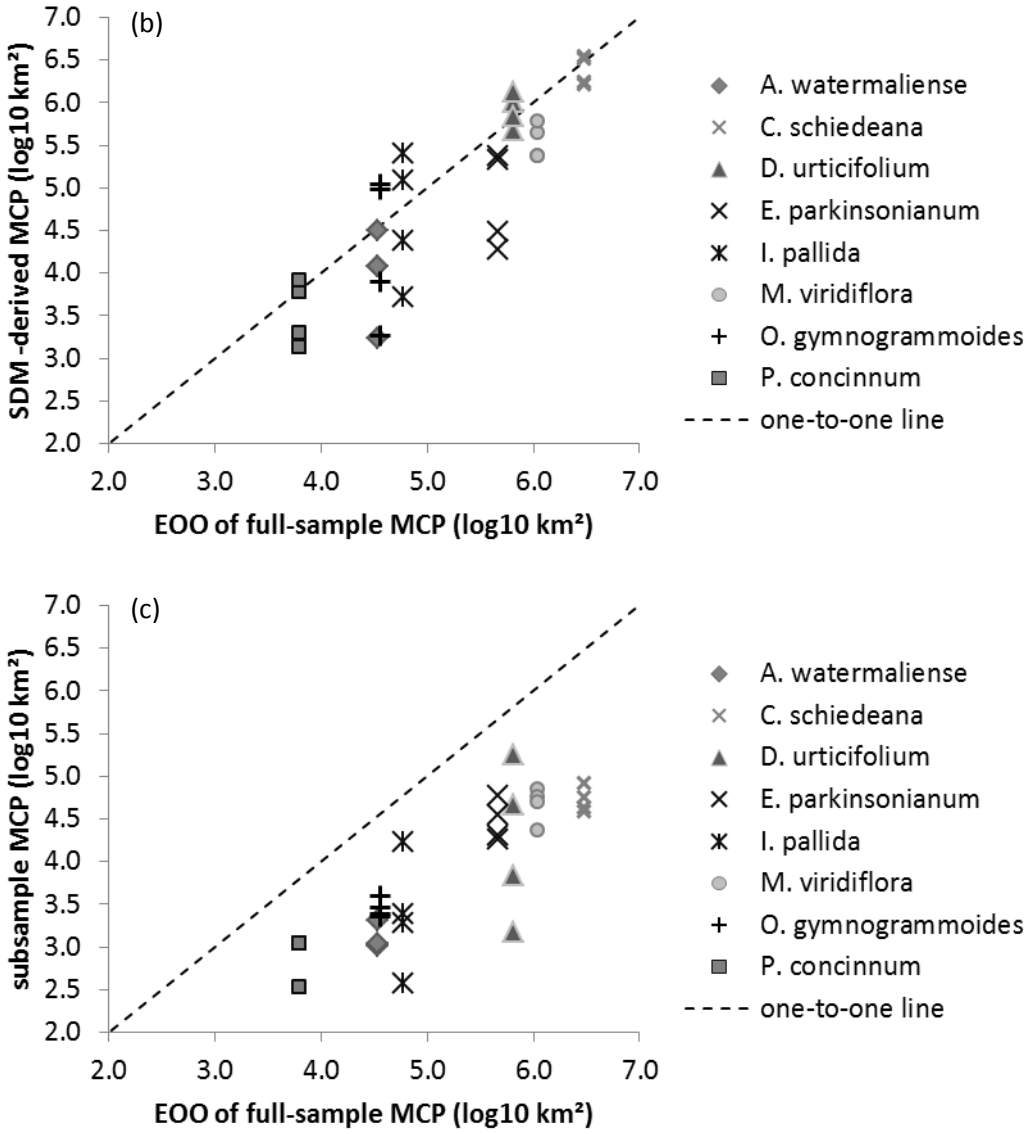


Two examples of SRLI species with a low number of occurrences and low geographical simiarity values. (a) predictions tend to follow a distinct ecological gradient and also go to the edge of the study extent; (b) a small proporation of locality data is included in SDM-derived MCP.

Appendix 9

We performed a preliminary investigation into the accuracy of SDM estimated EOOs when using geographically biased subsets of data. For this investigation we used the species with well-known distributions. Occurrence data for all species were partitioned into four bins using regional watershed boundaries. We chose watershed boundaries to represent the sort of geographical feature that might limit the range over which species are searched for while still allowing us to partition the data and leave a reasonable number of samples with which to train SDMs. Within each geographically biased group, 10 data points were randomly sampled and MaxEnt models were built using 10 data points without replication (these numbers selected after the insights of the other analyses in this paper). Otherwise, the modelling method used was identical to that used elsewhere in our paper. We only used 8 out of the 20 well-known species because only these could be divided into the watershed boundaries whilst leaving 10 data points with which to train SDMs.





EOO comparison with MaxEnt models built from geographically biased subsets; a) example of how species data were partitioned into four groups based on watershed boundaries; b) the relationship between the EOO of the full-sample MCP (log scale, km²) to the EOO of the MCP (log scale, km²) estimated from SDMs; c) the relationship between the EOO of the full-sample MCP (log scale, km²) to the EOO of the geographically biased subsample (log scale, km²) estimated from SDMs.