Diversity and Distributions, (Diversity Distrib.) (2015) 21, 595–608

BIODIVERSITY RESEARCH

Accounting for spatially biased sampling effort in presence-only species distribution modelling

Jessica Stolar* and Scott E. Nielsen

Department of Renewable Resources, University of Alberta, 751 General Services Building, Edmonton, AB T6G 2H1, Canada

ABSTRACT

Aim Presence-only datasets represent an important source of information on species' distributions. Collections of presence-only data, however, are often spatially biased, particularly along roads and near urban populations. These biases can lead to inaccurate inferences and predicted distributions. We demonstrate a new approach of accounting for effort bias in presence-only data by explicitly incorporating sample biases in species distribution modelling.

Location Alberta, Canada.

Methods First, we used logistic regression to model sampling effort of recorded rare vascular plants, bryophytes and butterflies in Alberta. Second, we simulated presence/absence data for nine 'virtual' species based on three relative occurrence thresholds – common, rare and very rare – for each taxonomic group. We sampled these virtual species using our bias model to represent typical sampling effort characteristic of presence-only datasets. We then modelled the distributions of these virtual species using logistic regression and attempted to recover their original simulated distributions using a sample weighting term (prior weight) estimated as the inverse of probability of sampling. Bias-adjusted model estimates were compared to those obtained from random samples and biased samples without adjustment. We also compared prior-weight adjustment to bias-file and target-group background approaches in Maxent.

Results Sample weighting recovered regression coefficients and mapped predictions estimated from unbiased presence-only data and improved model predictive accuracy as evaluated by regression and correlation coefficients, sensitivity and specificity. Similar model improvements were achieved using the Maxent bias-file method, but results were inconsistent for the target-group background approach.

Main conclusions These results suggest that sample weighting can be used to account for spatially biased presence-only datasets in species distribution modelling. The framework presented is potentially widely applicable due to availability of online biodiversity databases and the flexibility of the approach.

Keywords

Biased sampling effort, Maxent bias file, presence-only data, prior sample weight, species distribution modelling, virtual species.

*Correspondence: Jessica Stolar, Department of Renewable Resources, University of Alberta, 751 General Services Building, Edmonton, AB T6G 2H1, Canada
E-mails: stolar@ualberta.ca; scottn@ualberta.ca

INTRODUCTION

Climate change, land use change and habitat fragmentation represent serious threats to biodiversity (Boyd, 2003). To better understand the consequences of these changes and to make recommendations for conservation, species distribution modelling has become an increasingly important tool. Over the past two decades, valuable contributions have been made towards improving model predictions (Austin, 2002; Guisan & Thuiller, 2005; Elith & Leathwick, 2009). However, challenges remain, including how best to incorporate available information when data sources are not comprehensive across the study area (Araújo & Guisan, 2006; Jiménez-Valverde *et al.*, 2008; Beale & Lennon, 2011).

DOI: 10.1111/ddi.12279
http://wileyonlinelibrary.com/journal/ddi

Diversity and Distributions
A Journal of Conservation Biogeography

When modelling species distributions using presence-only data within a presence/background or presence/pseudo-absence design, one critical assumption is that presence locations have been collected without bias (Araújo & Guisan, 2006; Peterson *et al.*, 2011). This is rarely the case, however, as presence-only datasets are often derived from herbaria, entomological collections and museum records. These observations are susceptible to spatial bias with respect to their collection effort (e.g. closer to roads, cities and protected areas) and lack documentation of true absences of species (Schulman *et al.*, 2007; Rocchini *et al.*, 2011). A geographic bias in the sampling of species' occurrences can translate into a bias in the environmental space in which the species' distributions are modelled. Thus, sample selection bias becomes problematic when species are not sampled over the full range of environmental conditions in which they occur (Phillips *et al.*, 2009; Peterson *et al.*, 2011). As a result, spatially biased sampling effort yields model estimates with increased false negative rates (i.e. decreased sensitivity) (Tyre *et al.*, 2003; Botts *et al.*, 2011; Hanspach *et al.*, 2011). For instance, a region may have few occurrence records not because there are truly only a few species in the area, but rather because sampling effort has not been geographically and thus sufficiently environmentally extensive (Fagan & Kareiva, 1997).

Although numerous authors have identified sample selection bias in presence-only datasets as a limitation to the accuracy of species distribution models and subsequent applications (Dennis & Thomas, 2000; Reddy & Dávalos, 2003; Araújo & Guisan, 2006; Hortal *et al.*, 2008; Boakes *et al.*, 2010; Peterson *et al.*, 2011), only more recently have solutions been explored for identifying (e.g. herbarium bias; Loiselle *et al.*, 2008) or accounting for effort bias within the modelling framework itself (Beale & Lennon, 2011; McInerny & Purves, 2011; Fourcade *et al.*, 2014). Schulman *et al.* (2007) generated a 'collecting activity landscape' (p. 1395) based on the density of collecting localities and collecting intensity at each locality for herbarium records in the Amazon. Single collecting localities were then used to modify estimates of species' ranges such that adjusted range estimates varied inversely with collecting activity. Bayesian approaches (Eddy, 2004) incorporating collection activity in species distribution model estimates have also been investigated. Latimer *et al.* (2006) and Royle *et al.* (2007) demonstrated a hierarchical framework incorporating spatially explicit details such as the complexity of irregular sampling intensity (termed 'spatial coverage bias' by Royle *et al.*, 2007) (see also Gelfand *et al.*, 2003; Argáez *et al.*, 2005; Latimer *et al.*, 2006; Ward *et al.*, 2009; Di Lorenzo *et al.*, 2011; Hui *et al.*, 2011; Golini, 2012). Bias correction features are available in Maxent, the widely used, machine learning approach to species distribution modelling that uses presence/background data to generate model estimates (Dudík *et al.*, 2004, 2007; Phillips *et al.*, 2004, 2006). Both options take into account sampling effort, which is then used to bias the selection of the background points (Phillips *et al.*, 2009). The first approach, the 'bias-file' feature, can be used to input a layer representing sampling effort (i.e. relative collection intensity). Instead of randomly selecting the background points from the study area, this option allows the program to select points that reflect the sampling distribution of collection effort; thus, selection of background points is subjected to the same bias as the occurrence data. Syfert *et al.* (2013) demonstrated the importance and efficacy of the bias-file option within a Maxent framework and found that correcting for bias had more of an impact on model fit than the type of response curves permitted to train the model. The second approach uses the occurrence localities of 'target groups' (Phillips *et al.*, 2009, p. 181) as the background points when modelling the distribution of a species within that particular target group. This approach assumes presence data for all species within a target group are subject to the same effort bias.

Similarly, Zaniewski *et al.* (2002) assessed the effects of environmentally weighted distribution of pseudo-absences on model estimates. Weighted pseudo-absence points performed comparably to presence/absence estimates from generalized additive modelling (GAM) (Hastie & Tibshirani, 1986) and better than those from randomly generated pseudo-absence points or using an ecological niche factor analysis (ENFA) approach (Hirzel *et al.*, 2002). In a related study, Engler *et al.* (2004) used ENFA to estimate environmentally weighted pseudo-absence data representing locations more likely to contain true absences. They found this method improved model predictions obtained from a generalized linear model over those estimated using random pseudo-absence points or via the ENFA inferred from presences only. Lobo *et al.* (2010) presented a continuation of this approach within a presence/absence framework using modelled estimates of sampling bias to generate pseudo-absences for GAM.

Despite their associated modelling challenges, presence-only data represent a vast source of information on species' distributions and biodiversity, especially when efforts to resample an area in a systematic manner would be too costly and/or labour-intensive. Online digital versions of presence-only data [e.g. Global Biodiversity Information Facility (GBIF)] are also becoming increasingly accessible (Graham *et al.*, 2004; Peterson *et al.*, 2011). For instance, the GBIF contains over 388 million biodiversity records from natural history museum collections, insect collections, herbaria and bird surveys with more than 340 million records georeferenced (GBIF, 2012), the majority of which are subject to effort bias (Otegui *et al.*, 2013). There remains a need, however, to further explore methods of accounting for sample selection bias in a species distribution modelling algorithm that explicitly adjusts model predictions within a more well-known modelling framework. Providing additional avenues to resolving this issue would represent an important step towards better using these data for exploring questions related to spatial patterns of biodiversity.

Given (1) the need to further address sample selection bias in presence-only data sets used in species distribution

modelling, (2) the extensive amount of presence-only datasets available world-wide, and (3) the increasing need for accurate species distribution models for resource management, the objectives of this study were to develop and examine the effectiveness of a framework of applying sample weighting based on collection effort to account for sample selection bias of presence-only data in species distribution models.

METHODS

Estimating sample selection bias

The first step in accounting for sample selection bias was to parameterize sampling effort. This was estimated in a manner analogous to modelling a species' distribution, but instead considered 'effort' as a function of variables that influence where experts and citizen scientists are likely to search for particular species [e.g. target groups in Maxent (Phillips *et al.*, 2009)]. For this study, we used occurrence data from the Alberta Conservation Information Management System (ACIMS) database, which tracks biodiversity in the province of Alberta, Canada (Fig. 1) (Government of Alberta, 2012). Based on occurrence localities of rare (i.e. NatureServe S1- to S3-ranked; NatureServe, 2013) species of vascular plants, bryophytes and butterflies, the probability of sampling was estimated for each taxonomic group across Alberta using logistic regression (Keating & Cherry, 2004).

Occurrence data (as polygon shapefiles) were acquired from ACIMS (Government of Alberta, 2012), and the centroid of each polygon was calculated to obtain point location data (see e.g. Guisan & Zimmermann, 2000). Predictor variables used to estimate sampling effort included road density, Euclidean distance to roads, herbaria and insect collections (i.e. locations of experts) and resource extraction sites [as rare plant inventories are often conducted prior to extraction activity in Alberta (PERG, 2009)], population density (Center for International Earth Science Information Network, International Food Policy Research Institute, the World Bank and Centro Internacional de Agricultura Tropical, 2004), terrain ruggedness (Evans, 2004) and protection status (binary) (Government of Canada, 2008) (see Fig. S1 in Supporting

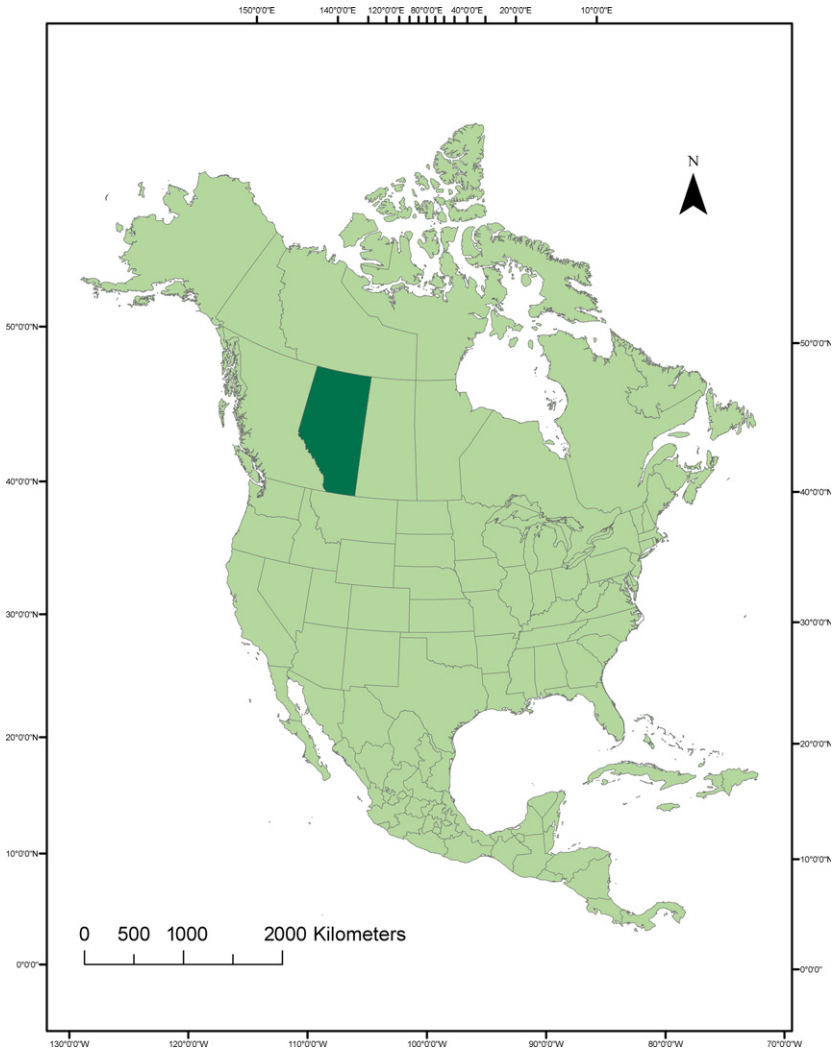

Figure 1 Orientation of Alberta (dark green), Canada, within North America.

Information). All of these factors were independent of the environmental data later used to model species' distributions. A principal component analysis was conducted to remove potential correlations between predictor variables, which were stored as raster data at a resolution of 250 m. Rivers, lakes and aquatic environments (Alberta Biodiversity Monitoring Institute, 2012) were omitted from the analysis by applying a land mask to the variables. Random background points [n = 10,000, as recommended by Barbet-Massin *et al.* (2012)] were generated using the Geospatial Modelling Environment (GME) (Beyer, 2012). All spatial information and statistical analyses were processed using ArcMap 10 (ESRI, 2011) and Stata 11 (StataCorp, 2009). See Fig. S2, Table S1, and Appendix S1 for maps, model summaries, and further discussion of sampling effort, respectively.

Simulating truth in GIS: Virtual species' distributions

As model validation requires an independent testing dataset (Peterson *et al.*, 2011), evaluation of model performance of bias-corrected model predictions requires a testing dataset to be free of bias (e.g. derived instead from random, systematic or stratified sampling). As we acknowledge that our ACIMS occurrence data were spatially biased in the first place (see above), simulated species' distributions were required to assess this bias correction method. Several authors have generated virtual species' distributions to examine key methodological questions related to species distribution modelling (Hirzel *et al.*, 2001; Dudík *et al.*, 2005; Elith & Graham, 2009; Phillips *et al.*, 2009; Jiménez-Valverde, 2012; Li & Guo, 2013). See Appendix S2, Fig. S3 and Table S2 for detailed methods on this procedure. This 'virtual ecologist' approach (reviewed by Miller, 2014) thereby allowed the 'true' species' distributions to be known, such that (1) random, independent test points could be sampled to evaluate model performance, and (2) subsequent model performance of biased and bias-adjusted models could be compared with 'truth' (i.e. random case).

Generating binary predictions and spatially biased samples

As 'truth' was known (i.e. simulated) and sampling effort had been estimated using logistic regression (i.e. effort models), we were able to subsample each virtual species in a manner that reflected the effort bias in the ACIMS database. First, we used Geospatial Modelling Environment (GME) software (Beyer, 2012) to generate 10,000 random points in Alberta (one separate set for each taxonomic group) and to query the value of the probability of presence at each point. From these queried probabilities, binary predictions were generated for three relative occurrence thresholds: common ($\mu-\sigma$), rare ($\mu+\sigma$) and very rare ($\mu+2\sigma$), where μ = mean and σ = SD of the probability of presence for each individual taxon. Thus, presence/absence data were simulated for three virtual species for each taxon (i.e. one at each relative occurrence threshold; nine species in total; Fig. S4). As a presence/background modelling framework was desired for this study as a representative modelling approach for presence-only data, absence data were omitted and three new sets of random points (n = 10,000) were generated as background data using GME.

Next, presence data were sampled in a biased manner such that the probability of retaining an occurrence location was based on the probability of sampling effort modelled for each taxon. Biased sampling was performed using the 'rbinom' command in the statistical software package, R (R Core Team, 2012). A random value was generated for each point from a binomial distribution subject to the probability value of the effort model at that location (UCLA Statistical Consulting Group, 2012). This step represented collecting location data for species, as is the case, for example, with data for herbarium and museum collections. This biased sampling procedure was replicated 100 times for each of the nine virtual species – three taxonomic groups at three relative occurrence levels – with each replicate representing a unique set of biased sampling points.

Species distribution modelling: Biased, adjusted and 'truth' models

Distribution of each species was estimated using logistic regression in Stata (StataCorp, 2009). The same predictor variables that were used to generate the virtual species' distributions, including any squared terms and/or interactions, were used to estimate their distribution from biased samples. Although this information is not available *a priori* under normal circumstances other than following traditional model selection procedures, it was used to examine the current method of improving model estimates by accounting for effort bias, not to study how bias may affect model selection.

Background points (n = 10,000) were selected at random as indicated in the previous section. Therefore, the training data consisted of presence locations, sampled in a biased manner to reflect the effort model, and 10,000 background points. Known presence locations were omitted if not selected as part of the biased sample. Logistic regression was used to estimate the distribution of each species under two scenarios: a biased (naïve) model and a bias-adjusted model. The latter was achieved using the 'pweight' (prior sample weights) option in Stata, which adjusts model estimates by weighting sample points according to the inverse of the probability of their inclusion in the sample (i.e. the effort model in this case) (StataCorp, 2012). As a result, biased locations were weighted, while the randomly sampled background points were not adjusted. Frair *et al.* (2004) were successful in employing a sample weighting approach to correct GPS collar bias in resource selection function models. This procedure was run for 100 iterations of each method (i.e. for each replicate of biased presences) to preclude results due to chance.

Additionally, 'truth' was modelled for each species using unbiased data. In this scenario, the known presence locations generated by applying relative occurrence thresholds to

random locations (i.e. prior to biased sub-sampling) and the random background locations were modelled using logistic regression with the same corresponding set of environmental predictors as the biased and pweight-adjusted models for each of the nine species. Only one iteration was required for these truth models. It is important to distinguish modelling 'truth' in this step from the initial simulation of 'truth' outlined in the two previous sections. In the current step, 'truth' was modelled as a reference when evaluating the effectiveness of the bias adjustment method described in the following section (i.e. comparison of regression coefficients, model performance).

Model evaluation: Was bias corrected?

As the goal of this study was to demonstrate that prior sample weights can be used directly in a modelling algorithm to correct for spatial biases in sampling effort, we compared model performance – area under the curve of the receiver operating characteristic (ROC AUC) (Swets, 1988), sensitivity and specificity (Allouche *et al.*, 2006) – and regression coefficients across all three scenarios: biased, pweight-adjusted and 'truth'. Miller (2014) highlighted the necessity of virtual species in determining whether a given modelling approach yields biased estimates, as indicated by significant differences between regression coefficients for truth and the model in question. Pearson's product moment correlation coefficients were also calculated as indices of similarity between model outputs (biased, pweight-adjusted) and truth over 10,000 random points on the landscape. See Appendix S3 for details on the model testing procedure.

The mean of regression coefficients, correlation coefficients, AUC, sensitivity and specificity was calculated over all 100 iterations for the biased and pweight-adjusted scenarios. These values were compared with the truth scenario, for which only one set of model evaluation statistics was required as a further subset of points had not been sampled (see previous section). Therefore, we used each truth layer for (1) an independent testing set from which random locations could be sampled for model testing, and (2) unbiased models to facilitate a comparison of regression and correlation coefficients for our biased and pweight-adjusted models.

Comparison with Maxent: Bias file and target-group background approaches

We also compared the prior-weight adjustment method in logistic regression with the bias correction options available in Maxent. For this analysis, we used the same occurrence data (either random or biased) and environmental predictor variables as described in the previous sections for logistic regression. Species were modelled under four scenarios: truth (i.e. random), biased, bias-file adjustment (using the same effort models as were used in the prior-weight adjustment approach) and target-group background adjustment (pooling all biased data points for a given taxonomic group). In all cases, linear, quadratic and product features were allowed for model parameterization, the logistic output format was selected, and regularization parameters were set to default. Model evaluation, comprising comparisons of correlation coefficients, AUC, sensitivity and specificity, was carried out in the same manner as described in the previous section for logistic regression.

RESULTS

Model estimates: Effects of effort bias and prior-weight adjustment on logistic regression coefficients

Overall, logistic regression coefficients estimated from biased data were significantly different (z-test; $P < 0.0001$) from those obtained via random sampling (Table 1; Tables S3 & S4; Figs S5–S7). In contrast, regression coefficients for pweight-adjusted models did not differ significantly from truth (at $\alpha/2 = 0.025$ for two-way comparison of z-tests). Minor exceptions to this trend for prior-weight adjustments were observed for the common butterfly species, where significant differences from truth were observed for β_o weights of mean annual temperature ($z = 3.273$, $P < 0.001$) and one category of natural region ($z = 4.319$, $P < 0.0001$), as well as the y-intercept ($z = -4.260$, $P < 0.0001$). Intercept terms were, however, consistently lowest for biased models, thereby resulting in lower overall predicted probabilities as compared with pweight-adjusted and truth models (Fig. 2; Figs S8 & S9) and decreased model sensitivity (see below). Fundamentally, regression coefficients and resulting spatial distributions were more consistent between truth and bias-adjusted models than biased models. Truth was thus effectively 'recovered' using the prior-weight adjustment method.

Across all three taxa, biased model estimates of regression coefficients exhibited relatively greater deviation from truth as relative occurrence increased. For instance, anomalies in biased coefficient estimates for common species ranged from −466% to 104% for vascular plants, −9863% to 144% for bryophytes and −2323% to 2153% for butterflies. In contrast, deviation of biased coefficient estimates from truth for very rare species ranged from −49% to −2% for vascular plants, −36% to −3% for bryophytes and −112% to −13% for butterflies. Furthermore, deviation between pweight-adjusted estimates and truth at common relative occurrence ranged from −9% to 1% for vascular plants, −34% to 10% for bryophytes and −58% to 100% for butterflies. At very rare relative occurrence for prior-weight adjustment, deviation from truth was even less with −5% to 1% for vascular plants, −4% to 1% for bryophytes and −88% to −2% for butterflies. Thus, both the requirement for bias correction and the resulting effect of prior-weight adjustment increased with relative occurrence.

Model evaluation and comparison with Maxent

For both the logistic regression and Maxent approaches, sensitivity among biased models was significantly lower

Table 1 Model summaries from logistic regression analysis of the distribution of virtual vascular plant species at common, rare and very rare relative occurrence levels across three different modelling approaches [random (i.e. truth), biased and pweight adjusted].

	Common			Rare			Very rare		
Variable	Truth	Biased	Adjusted	Truth	Biased	Adjusted	Truth	Biased	Adjusted
CTI	0.067 (0.016)	−0.025 (0.002)	0.062 (0.003)	0.315 (0.029)	0.241 (0.003)	0.314 (0.003)	0.603 (0.093)	0.458 (0.011)	0.575 (0.011)
MAP	0.196 (0.019)	0.400 (0.001)	0.195 (0.002)	1.452 (0.056)	1.162 (0.003)	1.451 (0.004)	1.736 (0.139)	1.703 (0.007)	1.736 (0.008)
MAT	0.425 (0.018)	0.647 (0.001)	0.428 (0.003)	2.449 (0.069)	1.993 (0.003)	2.449 (0.004)	3.749 (0.262)	3.549 (0.016)	3.802 (0.018)
MAT × MAP	0.057 (0.015)	0.013 (0.001)	0.058 (0.001)	0.571 (0.045)	0.361 (0.003)	0.575 (0.004)	0.382 (0.094)	0.196 (0.005)	0.382 (0.007)
y-int	−0.281 (0.016)	−1.591 (0.002)	−0.284 (0.003)	−3.620 (0.077)	−4.139 (0.004)	−3.620 (0.005)	−9.258 (0.498)	−9.618 (0.027)	−9.349 (0.032)
n	18,177	12,543	12,543	11,743	10,688	10,688	10,128	10,066	10,066

SEs (indicated in brackets) are of regression coefficient estimates for truth, whereas regression coefficients and SE of the mean are averaged over 100 runs (each representing a separate biased sample) for the latter two approaches (biased and adjusted). Mapped outputs of predicted values from these results are shown in Fig. 2.
CTI, compound topographic index; MAP, mean annual precipitation; MAT, mean annual temperature.

(one-sample *z*-test; $P < 0.05$), respectively, for all species as compared with truth (Table 2; Tables S5 & S6). Furthermore, decreased sensitivity among biased models increased with relative occurrence across all taxa. In general, sensitivity and specificity among adjusted models (either prior-weight or bias-file adjustment) were not significantly different from truth as compared with biased models at $\alpha/2 = 0.025$ for two-way comparison (logistic regression) and $\alpha/3 = 0.0167$ for three-way comparison (Maxent) of *z*-tests. For the prior-weight adjustment approach in logistic regression, exceptions for which there was still a significant difference included sensitivity and specificity for the common butterfly species and very rare bryophyte species. For the bias-file approach in Maxent, exceptions for which there was still a significant difference included sensitivity for the very rare bryophyte and butterfly species and specificity for the rare bryophyte, common butterfly and common vascular species.

Area under the curve scores did not reveal consistent differences among the scenarios (Table 2; Tables S5 & S6), with nearly all models exhibiting very good discrimination for each of the scenarios (Pearce & Ferrier, 2000). The only exceptions were the target-group background models for common bryophytes and butterflies, which are discussed below. More meaningfully, however, with respect to examining the similarity of model outputs, pweight and bias-file adjusted models were more significantly correlated with truth than those obtained for biased models (Table 2; Tables S5 & S6; Fig. 3; Figs S10 & S11). Two exceptions here were for the bias-file adjustments of very rare bryophyte and vascular plant species, which were marginally less significantly correlated with truth than the biased case.

In our experiment, the target-group background approach in Maxent yielded inconsistent impacts on model predictions. In some cases, this approach resulted in similar model adjustments as the bias file (e.g. common vascular species; Table 2; Fig. 3). Conversely, for some species (e.g. common bryophyte and common butterfly species; Tables S5 & S6; Figs S10 & S11), the target-group background adjustment yielded model estimates less consistent with truth than the biased scenario.

DISCUSSION

Assessing the effectiveness of bias-corrected presence-only species distribution models

Bias-corrected models based on prior weights of relative sample probabilities correctly estimated the true logistic regression coefficients of the virtual species' distributions, whereas unadjusted models demonstrated biases in regression coefficients and notably the *y*-intercept, which was consistently lower as compared with that of the 'truth' model, and thus lower overall predicted probabilities. This denotes a general decrease in sensitivity (i.e. higher false negative rate) in models estimated from spatially biased occurrence data (as reported by Tyre *et al.*, 2003; Botts *et al.*, 2011; Hanspach

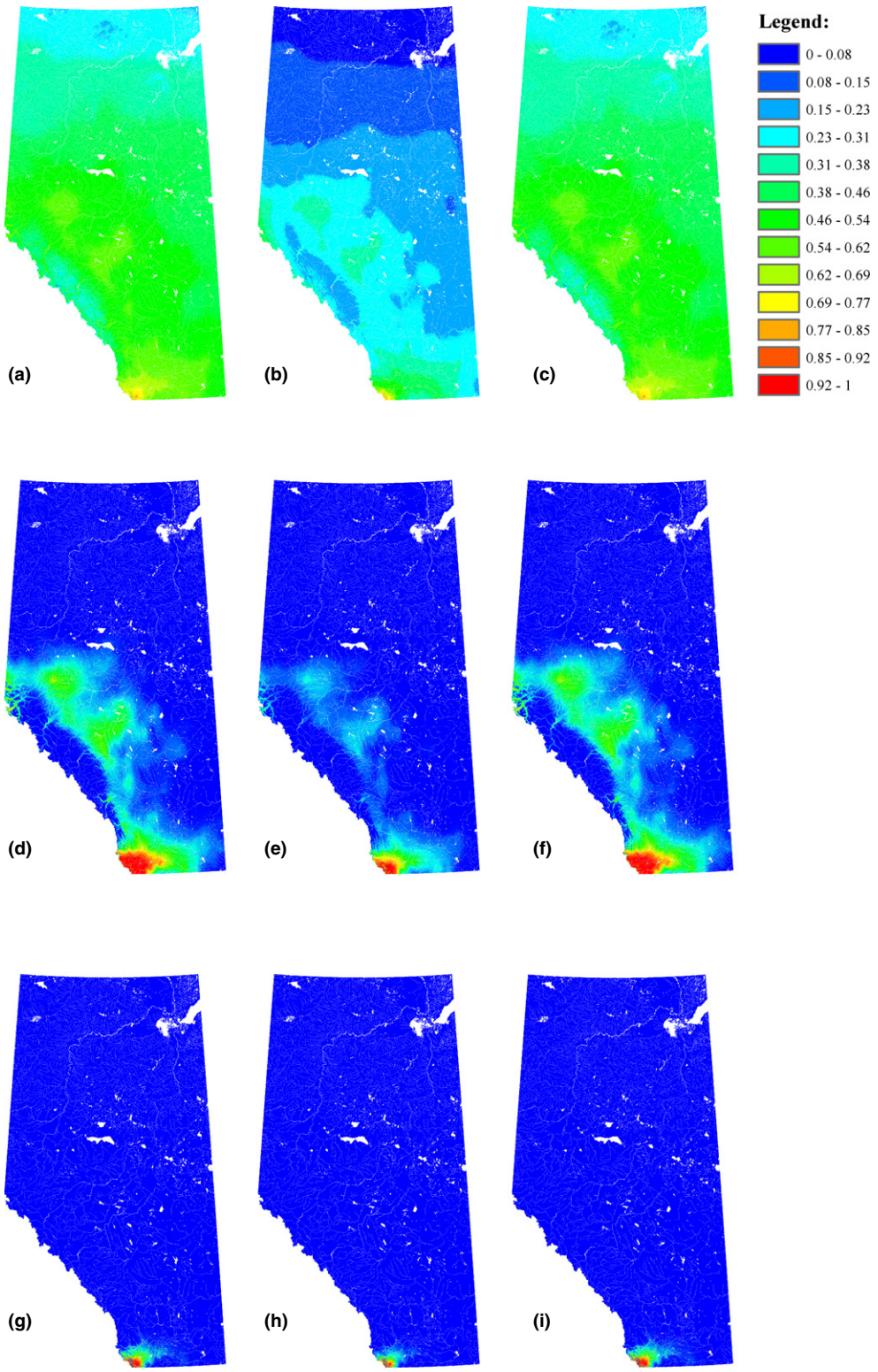

Figure 2 Spatial distributions of logistic regression estimates for a common (a–c), rare (d–f) and very rare (g–i) virtual vascular plant species under random, biased and pweight-adjusted (left to right) modelling scenarios. Warmer colours indicate higher probability of suitable habitat. See Table 1 for additional summary of model outputs and Table 2 for model evaluation.

et al., 2011) and the need for bias correction in interpreting environmental relationships using presence-only data. This was indeed the effect of biased sampling effort on model performance, as indicated by the significant decrease in sensitivity across all species, which was indicative of under-prediction of presences (Pearson, 2007). For the models

Table 2 Model evaluation metrics for virtual vascular plant species at common, rare and very rare relative occurrence levels under different scenarios (logistic regression GLM: truth, biased and pweight-adjusted; Maxent: truth, biased, bias-file adjusted and target-group background adjusted).

		GLM			Maxent			
Vascular		Truth	Biased	Adjusted	Truth	Biased	Bias-file adj.	TGB adj.
Common	ROC AUC	0.9973 (0.0016)	**0.9238 (0.0003)**	0.9950 (0.0004)	0.9963 (0.0027)	**0.9600 (0.0004)**	0.9904 (0.0002)	**0.9834 (0.0003)**
	Sensitivity	0.9600 (0.0196)	**0.0599 (0.0003)**	0.9549 (0.0039)	0.9700 (0.0171)	**0.8080 (0.0006)**	0.9756 (0.0013)	**0.9152 (0.0019)**
	Specificity	0.9600 (0.0196)	0.9934 (0.0009)	0.9237 (0.0095)	0.9700 (0.0171)	**0.9253 (0.0009)**	**0.8651 (0.0045)**	0.9498 (0.0031)
	Pearson's *r*		0.9135 (0.0008)	0.9965 (0.0003)		0.9101 (0.0006)	0.9852 (0.0003)	0.9406 (0.001)
Rare	ROC AUC	0.9997 (0.0003)	**0.9988 (0.0001)**	0.9997 (< 0.0001)	0.9985 (0.0011)	0.9975 (0.0001)	0.9984 (0.0001)	0.9974 (0.0001)
	Sensitivity	0.9900 (0.0099)	**0.6485 (0.0338)**	0.9908 (0.0007)	0.9800 (0.0140)	**0.9342 (0.0014)**	0.9792 (0.0012)	0.974 (0.0013)
	Specificity	0.9900 (0.0099)	0.9939 (0.0007)	0.9899 (0.0009)	0.9800 (0.0140)	0.9899 (0.0002)	0.9803 (0.0002)	0.9628 (0.0006)
	Pearson's *r*		0.9492 (0.0003)	0.9993 (0.0001)		0.9954 (0.0001)	0.9973 (0.0001)	0.9883 (0.0002)
Very rare	ROC AUC	0.9993 (0.0006)	0.9985 (0.0001)	0.9991 (0.0001)	0.9999 (0.0001)	**0.9990 (0.0001)**	**0.9982 (0.0001)**	**0.992 (0.0001)**
	Sensitivity	0.9900 (0.0099)	**0.9198 (0.0026)**	0.9789 (0.0013)	0.9900 (0.0099)	**0.9644 (0.0017)**	0.9909 (0.0009)	0.9781 (0.0022)
	Specificity	0.9900 (0.0099)	0.9927 (0.0004)	0.9825 (0.0009)	0.9900 (0.0099)	0.9899 (0.0002)	0.9688 (0.0004)	**0.9616 (0.0004)**
	Pearson's *r*		0.9717 (0.0009)	0.9968 (0.0004)		0.9803 (0.0005)	0.9425 (0.0012)	0.8589 (0.0016)

Other than for truth, SEs (indicated in brackets) represent SE of the estimate averaged over 100 runs (each representing a separate biased sample). 200 independent test points (100 presences and 100 absences) were used to calculate ROC AUC, sensitivity and specificity (significant differences indicated in bold type). Pearson's *r* correlation coefficient (comparing with truth) was calculated from 10,000 random points across the landscape. All correlations are significant at $\alpha = 0.05$ (d.f. = 9998).

presented here, this observation is supported by lower regression coefficients and particularly lower β_o values observed for the biased models as compared with truth, as indicated above. Despite higher logistic regression coefficients for some environmental predictors, the drastically lower *y*-intercept in each biased case was sufficient to yield under-predicted probabilities of presence of suitable habitat among biased models. As the majority of biased predicted values were therefore absences, the false positive rate was lower and thus the true negative rate (specificity) increased.

Asymmetry in biased model prediction errors also supports the assessment of model performance using sensitivity, specificity and their derivatives [(e.g. true skill statistic (Allouche *et al.*, 2006)] instead of the threshold-independent AUC. As emphasized by Lobo *et al.* (2008) and Jiménez-Valverde (2012), the appropriateness and usefulness of AUC as an evaluation metric in species distribution modelling is highly dependent on modelling objectives and context. In particular, the use of AUC with presence/background evaluation data has been discouraged since doing so violates AUC theory (Jiménez-Valverde, 2012; Miller, 2014) and inflates both the number of false absences (Lobo *et al.*, 2008) and predictive accuracy for rare species (Boyce *et al.*, 2002; Phillips *et al.*, 2009; see below for discussion of relative occurrence area). AUC is not necessarily an indicator of goodness-of-fit, but is rather a discrimination index (Lobo *et al.*, 2008). As a result, it does not take into account probability values predicted by a given model, but rather indicates whether the model predicts a given presence to have a higher probability value than an absence. Thus, discriminatory power (i.e. high AUC score) is not necessarily indicative of a well-fitted model and vice versa (Lobo *et al.*, 2008).

Although we applied AUC evaluation using presence/absence test data in this case, the ability of the model to discriminate correctly across all thresholds was not matched by our results from the Pearson correlation analysis of the true similarity of model outputs. Similar results were reported by Fourcade *et al.* (2014), who evaluated their models using AUC and three measures of correlation between model outputs. They found AUC to be an unreliable metric for detecting model accuracy, noting that low ΔAUC between biased and bias-corrected models was not reflected in a correspondingly high degree of correlation between the two model types. Similarly to our correlation results, Fourcade *et al.* (2014) also noted that directly comparing model outputs (via overlap in environmental and geographic space) was more meaningful than AUC in evaluating the effectiveness of each bias correction method in recovering unbiased estimates of the species' known distributions.

For our results, the importance of using bias-adjusted models of species' distributions varied depending on their relative occurrence area (i.e. ratio between species' occurrence and study extent) (Lobo *et al.*, 2008) with the effect being most pronounced for common species. This suggests that spatially biased occurrence data may not be as

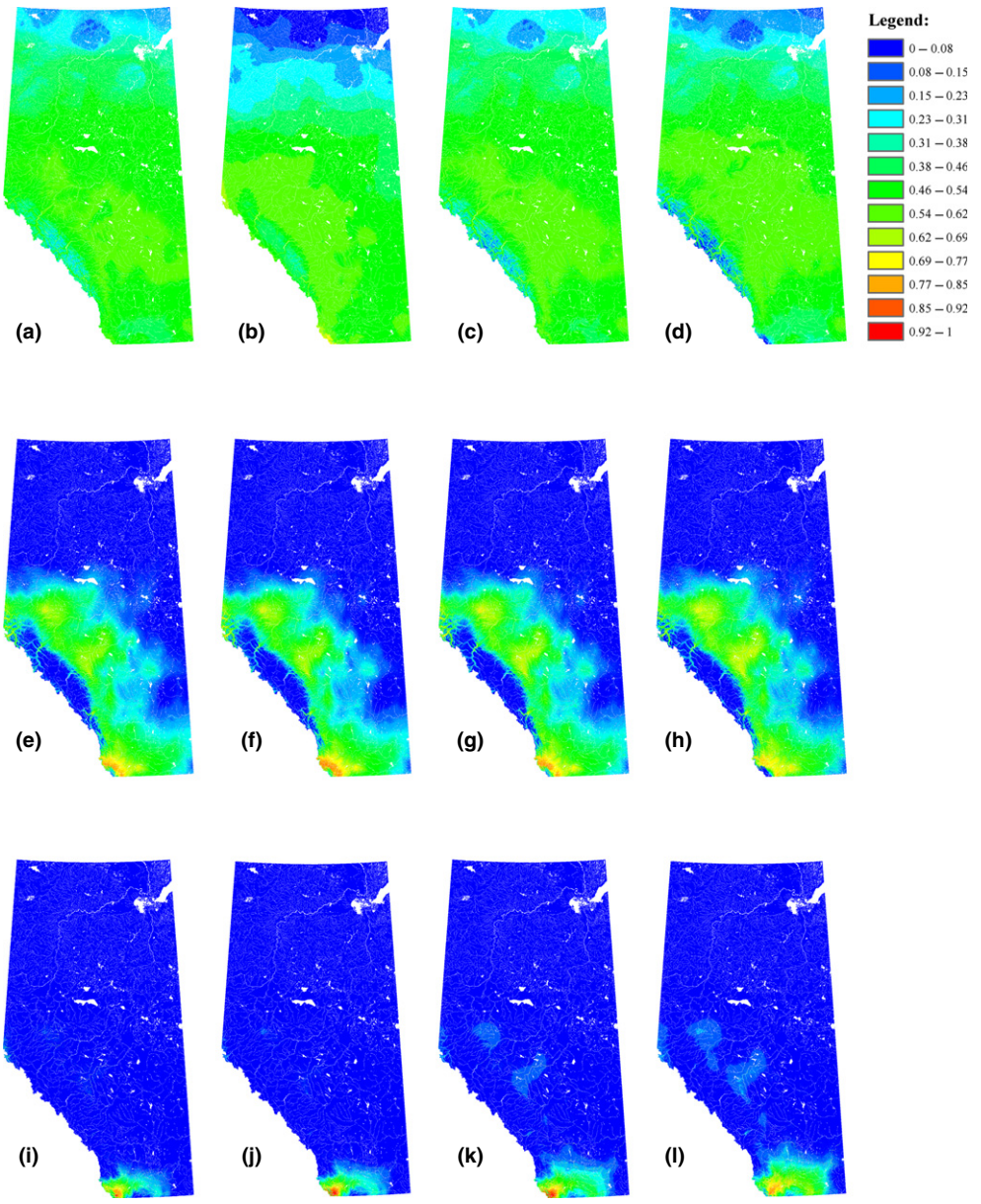

Figure 3 Spatial distributions of Maxent estimates for a common (a–d), rare (e–h) and very rare (i–l) virtual vascular plant species under random, biased, bias-file adjusted and target-group background adjusted (left to right) modelling scenarios. Warmer colours indicate higher probability of suitable habitat. See Table 2 for model evaluation.

problematic for rarer species, especially those that are habitat specialists (see also Phillips *et al.*, 2009; Chefaoui *et al.*, 2011). In our study, this tendency likely had less to do with contamination of background points, which was calculated to be < 1% in each case, and can be further explained by the fact that imposing our effort bias on common species removed a greater proportion of presences as compared with rare and very rare species. As emphasized by Lobo *et al.* (2008), different species occupy different proportions of a given study extent (i.e. relative occurrence area). Thus, species with a lower relative occurrence area will have a proportionately larger number of absences in the region from which training and testing data are drawn. As a result, AUC and specificity scores (as well as Pearson's correlation comparisons presented here) become artificially inflated for relatively rarer species, which appear to be more accurately modelled than their counterparts with a greater relative occurrence area. To this end, it is thus not entirely consequential to make generalizations about the relative impacts of bias and the requisite for bias correction based on relative occurrence area. Li & Guo (2013) recently proposed alternative accuracy metrics (based on the *F*-statistic) for use with presence/background test data that also facilitate threshold selection and estimation of relative occurrence area.

Tackling effort bias in species distribution modelling

Analytical techniques are needed to account for sample selection bias in presence-only modelling of species' distributions providing greater utility of increasingly accessible online presence data (Graham *et al.*, 2004; Peterson *et al.*, 2011). Our method is similar to that presented by Schulman *et al.* (2007), who adjusted predicted values of species' distributions to vary inversely with sampling intensity by down-weighting samples from locations of with higher sampling intensity. Our approach has expanded on this method to directly adjust estimates of species' distributions within the statistical model. Phillips *et al.* (2009) suggested such approaches when dealing with spatially biased occurrence records, but noted the difficultly in implementing it when sampling effort is unknown. We have demonstrated a technique for both estimating sampling effort and using that information to adjust resulting model parameters, inferences and model predictions.

Bias adjustment methods using prior weights of sampling effort presented here are conceptually similar to the FactorBiasOut technique proposed for Maxent (Dudík *et al.*, 2005). Their approach estimates the product of the probabilities of true species' distribution and sample selection distribution and subsequently uses Kullback–Leibler divergence (also known as relative entropy) to estimate and factor out the sample selection distribution, thereby approximating the species' distribution. The target-group background approach was presented as an approximation of this method (e.g. Ponder *et al.*, 2001; Anderson, 2003) and further explored by Phillips *et al.* (2009). They identified target groups as species within the same broad taxonomic groups (e.g. vascular plants, birds), for which collector expertise, method of collection and, subsequently, distribution of sampling effort would be the same – analogous to our approach of modelling effort for broad taxonomic groups. In their experiment, use of the target-group background samples improved otherwise biased estimates of species' distributions, as measured by AUC. Their results suggested that previous discrepancies between model performance reported by Elith *et al.* (2006) were attributable to the presence of biased data in the models such that differences attributed to modelling algorithm were negligible once sample selection bias was accounted for using target-group background points. Our results did not show the same level of consistent improvement when applying the target-group background approach (see also Fourcade *et al.*, 2014). Rather, our application of the Maxent bias file using modelled sampling effort to bias background points yielded comparatively better bias correction. In cases where Maxent is a more appropriate modelling algorithm than logistic regression [e.g. for smaller sample sizes (Pearson *et al.*, 2007)], we recommend using the bias-file adjustment as described here, particularly when the target group is not necessarily large enough to yield an adequate number of background points.

As target-group background points can also include the occurrence points of the species being modelled, Phillips *et al.* (2009) also examined the effects of removing the 'contaminated' background points [called 'non-overlapping background' (p. 195)] from the model and found no difference in model performance. In light of the effectiveness of their target-group background approach to dealing with biased sampling effort, they noted their predictions represented the realized species' distributions and cautioned against extrapolating model estimates to other regions or environmental conditions not surveyed as part of the analysis. The latter stipulation also applies to the sample weighting approach presented here. Model adjustments based on the spatial distribution of sampling were specific to the region for which effort models were estimated.

CONCLUDING REMARKS

We saw the potential to further explore the issue of spatially biased occurrence data in a well-known and highly flexible GLM framework. This technique offers a potential alternative to dealing with sample selection bias and thereby accounting for one of several sources of uncertainty in species distribution modelling (Beale & Lennon, 2011; Rocchini *et al.*, 2011). Given that prior-weight adjusted models had improved model predictive accuracy and estimation of regression coefficients and mapped predictions, we offer this method for consideration when using presence-only data for species distribution modelling. The method we have presented here is potentially widely applicable due to the wealth of information available from online biodiversity databases and the flexibility of the approach.

ACKNOWLEDGEMENTS

Funding for this study was provided by the Natural Science and Engineering Research Council of Canada (NSERC Alexander Graham Bell Canada Graduate Scholarship), the University of Alberta (Faculty of Graduate Studies and Research), the Foundation for Orchid Research and Conservation and the Climate Change and Emissions Management Corporation through the Alberta Biodiversity Monitoring Institute. We further thank Drs Guillaume Blanchet and David Roberts for their assistance with R programming language and the four anonymous referees whose helpful suggestions improved this manuscript.

REFERENCES

Alberta Biodiversity Monitoring Institute (2012) ABMI Wall-to-wall land cover map. Available at: http://www.abmi.ca/abmi/rawdata/geospatial/gisdownload.jsp?categoryId=2&subcategoryId=4 (accessed 12 August 2014).

Allouche, O., Tsoar, A. & Kadmon, R. (2006) Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS). *Journal of Applied Ecology*, **43**, 1223–1232.

Anderson, R.P. (2003) Real vs. artefactual absences in species distributions: tests for *Oryzomys albigularis* (Rodentia: Muridae) in Venezuela. *Journal of Biogeography*, **30**, 591–605.

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

Argáez, J.A., Christen, J.A., Nakamura, M. & Soberón, J. (2005) Prediction of potential areas of species distributions based on presence-only data. *Environmental and Ecological Statistics*, **12**, 27–44.

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

Barbet-Massin, M., Jiguet, F., Albert, C.H. & Thuiller, W. (2012) Selecting pseudo-absences for species distribution models: how, where and how many? *Methods in Ecology and Evolution*, **3**, 327–338.

Beale, C.M. & Lennon, J.J. (2011) Incorporating uncertainty in predictive species distribution modelling. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **367**, 247–258.

Beyer, H.L. (2012) Geospatial modelling environment. Spatial Ecology LLC. Available at: http://www.spatialecology.com/gme/ (accessed 14 December 2012).

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

Botts, E.A., Erasmus, B.F.N. & Alexander, G.J. (2011) Geographic sampling bias in the South African Frog Atlas Project: implications for conservation planning. *Biodiversity and Conservation*, **20**, 119–139.

Boyce, M.S., Vernier, P.R., Nielsen, S.E. & Schmiegelow, F.K.A. (2002) Evaluating resource selection functions. *Ecological Modelling*, **157**, 281–300.

Boyd, D.R. (2003) *Unnatural law: rethinking Canadian environmental law and policy*. UBC Press, Vancouver, BC.

Center for International Earth Science Information Network (CIESIN), International Food Policy Research Institute (IFPRI), the World Bank & Centro Internacional de Agricultura Tropical (CIAT) (2004) *Global Rural-Urban Mapping Project (GRUMP): urban/rural population grids*. CIESIN, Columbia University, Palisades, NY. Available at: http://sedac.ciesin.columbia.edu/data/collection/grump-v1 (accessed 12 August 2014).

Chefaoui, R.M., Lobo, J.M. & Hortal, J. (2011) Effects of species' traits and data characteristics on distribution models of threatened invertebrates. *Animal Biodiversity and Conservation*, **34**, 229–247.

Dennis, R. & Thomas, C. (2000) Bias in butterfly distribution maps: the influence of hot spots and recorder's home range. *Journal of Insect Conservation*, **4**, 73–77.

Di Lorenzo, B., Farcomeni, A. & Golini, N. (2011) A Bayesian model for presence-only semicontinuous data, with application to prediction of abundance of *Taxus Baccata* in two Italian regions. *Journal of Agricultural, Biological, and Environmental Statistics*, **16**, 339–356.

Dudík, M., Phillips, S.J. & Schapire, R.E. (2004) Performance guarantees for regularized maximum entropy density estimation. *Proceedings of the 17th Annual Conference on Computational Learning Theory*, pp. 472–486. ACM Press, New York.

Dudík, M., Phillips, S.J. & Schapire, R.E. (2005) Correcting sample selection bias in maximum entropy density estimation. *Advances in neural information processing systems 18* (ed. by Y. Weiss, B. Scholkopf and J.C. Platt), pp. 323–330. MIT Press, Cambridge, MA.

Dudík, M., Phillips, S.J. & Schapire, R.E. (2007) Maximum entropy density estimation with generalized regularization and an application to species distribution modeling. *Journal of Machine Learning Research*, **8**, 1217–1260.

Eddy, S.R. (2004) What is Bayesian statistics? *Nature Biotechnology*, **22**, 1177–1178.

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**, 66–77.

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., Graham, C.H., Anderson, R.P. *et al.* (2006) Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, **29**, 129–151.

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**, 263–274.

ESRI (2011) *ArcGIS desktop: release 10*. Environmental Systems Research Institute, Redlands, CA.

Evans, J. (2004) *Topographic ruggedness index*. Available at: http://arcscripts.esri.com/details.asp?dbid=12435 (accessed 12 August 2014).

Fagan, W.F. & Kareiva, P.M. (1997) Using compile species lists to make biodiversity comparisons among regions: a test case using Oregon butterflies. *Biological Conservation*, **80**, 249–259.

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

Frair, J.L., Nielsen, S.E., Merrill, E.H., Lele, S.R., Boyce, M.S., Munro, R.H.M., Stenhouse, G.B. & Beyer, H.L. (2004) Removing GPS collar bias in habitat selection studies. *Journal of Applied Ecology*, **41**, 202–212.

GBIF (2012) Global Biodiversity Information Facility. Available at: http://data.gbif.org/welcome.htm (accessed 14 December 2012).

Gelfand, A.E., Silander, J.A., Jr, Wu, S., Latimer, A., Lewis, P.O., Rebelo, A.G. & Holder, M. (2003) Explaining species distribution patterns through hierarchical modeling. *Bayesian Analysis*, **1**, 41–92.

Golini, N. (2012) *Bayesian modeling of presence-only data*. PhD Thesis. Università di Roma, Sapienza.

Government of Alberta (2012) Alberta Conservation Information Management System. Available at: http://www.albertaparks.ca/albertaparksca/management-land-use/alberta-conservation-information-management-system-(acims).aspx (accessed 31 August 2011).

Government of Canada, Natural Resources Canada, Canada Centre for Remote Sensing (2008) Atlas of Canada 1,000,000 national frameworks data, protected areas. Available at: http://www.geogratis.gc.ca/download/frameworkdata/protected_areas/ (accessed 12 August 2014).

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

Guisan, A. & Thuiller, W. (2005) Predicting species distribution: offering more than simple habitat models. *Ecology Letters*, **8**, 993–1009.

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

Hanspach, J., Kühn, I., Schweiger, O., Pompe, S. & Klotz, S. (2011) Geographical patterns in prediction errors of species distribution models. *Global Ecology and Biogeography*, **20**, 779–788.

Hastie, T. & Tibshirani, R. (1986) Generalized additive models. *Statistical Science*, **1**, 297–310.

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

Hirzel, A.H., Hausser, J., Chessel, D. & Perrin, N. (2002) Ecological-niche factor analysis: how to compute habitat-suitability maps without absence data? *Ecology*, **83**, 2027–2036.

Hortal, J., Jiménez-Valverde, A., Gómez, J.F., Lobo, J.M. & Baselga, A. (2008) Historical bias in biodiversity inventories affects the observed environmental niche of the species. *Oikos*, **117**, 847–858.

Hui, C., Foxcroft, L.C., Richardson, D.M. & MacFadyen, S. (2011) Defining optimal sampling effort for large-scale monitoring of invasive alien plants: a Bayesian method for estimating abundance and distribution. *Journal of Applied Ecology*, **48**, 768–776.

Jiménez-Valverde, A. (2012) Insights into the area under the receiver operating characteristic curve (AUC) as a discrimination measure in species distribution modelling. *Global Ecology and Biogeography*, **21**, 498–507.

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

Keating, K.A. & Cherry, S. (2004) Use and interpretation of logistic regression in habitat-selection studies. *Journal of Wildlife Management*, **68**, 774–789.

Latimer, A.M., Wu, S., Gelfand, A.E. & Silander, J.A., Jr (2006) Building statistical models to analyze species distributions. *Ecological Applications*, **16**, 33–50.

Li, W. & Guo, Q. (2013) How to assess the prediction accuracy of species presence-absence models without absence data? *Ecography*, **36**, 788–799.

Lobo, J.M., Jiménez-Valverde, A. & Real, R. (2008) AUC: a misleading measure of the performance of predictive distribution models. *Global Ecology and Biogeography*, **17**, 145–151.

Lobo, J.M., Jiménez-Valverde, A. & Hortal, J. (2010) The uncertain nature of absences and their importance in species distribution modelling. *Ecography*, **33**, 103–114.

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

McInerny, G.J. & Purves, D.W. (2011) Fine-scale environmental variation in species distribution modelling: regression dilution, latent variables and neighbourly advice. *Methods in Ecology and Evolution*, **2**, 248–257.

Miller, J.A. (2014) Virtual species distribution models: using simulated data to evaluate aspects of model performance. *Progress in Physical Geography*, **38**, 117–128.

NatureServe (2013) NatureServe conservation status. Available at: http://www.natureserve.org/explorer/ranking.htm (accessed 8 November 2013).

Otegui, J., Ariño, A.H., Encinas, M.A. & Pando, F. (2013) Assessing the primary data hosted by the Spanish Node of the Global Biodiversity Information Facility (GBIF). *PLoS One*, **8**, e55144.

Pearce, J. & Ferrier, S. (2000) Evaluating the predictive performance of habitat models developed using logistic regression. *Ecological Modelling*, **133**, 225–245.

Pearson, R.G. (2007) Species' distribution modeling for conservation educators and practitioners. Synthesis. American Museum of Natural History. Available at: http://ncep.amnh.org/resources.php?globalnav=resources&sectionnav=modules&sectionsubnav=module_files&module_id=361 (accessed 17 September 2014).

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.

PERG (2009) Peatlands and oil sands. Available at: http://www.gret-perg.ulaval.ca/no_cache/research/research-scope/peatlands-and-oil-sands/?tx_centrerecherche_pi1%5BshowUid%5D=770 (accessed 19 November 2014).

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

Phillips, S.J., Dudík, M. & Schapire, R.E. (2004) A maximum entropy approach to species distribution modeling. *Proceedings of the Twenty-First International Conference on Machine Learning* (ed. by R. Greiner and D. Schuurmans), pp. 655–662. ACM Press, New York.

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., Dudík, M., Elith, J., Graham, C.H., Lehmann, G.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.

Ponder, W.F., Carter, G.A., Flemons, P. & Chapman, R.R. (2001) Evaluation of museum collection data for use in biodiversity assessment. *Conservation Biology*, **15**, 648–657.

R Core Team (2012) *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. Available at: http://www.R-project.org/ (accessed 14 December 2012).

Reddy, S. & Dávalos, L.M. (2003) Geographical sampling bias and its implications for conservation priorities in Africa. *Journal of Biogeography*, **30**, 1719–1727.

Rocchini, D., Hortal, J., Lengyel, S., Lobo, J.M., Jiménez-Valverde, A., Ricotta, C., Bacaro, G. & Chiarucci, A. (2011) Accounting for uncertainty when mapping species distributions: the need for maps of ignorance. *Progress in Physical Geography*, **35**, 211–226.

Royle, J.A., Kéry, M., Gautier, R. & Schmid, H. (2007) Hierarchical spatial models of abundance and occurrence from imperfect survey data. *Ecological Monographs*, **77**, 465–481.

Schulman, L., Toivonen, T. & Ruokolainen, K. (2007) Analysing botanical collecting effort in Amazonia and correcting for it in species range estimation. *Journal of Biogeography*, **34**, 1388–1399.

StataCorp (2009) *Stata statistical software: release 11*. StataCorp LP, College Station, TX.

StataCorp (2012) Stata 12 help for pweight. Available at: http://www.stata.com/help.cgi?pweight (accessed 14 December 2012).

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

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

Tyre, A.J., Tenhumberg, B., Field, S.A., Niejalke, D., Parris, K. & Possingham, H.P. (2003) Improving precision and reducing bias in biological surveys: estimating false-negative error rates. *Ecological Applications*, **13**, 1790–1801.

UCLA Statistical Consulting Group (2012) R learning module: probabilities and distributions. Available at: http://www.ats.ucla.edu/stat/r/modules/prob_dist.htm (accessed 29 October 2012).

Ward, G., Hastie, T., Barry, S., Elith, J. & Leathwick, J.R. (2009) Presence-only data and the EM algorithm. *Biometrics*, **65**, 554–563.

Zaniewski, A.E., Lehman, A. & Overton, J.McC (2002) Predicting species spatial distributions using presence-only data: a case study of native New Zealand ferns. *Ecological Modelling*, **157**, 261–280.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article:

Appendix S1 Modelled estimates of sampling effort.

Appendix S2 Simulating truth in GIS: Virtual species' distributions.

Appendix S3 Model evaluation: Testing data.

Appendix S4 Supporting references.

Table S1 Model estimates from logistic regression analysis of sampling effort.

Table S2 Simulated regression coefficients used to generate virtual distributions.

Table S3 Model summaries from logistic regression analysis of the distribution of virtual bryophyte species.

Table S4 Model summaries from logistic regression analysis of the distribution of virtual butterfly species.

Table S5 Model evaluation metrics for virtual bryophyte species.

Table S6 Model evaluation metrics for virtual butterfly species.

Figure S1 Choropleth maps of variables used to estimate spatially-biased sampling effort.

Figure S2 Mapped predictions of effort models.

Figure S3 Environmental predictors used to simulate distributions.

Figure S4 Spatial distributions of virtual vascular plant, bryophyte, and butterfly species generated.

Figure S5 Esimates of logistic regression coefficients for virtual vascular plant species.

Figure S6 Estimates of logistic regression coefficients for virtual bryophyte species.

Figure S7 Estimates of logistic regression coefficients for virtual butterfly species.

Figure S8 Spatial distributions of logistic regression estimates for a common, rare, and very rare virtual bryophyte species under random, biased, and pweight-adjusted modelling scenarios.

Figure S9 Spatial distributions of logistic regression estimates for a common, rare, and very rare virtual butterfly species under random, biased, and pweight-adjusted modelling scenarios.

Figure S10 Spatial distributions of Maxent estimates for a common, rare, and very rare virtual bryophyte species under random, biased, bias-file adjusted, and target-group background adjusted modelling scenarios.

Figure S11 Spatial distributions of Maxent estimates for a common, rare, and very rare virtual butterfly species under random, biased, bias-file adjusted, and target-group background adjusted modelling scenarios.

BIOSKETCHES

Jessica Stolar is a PhD student in the Department of Renewable Resources at the University of Alberta. Her research interests are impacts of land use and climate change on Alberta's rare vascular plants and butterflies for conservation planning.

Scott E. Nielsen is an Alberta Biodiversity Conservation Chair and an Associate Professor in the Department of Renewable Resources at the University of Alberta. His laboratory studies the conservation ecology of species to ecosystems with the goal of understanding the processes affecting their distribution, dynamics and interactions.

Author contributions: S.N. and J.S. formulated the ideas; J.S. collected the data; J.S. led the analyses, assisted by S.N.; J.S. led the writing, assisted by S.N.; and all authors carefully revised and corrected the final versions of the manuscript.

Editor: Janet Franklin