Global Ecology and Biogeography, (Global Ecol. Biogeogr.) (2016) 25, 1022–1036

Global Ecology and Biogeography
A Journal of Macroecology

MACROECOLOGICAL METHODS

# Improving niche and range estimates with Maxent and point process models by integrating spatially explicit information

Cory Merow[1]*, Jenica M. Allen[2], Matthew Aiello-Lammens[3,4] and John A. Silander, Jr[5]

[1]*Ecology and Evolutionary Biology, University of Connecticut, Storrs, CT 06269-3043,* [2]*Natural Resources and the Environment, University of New Hampshire, Durham, NH 03824,* [3]*Ecology and Evolutionary Biology, University of Connecticut, Storrs, CT,* [4]*Environmental Studies and Science, Pace University, Pleasantville, NY 10570,* [5]*Ecology and Evolutionary Biology, University of Connecticut, Storrs, CT 06269-3043*

*Correspondence: Cory Merow, Ecology and Evolutionary Biology, University of Connecticut, 75 North Eagleville Road, Storrs, CT 06269-3043.
E-mail: cory.merow@gmail.com

## ABSTRACT

**Aim** Accurate spatial information on species occurrence is essential to address global change. Models for presence-only data are central to predicting species distributions because these represent the only geographical information available for many species. In this paper we introduce extensions to incorporate a variety of types of additional spatially explicit sources of information in Maxent and Poisson point process models. This spatial information comes from the output of other statistical or conceptual models.

**Innovation** Our approach relies on minimizing the relative (or cross) entropy (known as Minxent) between the predicted distribution and a prior distribution. In many scenarios, researchers have some additional information or expectations about the species distribution, such as outputs from previous models. Here, we show how to use this information to improve predictions of both niche models and spatial distributions, depending on what types of spatially explicit prior information is available and how it is incorporated in the model.

**Main conclusions** We illustrate applications of Minxent that include models for sampling bias, explicitly incorporating dispersal/other ecological processes, combining native and invasive range data, incorporating expert maps, and borrowing strength across taxonomic relatives. These applications focus on addressing biological scenarios where range modelling is extremely challenging – non-equilibrium species distributions and rare and narrowly distributed species – due to data limitations. When data are limited, we are typically forced to make informal assumptions or lean on predictions of other models in order to obtain useful predictions; our applications of Minxent provide a formal way of describing these assumptions and connections to other models.

**Keywords**
**Data fusion, dispersal, ecological niche model, expert map, invasion, maximum entropy, sampling bias, species distribution model.**

## INTRODUCTION

The need to map species distributions has never been greater. Conservation planning, biodiversity studies and biogeography rely on accurate spatial information about where species occur. Occurrence models are central to such predictions, because for many species the only geographical information available consists of presence data. Accordingly, the use of presence-only modelling strategies over the last decade has grown rapidly. The most popular tool, Maxent (Phillips *et al.*, 2006), has been cited over 4700 times since 2006 (Google Scholar). However, challenges persist in building reliable models: sampling bias and small sample sizes limit the models one can build, while their correlative nature raises

DOI: 10.1111/geb.12453

questions about the reliability of their predictions. A possible way to address some of these issues is to use mechanistic models (Kearney & Porter, 2009; Buckley *et al.*, 2010), but this is not feasible for most distribution modelling endeavours given the data required. The next best option, for studies that lack critical data on mechanism, is to incorporate all available sources of information, which may include additional phenomenological information or a subset of the most important mechanisms. Existing information – such as occurrence patterns of related species, expert range maps, or simulation models describing ecological processes – may address some of the shortcomings of presence-only data, but is not readily incorporated into existing models. In this paper, we introduce extensions of the Maxent modelling framework (Phillips *et al.*, 2006), and more generally Poisson point process models (Warton & Shepherd, 2010; Chakraborty *et al.*, 2011; Renner & Warton, 2013; Renner *et al.*, 2015), to incorporate additional sources of geographically explicit information in models for species environmental niches and geographical ranges/distributions (which we use interchangeably here).

Our approach to incorporating spatially explicit prior information is motivated by the principle of minimum discrimination information (Kullback, 1959) – one possesses some prior notion which, when confronted with data, is updated to obtain a posterior belief. These beliefs are cast as probability distributions that characterize the precision of the prior understanding. The principle of minimum discrimination information postulates that one should select the posterior distribution which is consistent with the data and diverges as little as possible from the prior distribution (Kullback, 1959; Kesavan & Kapur, 1989; Kapur & Kesavan, 1990). Minimizing the difference between the prior and the posterior is variously known as minimizing the relative entropy, cross entropy, Kullback–Liebler divergence (Kullback, 1959) or information divergence (Caticha, 2008). In species distribution modelling (SDM), when priors are uniform in geographical space, we refer to the model as Maxent. When priors are non-uniform in geographical space we henceforth refer to the models as Minxent (minimum cross (x) entropy; cf. Kapur & Kesavan, 1990), while noting that Minxent is not a new modelling approach but rather a generalization of Maxent (technical details below). Here, we show how the principle of minimum discrimination information leads to a straightforward adaptation to Poisson point process models; namely that the spatial prior corresponds to an offset term commonly used in statistical models (Warton *et al.*, 2014).

A Minxent approach to SDMs is particularly useful if one has additional information about a species' distribution not contained in the presence data. Hence, although we use the general term 'prior' to describe spatially explicit information independent of the presence data, it does not necessarily suggest subjective beliefs, but rather other sources of information. In practice, this information must be cast as a map in geographical space. Minxent can then predict the species distribution that is as similar as possible to this map, but which accounts for any differences that are characterized by the presence data (details in 'The model'). For example, previous models may be available that are driven by other data sources. It is useful to distinguish between priors that characterize a species' niche, which implies interest in the covariates that predict occurrence and transferring the model in time or space, from those that focus on a species' range/distribution, which implies focus on the spatial pattern of the realized distribution. As an example of a niche model, in the section 'Application 3: incorporating native range data to forecast invasions' we develop a prior from a model fitted in a species' native range and projected onto the introduced range. Such a niche model characterizes the potential distribution and is transferrable in space or time (cf. Anderson, 2013), to the extent that the model assumptions are met. When the map characterizes a strictly spatial process, for example a model of dispersal probability in 'Application 2: dispersal information', the prediction can be interpreted as a geographical range or distribution. Such spatial models characterize a realized distribution and are not transferable (cf. Anderson, 2013). Notably, depending upon how they are handled in the model, priors are flexible in their ability to both characterize confounding processes that should be factored out of predictions (e.g. sampling bias) and biological processes that help to describe distributions (e.g. dispersal).

To understand the types of applications where Minxent may be useful it is important to note a key difference between a Minxent prior and the common application of priors on Bayesian model parameters. In Minxent, priors do not directly constrain the values of the parameters (i.e. $\boldsymbol{\lambda}$, defined below), but rather place a constraint on the spatial projection of the model. This difference is valuable because it may be challenging to define priors on parameter values governing a species' niche, for example how occurrence relates to summer precipitation. Rather, a researcher may have a better understanding of the mapped species' distribution, making it more natural to formulate priors in geographical space. Many possible sources of information could constrain the spatial projection. When modelling invasions, where the species is not in equilibrium with the environment, projecting the distribution based on patterns of dispersal that affect the accessible habitat could be a useful prior. Alternatively, an expert map could provide a more reasonable first guess than a uniform distribution when attempting to predict a species' geographical distribution. Minxent allows one to include multiple such sources of spatial information; these applications are illustrated below.

In the remainder of this paper, we explain how to build Minxent models. We discuss the technical and practical aspects of building and interpreting these models and illustrate them with three applications in the main text and three more in the Appendix S6 to address common challenges encountered when building SDMs, with a particular focus on the challenges of predicting distributions of invasive or rare species.

## THE MODEL

### How Minxent works

While the key details of Maxent (Phillips & Dudik, 2008; Elith *et al.*, 2010, 2011; Yackulic *et al.*, 2012; Merow *et al.*, 2013; Merow & Silander, 2014), and Poisson point process models (PPPMs) more generally, are treated at length elsewhere (Warton & Shepherd, 2010; Chakraborty *et al.*, 2011; Fithian & Hastie, 2013; Renner & Warton, 2013; Warton *et al.*, 2014; Renner *et al.*, 2015), we highlight a few points that will help to clarify how they incorporate spatially explicit prior information. In short, Maxent and PPPMs are formally equivalent in the limit of increasing spatial resolution; however, Maxent imposes a specific way of handling a number of modelling decisions, including background selection and weighting, scale dependence based on the gridded landscape and use of various tools common in machine learning (details in Fithian & Hastie, 2013; Renner & Warton, 2013; Renner *et al.*, 2015). Both modelling approaches assume that the presence locations are independent of one another, that the intensity of presence records varies spatially and that the intensity varies log-linearly with environmental predictors (Renner *et al.*, 2015). To estimate this intensity, Maxent/PPPMs contrast the environmental conditions at $m$ presence locations against those at $n$ background locations, irrespective of whether the background locations were sampled. The dependence on background locations necessitates careful consideration of how those data are sampled, a topic addressed elsewhere (VanDerWal *et al.*, 2009; Anderson & Raza, 2010; Anderson, 2013). We denote the predicted intensity across the landscape as relative occurrence rate (ROR; cf. Fithian & Hastie, 2013; Merow *et al.*, 2013). It is helpful to interpret ROR as a multinomial distribution in geographical space, where the probabilities describe which cells are most likely to contain a presence (Merow *et al.*, 2013). ROR sums to unity across all cells in the region used for model fitting (i.e. for a landscape of $N$ cells, and ROR in cell $i$ denoted $P_i$, $\sum_{i=1}^{N} P_i = 1$). Notably, cells with low ROR may still have a high absolute probability of presence, but have lower relative probability than other cells in the region. The ROR is Maxent's so-called *raw output*. We do not consider the logistic transformation of Maxent's output (cf. Phillips & Dudik, 2008) due to the many issues that have been raised with it (Royle *et al.*, 2012; Merow *et al.*, 2013), but rather focus on raw output.

In the SDM context, the principle of maximum entropy asserts that predictions should remain as uniform as possible in geographical space, while also being consistent with constraints imposed by the occurrence data (further details in Merow *et al.*, 2013). Underlying this assertion is a prior distribution; one has no knowledge of the species' environmental preferences, so the species is equally likely to occur anywhere. However, there are situations when this ignorance is unjustified, and it is desirable to include some additional information. Minxent incorporates spatially explicit information in the form of rasters that express the prior probability (ROR) of observing a presence in each grid cell. For a landscape of $N$ cells, the value of the prior distribution in cell $i$, denoted $Q_i$, must obey $\sum_{i=1}^{N} Q_i = 1$ where $0 < Q_i < 1 \; \forall i$. Exactly as with the predicted ROR described above, the $Q_i$ describe the prior relative rate of observing occurrences. In the applications that follow, the $Q_i$ are the predictions from other types of occurrence models, rescaled to sum to unity. For example, if sampling is biased such that some regions or environments are more intensely sampled than others, a larger number of presences observed in a particular environment may be due to increased sampling there, rather than preferred habitat (Phillips, 2008; Phillips *et al.*, 2009). If the species is equally likely to occur anywhere, presences should be observed in proportion to sampling probability. Differences in relative sampling probability can be reflected in a prior distribution, under the null assumption that the likelihood of a presence record in a particular environment is proportional to the sampling probability there.

A prior distribution more generally reflects information about the species distribution that is independent of the presence data and could reflect researchers' assumptions or output from independent models. To fit a model, Minxent minimizes the Kullback–Leibler divergence to produce a prediction that is maximally similar to the prior distribution while obeying constraints imposed by the presence data (Phillips *et al.*, 2006). We minimize, instead of maximize, the Kullback–Liebler divergence because it is defined with the opposite sign to the entropy. This minimization is equivalent to a maximum likelihood estimation of the parameters (Merow *et al.*, 2013). Let $\mathbf{z}(x_i)$ denote the vector of environmental predictors in cell $x_i$. Then, $P^*(\mathbf{z}(x_i))$ is the distribution we aim to predict, describing the ROR in each cell, while $Q(x_i)$ is the prior distribution. The Kullback–Leibler divergence takes the form

$$\sum_{i=1}^{N} P^*(\mathbf{z}(x_i)) \, \log \frac{P^*(\mathbf{z}(x_i))}{Q(x_i)} \tag{1}$$

where $N$ is the total number of cells. Minimizing this divergence results in the prediction $P^*(\mathbf{z}(x_i))$:

$$P^*(\mathbf{z}(x_i)) = Q(x_i) e^{\mathbf{z}(x_i)\boldsymbol{\lambda}} / C \tag{2}$$

where $\boldsymbol{\lambda}$ is a vector of fitted coefficients and $C$ is a constant that ensures normalization. Note that the Maxent software package returns $P^*(\mathbf{z}(x_i))/Q(x_i)$ (see Appendix S2 in Supporting Information for a workaround). Hence the primary difference between Minxent and Maxent is the inclusion of $Q(x_i)$ to describe spatial pattern. Consequently, Minxent should generally be used with the same precautions as Maxent (Elith *et al.*, 2011; Yackulic *et al.*, 2012; Halvorsen, 2013; Merow *et al.*, 2013; Merow & Silander, 2014).

The connection between Maxent and PPPMs simplifies the interpretation of the prior. The prior distribution of Maxent can be interpreted as an offset term in a log-linear model for ROR, as in Poisson regression (Warton *et al.*, 2014). This is apparent by taking the natural logarithm of both sides of equation (2):

$$\ln[P^*(\mathbf{z}(x_i))] = \ln(Q(x_i)) + \mathbf{z}(x_i)\boldsymbol{\lambda} - \ln(C). \qquad (3)$$

The value of connecting the prior and offset concepts is apparent below in 'Application 1: sampling bias', based on the common interpretation of using the offset to represent exposure. This connection also has practical implications; Minxent models can be fitted with standard generalized linear modelling software (see appendix of Renner *et al.*, 2015, for multiple options). Because the term 'prior' already has a well-established meaning in Bayesian statistics, which is rather different from that intended in entropy calculations, we henceforth refer to prior sources of information as offsets, following other applications of this modelling framework (Warton *et al.*, 2014).

It is useful to distinguish between what we term *nuisance offsets* and *informative offsets*. Nuisance offsets are those that should be factored out of the prediction because they contain information confounding predictions (e.g. 'Application 1: sampling bias'). Informative offsets contain additional biological information on the true distribution (e.g. 'Application 2: dispersal Information') or niche (e.g. 'Application 3: incorporating native range data to forecast invasions') and must be included in the prediction, as in equation (2). These differences affect whether model predictions can be interpreted as transferable niche models or as non-transferable geographical distributions (cf. Anderson, 2013). Models where the nuisance offset has been factored out can be interpreted in terms of an environmental niche ('Application 1: sampling bias'), while those that include an informative offset can predict geographical distributions (realized distribution in 'Application 2: dispersal information') or niches ('Application 3: incorporating native range data to forecast invasions') depending on whether the offset describes a spatial pattern or a niche, respectively. Maxent assumes the offset is a nuisance offset and factors it out, while PPPMs assume that the offset is an informative offset and do not factor it out. Hence if either fitting tool is used under the opposite assumption one has to factor the offset in/out accordingly.

The coefficients ($\boldsymbol{\lambda}$) from Minxent models should be interpreted differently from Maxent coefficients. In Maxent models, without offsets, the coefficients and associated response curves describe which predictors are most important in shaping a species' distribution. However, when informative offsets are used, some amount of that information is already contained in the offset. Minxent coefficients therefore describe how the prediction $P^*(\mathbf{z}(x_i))$ differs from the offset $Q(x_i)$. For example, a positive response to precipitation implies that a species' distribution is more sensitive to precipitation than assumed by the offset (Fig. 4).

## APPLICATIONS

A variety of applications become possible once it is recognized that one can specify relevant assumptions or additional information via spatially explicit offsets. Broadly, in the following examples, this spatial information is the output from another model, projected across the landscape where a prediction of a species' potential or realized distribution is desired. Discussion focuses on how to build a model for the offset, incorporate it into a Maxent/PPPM and interpret results in a variety of applications.

### Data

We illustrate applications using four data sets on invasive species in the USA. We chose this system because it represents species whose distributions are not at equilibrium, the commonness of sampling bias in invasive species databases and, the availability of different data sources. The largest dataset focuses on the north-eastern USA and was collected for the Invasive Plant Atlas of New England (IPANE; Bois *et al.*, 2011). Survey effort in the IPANE dataset is highly uneven over New England; however, IPANE represents an enormously valuable effort to inventory a large taxonomic group over a large spatial extent with attention to sampling intensity and spatial accuracy. A time series of occurrence records over the last century has also been compiled in the IPANE database from herbarium records.

To better characterize the environmental preferences of these species we built models that combine IPANE data with two data sets for the same species spanning the USA [EDDMapS (http://www.eddmaps.org/) and GBIF (http://www.gbif.org/)] and another data set from a large part of the species' native range in Japan (Phytosociological Releve Database, or PRDB; http://ss.ffpri.affrc.go.jp/labs/prdb/index-e.html). We use *Celastrus orbiculatus* (oriental bittersweet) and *Berberis thunbergii* (Japanese barberry) to illustrate our analyses. Both are invasive woody species native to east Asia that grow in forest understories and edges. We use climate data from the WorldClim database (Hijmans *et al.*, 2005) at the 5′ scale, which included the maximum temperature of the warmest month, temperature seasonality, isothermality, and mean annual precipitation. Additionally, we used human population and road density [US Census TIGER/line files (https://www.census.gov/geo/maps-data/data/tiger-line.html) and ISCGM Global Map v.2 (http://www.iscgm.org/)] to account for anthropogenic correlates of sampling bias; we expected places with greater accessibility to be biased toward greater sampling.

In addition to these applications, we present three additional applications in the Appendix S6, focusing on rare species in the Cape Floristic Region of South Africa.

### Application 1: sampling bias

#### *Background*

We begin by using Minxent to account for sampling bias, as this is perhaps the most critical improvement needed in many presence-only models (Phillips *et al.*, 2009; Chakraborty *et al.*, 2011; Yackulic *et al.*, 2012; Merow *et al.*, 2013). Occurrence data sets often contain some sampling bias, wherein some environmental conditions (near towns, roads, etc.) are more likely to have been sampled than others (Reddy & Dávalos, 2003; Graham *et al.*, 2004; Phillips *et al.*, 2009). When sampling is biased, one cannot differentiate whether species are observed in particular environments because those locations are preferable to the species or to the biologist (Phillips *et al.*, 2009; Sastre & Lobo, 2009; Wisz & Guisan, 2009). The probability of recording an individual in a cell can be decomposed into the product of the probability of sampling the cell and the ROR there (Phillips *et al.*, 2009; Yackulic *et al.*, 2012). This decomposition is readily understood in terms of a regression offset term in PPPM models (equation 3). Offsets are typically used in Poisson regression to account for exposure with count data; for example, twice as many counts are likely to be observed if they have had twice as many opportunities to occur (conditional on the covariates). In the context of sampling, we expect to observe twice as many presences in environments that have received twice as much sampling effort. Hence one can include sampling bias in Minxent by converting the presence counts to a *rate* of presence counts per unit sampling effort.

Modelling sampling bias typically employs target group sampling (Ponder *et al.*, 2001; Phillips *et al.*, 2009), in which the presence locations of species recorded using the same sampling protocol as the focal species (e.g. from the same database) are used to estimate sampling probability, under the assumption that those surveys would have recorded the focal species had it occurred there (Phillips *et al.*, 2009). Sampling intensity models can be built using Maxent/PPPM software by supplying the target group presences along with covariates describing sampling probability (e.g. road density). Modelling sampling in this way allows one to infer sampling probability at locations where target group records do not exist. This contrasts with the biased background approach typically used with Maxent (cf. Phillips *et al.* 2009), in which only locations with target group samples are included in the background. Inferring sampling can be valuable when recorders have probably sampled many locations without recording a species (e.g. for targeted sampling used with cryptic species) but can be misleading when the true sampling effort is well characterized by the target group (e.g. for surveys with standardized sampling routes). Importantly, the covariates that describe the sampling process should typically not be used to describe the ROR of the focal species because separate coefficients for the effect of a shared covariate (e.g. temperature) for both sampling and occurrence are not identifiable (details in Appendix S3). The resulting raw output (normalized) of the sampling model can be used as an offset in a subsequent Minxent model using the environmental covariates needed to describe occurrence of the focal species.

A similar approach can also be used to factor out detection probability. Detection probability, i.e. the probability that the species would have been recorded, conditional on having sampled the location, can be differentiated from sampling bias as $P(\text{presence record}) = P(\text{sampling}) * P(\text{detection}|\text{sampled}) * P(\text{occurrence})$. While it would be challenging to differentiate between $P(\text{sampling})$ and $P(\text{detection}|\text{sampled})$ with presence-only data, we can readily estimate their product simply by choosing not to remove duplicated presences in cells. In this case, replicated presences in a single cell are interpreted as higher sampling effort there, and hence a higher probability of detection (though this may be a better assumption for rare species of interest than for common species; Anderson, 2003). In the example below, we retained all target group observation locations for the bias models in order to model the product of sampling and detection probabilities. In 'Application 3: Incorporating native range data to forecast invasions', we show how to include a sampling offset and a informative offset in the same model.

### *Approach*

We predicted the potential distribution of two invasive plants, *C. orbiculatus* and *B. thunbergii* across New England (Figs 1 & S2). The density of occurrence points and sampling are both highest in southern New England (Fig. 1a), meaning that is it critical to account for sampling bias to infer whether a low density of occurrence points in the north is due to unsuitable habitat or a lack of sampling there. The target group sample included unique sampling locations of all 111 invasive species in the IPANE database (5490 locations). Presence data included 355 locations for *C. orbiculatus* and 373 locations for *B. thunbergii* after removal of duplicate observations within each 5′ grid cell (note that duplicates *are* removed for the occurrence model but not for the sampling model described above because duplicates were interpreted to indicate greater sampling effort). Anthropogenic factors in the bias models included human population and road density (Fig. 1c). To predict ROR, we used maximum temperature of the warmest month, temperature seasonality, isothermality and mean annual precipitation. We used linear, quadratic and product features for all models to create relatively simple response curves (cf. Merow *et al.*, 2014), with other Maxent software settings left at default values.

### *Results*

A typical Maxent model that ignores sampling bias (Fig. 1b) predicts a much smaller potential distribution than a model that accounts for sampling bias for *C. orbiculatus* (Fig. 1d) and *B. thunbergii* (Fig. S2). There are two potential sources of bias that inhibit our ability to predict the potential (equilibrium) distribution of these invasive species: (1) The citizen-science database that we used is very likely to be biased towards easily accessible locations, and (2) the species are probably still spreading northward, leading to sampling that is biased toward warmer, southern climates. Predictions that incorporate bias with only anthropogenic factors (Fig. 1d) are therefore best suited for predicting the current distribution of the species, while accounting for patterns of bias among volunteers. There remains the possibility that these species have already reached their equilibrium distribution; however, one cannot determine this from presence data alone (see one solution in 'Application 2: Dispersal Information'). With only presence data, the best we can do is describe the predictions conditional on different sets of assumptions about sources of bias (e.g. no bias versus bias near human

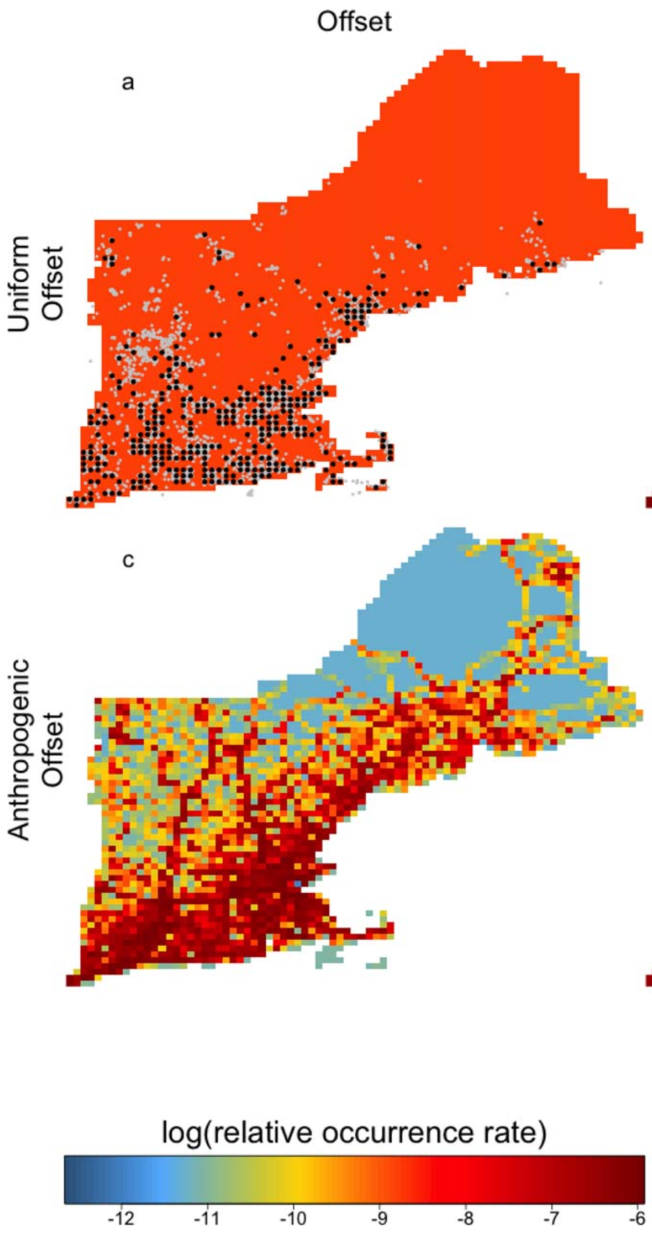

Figure 1 Differences in predictions for *Celastrus orbiculatus* when accounting for sampling bias. The left column indicates the offset used (describing independent spatially explicit information) in conjunction with occurrence data, shown as black dots in (a), to make the predictions of relative occurrence rate in the right column. Panel (a) shows the commonly used uniform offset while panel (c) shows an offset based on sampling effort, as predicted by human population and road density. The target group data are shown as grey dots in (a). Importantly, accounting for sampling bias leads to a larger predicted potential distribution (d) for this invasive species than the model that ignores sampling bias (b).

settlement versus. bias due to dispersal limitation in northern New England), in the hope of refining hypotheses to drive data collection that can differentiate among these assumptions. Although considering different possible assumptions complicates interpretation, it is critical that researchers understand and report how their predictions differ with assumptions.

## Application 2: dispersal information

### *Background*

Forecasting the distributions of species with shifting ranges is relevant to pressing issues, including invasions and responses to disturbance, land-use change or climate change (Václavík & Meentemeyer, 2009; Elith *et al.*, 2010). Invasions represent some of the most extreme range shifts because occurrence patterns may contain a strong historical signal (e.g. introduction points), which can lead to an underestimation of the range of suitable environmental conditions (Franklin, 2010; Gallien *et al.*, 2010; Merow *et al.*, 2011). Modelling range-shifting species is similar to the problem of having a biased sample; presences have not been observed in suitable habitat. To account for range dynamics, one could use dispersal models (e.g. Engler & Guisan, 2009; Merow *et al.*, 2011; Bocedi *et al.*, 2014) to obtain information on the probability of the species having reached particular locations to link realized (actual occurrence locations) and potential distributions (suitable, but not necessarily accessible). Dispersal predictions might result from models ranging from simple diffusion models that describe spread across a homogeneous landscape to simulation models with a complex system of rule-based movements (Holmes, 1993; Higgins & Richardson, 1996; Okubo & Levin, 2001; With, 2002; Hastings *et al.*, 2005). Such spread models, though imperfect, may be more useful for exploring dynamic patterns than the typical alternatives of assuming no dispersal limitation or unlimited dispersal.

### *Approach*

SDMs typically assume that the distribution being studied is at equilibrium; hence it is important to incorporate an understanding of dispersal to reflect the non-equilibrium distribution from which we have samples. A dispersal model can help to describe either the realized or the potential distribution, depending upon how the offset is handled. We focus on an example from Merow *et al.* (2011), who described the spread of *C. orbiculatus* across the New England landscape over the last century using a mechanistic model of dispersal by birds, bird and plant habitat preferences, and a simple density-dependent population growth model (Fig. 2a–d). The black dots in Fig. 2(a)–(d) (and m–p) show the data describing the temporal pattern of spread; this time series allowed us to fit models at each time period. Data collected as of 2009 are shown as black dots in Fig. 2(e)–(l), and provide our best estimate of the potential distribution. The dispersal model was developed independent of the presence data and climatic data supplied to Minxent. Consequently, the dispersal model over-/underpredicts spread in some locations and the objective here is to improve the predictions of temporal dynamics using the observed spatial occurrence pattern.

By running multiple simulations of the (stochastic) dispersal model, Merow *et al.* (2011) predicted the probability that the species occurs in a cell on the basis of the fraction of simulations with presence. The offset supplied to Minxent was a normalized version of this prediction across the landscape at a snapshot in time. The offset can be used in two ways. If the offset is factored out (as with sampling bias above), the model predicts the potential distribution (effectively accounting for the bias with which the species has sampled the landscape, so to speak; Fig. 2i–l). If instead, the offset is not factored out of the prediction (as a informative offset, similar to that constructed in 'Using expert range maps' and 'Higher-order taxonomic in the Supporting Information'), the model predicts the realized distribution at the instant in time to which the dispersal model corresponds (e.g. 1960 in Fig. 2n). Different snapshots from the dispersal model can be used to describe how the realized distribution changes over time. Hence, the dispersal model downweights locations with environmentally suitable conditions but which are inaccessible. For model building, Maxent software settings were left at default values, except that only linear and quadratic features were used to accommodate the smaller sample sizes.

### *Results*

Two improvements in predictions are noticeable in the application of Minxent: one in the time series of the realized distribution and one in the potential distribution. We can compare the predictions of the realized distribution predicted by the dispersal model from Merow *et al.* (2011; Fig. 2a–d) with those predicted by Minxent using the dispersal model as a informative offset (Fig. 2m–p). The models generally agree with each other and the occurrence data in southern New England; however, Minxent avoids overprediction in northern New England. Next, we compared the potential distributions predicted by Maxent (with a uniform offset; Fig. 2e–h) and Minxent (Fig. 2i–l; i.e. by treating the dispersal information as a nuisance offset at each time step.) Our best knowledge of the potential distribution of *C. orbiculatus* consists of an independent data set from the IPANE database collected from 2000–09 (black dots in Fig. 2e–l). Ideal predictions would be stable across time steps and closely match the distribution of IPANE data points in 2009, or potentially predict a larger range, since the species may still be spreading. Our results generally show an expanding distribution with increasing sample size (over time), with some variability in Minxent predictions across time steps (Fig. 2i–l), as we would expect based on the small sample sizes available in 1940 and 1960 (21 and 37 points, respectively). Minxent predictions better (third row, Fig. 3) match the validation data at any given time step, compared with the corresponding Maxent prediction at the same time step (second row, Fig. 3). Hence, it is apparent that by incorporating dispersal to describe bias in the accessibility of habitat, we are better able to describe locations at risk of invasion.

## Application 3: using native range data to forecast invasions

### *Background*

The distribution of invasive species in their introduced range may reflect only a portion of the species' environmental tolerance due to historical effects and limited dispersal. Hence, native range data may provide a better estimate of the species' true tolerance in the introduced region (e.g. Guisan & Thuiller, 2005). However, biotic interactions and dispersal limitation in the native range may also limit the realized climatic niche of a species, resulting in an apparent expansion of the species' climatic tolerances during invasion

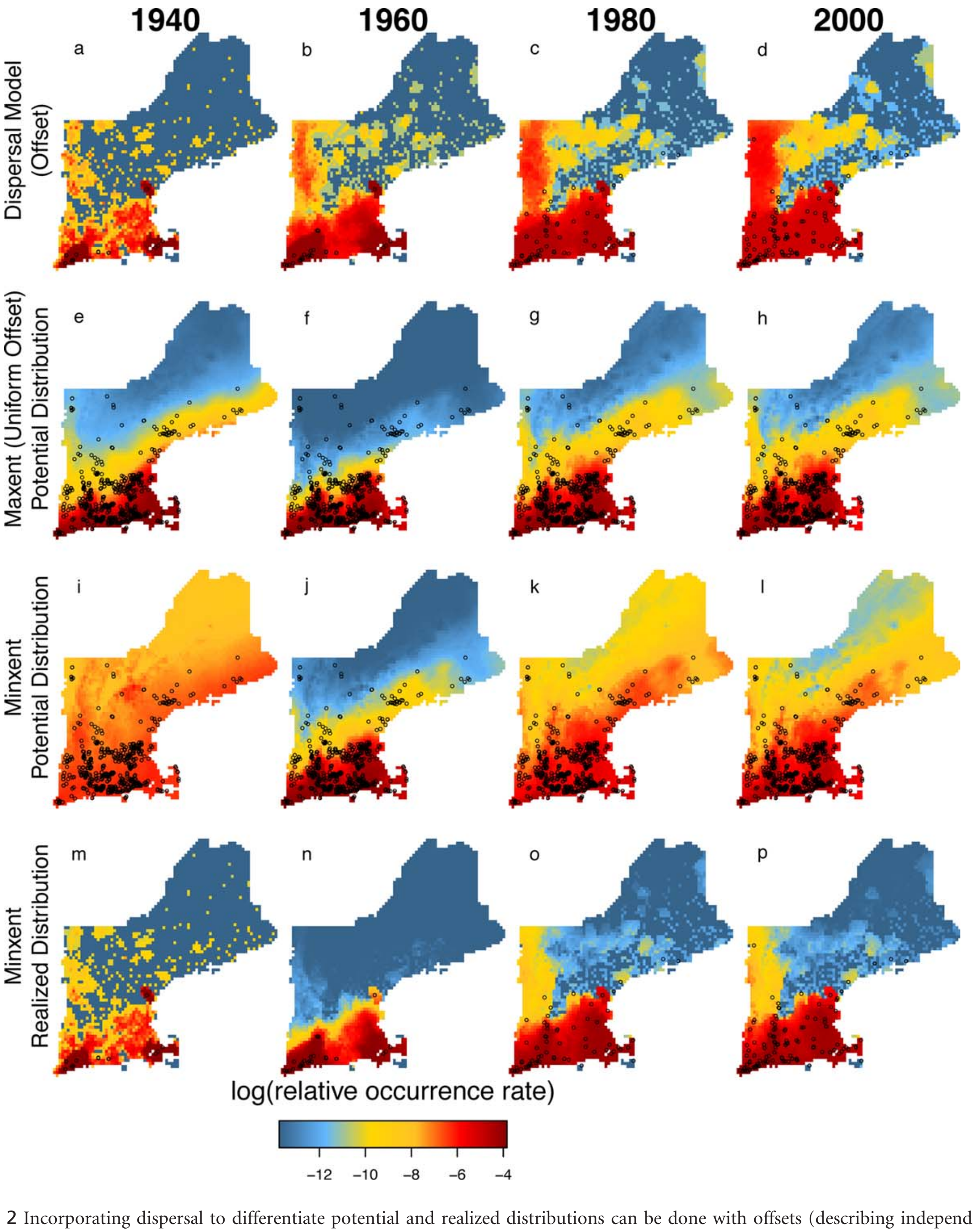
1940
1960
1980
2000
Dispersal Model (Offset)
Maxent (Uniform Offset) Potential Distribution
Minxent Potential Distribution
Minxent Realized Distribution
a
b
c
d
e
f
g
h
i
j
k
l
m
n
o
p
log(relative occurrence rate)
−12
−10
−8
−6
−4

Figure 2 Incorporating dispersal to differentiate potential and realized distributions can be done with offsets (describing independent spatially explicit information). (a)–(d) A dispersal model that estimates the relative probability that a species has reached each point on the landscape. (e)–(h) Predictions from Maxent with a uniform offset. (i)–(l) If the dispersal information is treated as nuisance offset (factored out of the prediction), model predictions represent the species' potential distribution. (m)–(p) If, instead, the offset is treated as biological (not factored out of the prediction), model predictions represent the species' realized distribution. In (a)–(d) and (m)--(p) black dots represent the known occurrence points at the given point in time, and were used to fit the Maxent/Minxent models at each time step. In (e)–(l), the black dots represent independent occurrence data which represent our best knowledge of the potential distribution as of 2009. Note that Minxent's predictions of potential distribution (i–l) match the independent occurrence data better than the corresponding Maxent predictions (e)–(h) at each time step. Similarly, Minxent's predictions of realized distribution (m–p) better match the time series of spread compared with the dispersal model alone (a–d).

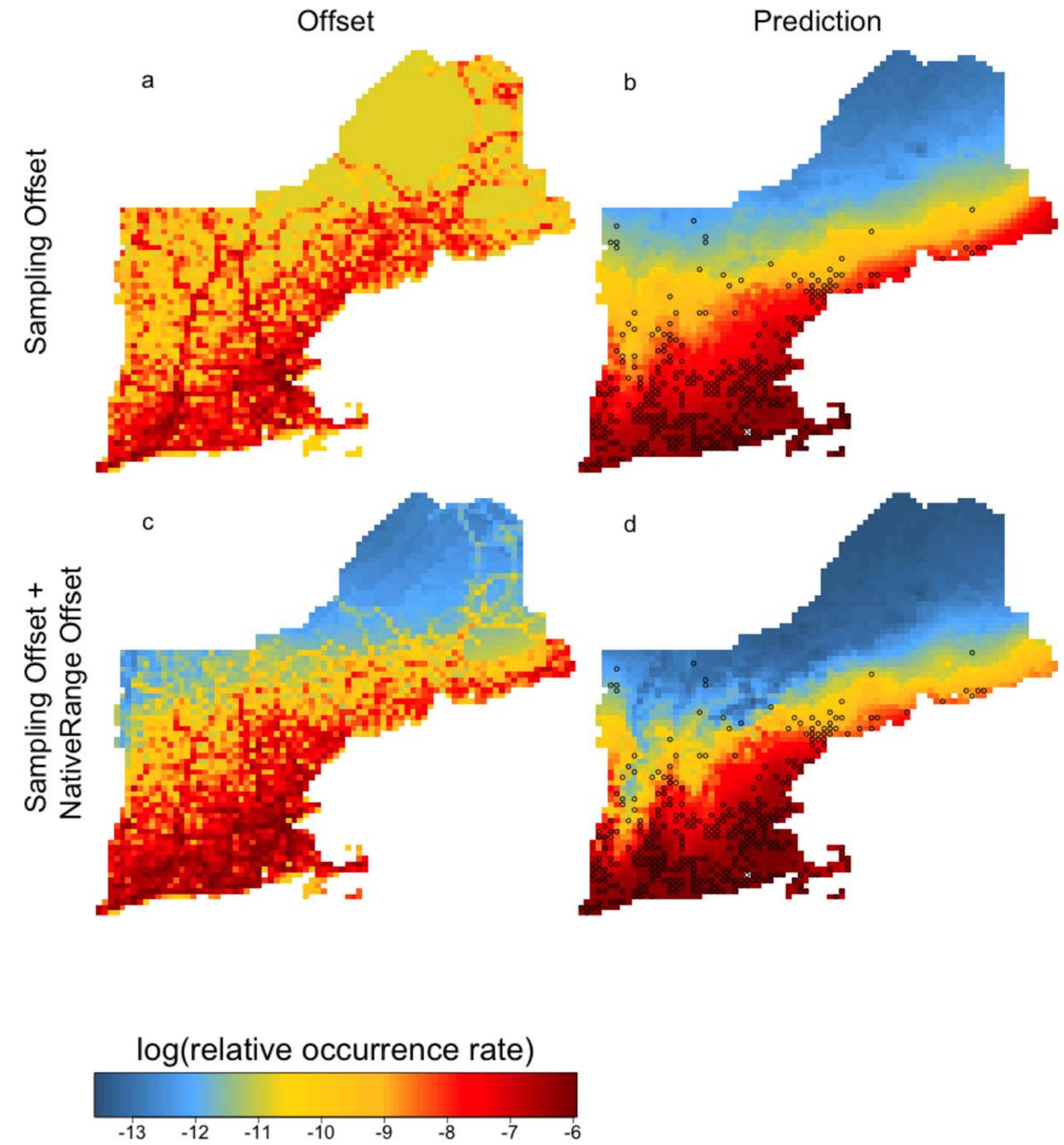

Figure 3 (a) A null model for the distribution of *Celastrus orbiculatus*, based on a target group sample as described in 'Application 1: sampling bias'. (b) The Minxent prediction for a model including sampling bias, but not native range information. (c) An offset that includes both sampling bias and native range information. Native range information derives from a model built in Japan, projected onto New England. (d) The Minxent prediction for a model including both sampling bias and native range data.

(Broennimann *et al.*, 2007; Early & Sax, 2014). Therefore, both the native and invasive range may miss portions of species climatic tolerances and it is desirable to use both invasive and native range data together. Simply combining these data sets in a single Maxent model is inadvisable due to differences in sample size, potential bias (cf. Appendix S6a) and uncertainty in how best to select background points drawn from two disjunct regions (e.g. how many from each). Instead, following the protocol in Appendix S6(a), one can fit a model in the native range, project it onto the introduced region (Fig. 3b) and use this as an offset. This approach can be particularly useful when few observations exist in the introduced region and more data-hungry approaches are not suitable (cf. Ibáñez *et al.*, 2009).

Given the flexibility of Minxent to capture a variety of information sources, it is natural to extend these models to include multiple sources. A key attribute of the Minxent predictions that makes them easy to work with is their multiplicative nature. Including multiple sources of information involves multiplying (or adding, if included in the linear predictor of the PPPM) all the relevant offsets together and normalizing the result. For example, in the following

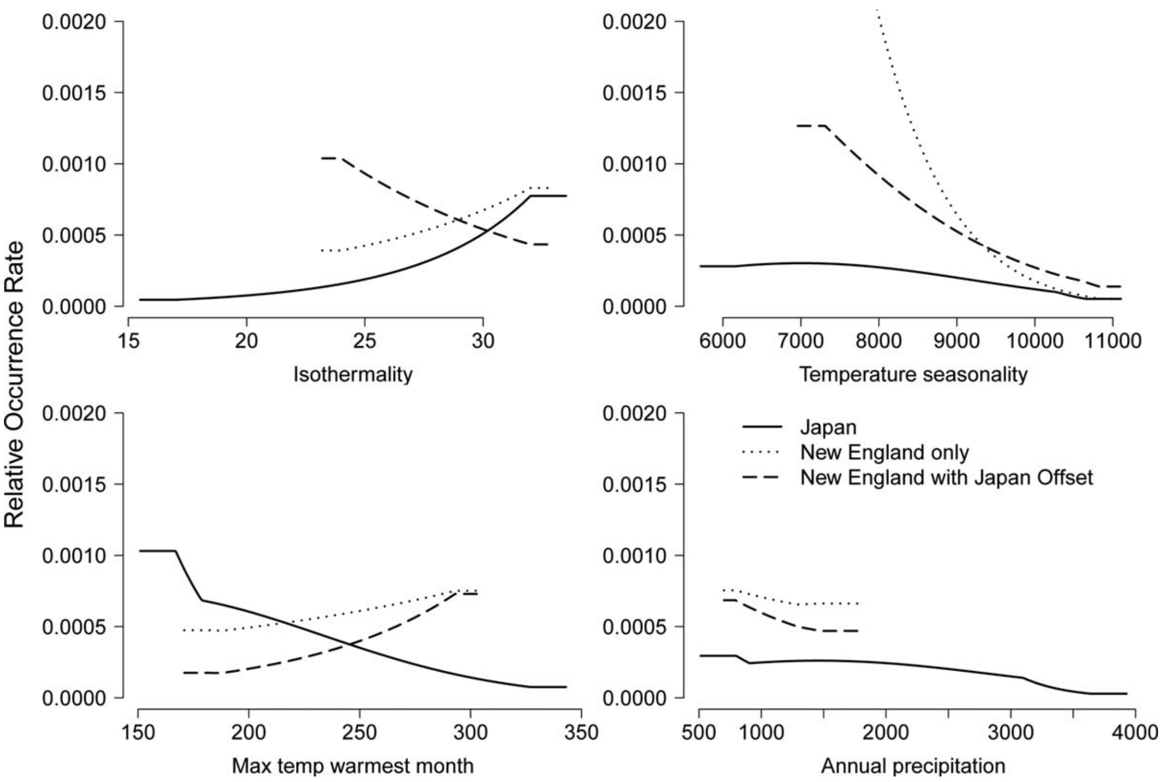

Figure 4 Modelled response curves for each climate predictor in each range for *Celastrus orbiculatus.* The range of climatic conditions in New England is contained within the spectrum of environments in the native range (Japan). The response to temperature seasonality and annual rainfall are similar in both the native and invaded ranges, but there appears to be a shift toward lower isothermality and warmer temperatures in the invaded range.

application, we include both sampling bias (‘Application 1: sampling bias’) and a native range offset (methodological details in Appendix S5).

*Approach*

We predicted the potential distributions of *C. orbiculatus* and *B. thunbergii* in New England, building on Application 1, but adding data from the native range. We accounted for sampling bias in the native range (Japan) by fitting sampling models based on human population and road density as recommended in Application 1 using a target group consisting of samples of *B. thunbergii*, *C. orbiculatus*, *Elaeagnus umbellata* and *Rosa multiflora*. We projected those models to New England and combined them with a sampling bias model based on a target group using all available occurrence records from IPANE (Fig. 3c). This offset was used in conjunction with occurrence data from IPANE, with duplicates removed at the 5′ scale. We used linear, quadratic and product features to create relatively simple response curves (cf. Merow *et al.*, 2014) from maximum temperature of the warmest month, temperature seasonality, isothermality and annual precipitation with other Maxent software settings left at default values.

*Results*

In a model without native range information, *C. orbiculatus* was most likely to occur in southern New England (Fig. 3b) despite sampling bias there (Fig. 3a). Occurrences in the native range suggest that *C. orbiculatus* is most likely to occur in southern New England, with low probability in northern regions (compare Fig. 3a,c). A comparison of models that ignore, versus include, the native range data (Fig. 3b versus 3d) shows that the native range offset serves to predict a more restricted range for *C. orbiculatus* (see similar results for *B. thunbergii* in Appendix S5). Due to the high sampling bias the model in Fig. 3(b) predicts that habitat is suitable in central New England, where there are few samples (samples shown as white circles in Fig. 3b,d). By including the native range data, it becomes apparent that central New England represents environments of low suitability. Note that the dispersal model in Application 2 predicts a similar pattern of low suitability in central New England (Fig. 2l) without accounting for native range information, but much higher suitability in coastal Maine. This coastal region is thus the highest priority for further data collection.

Model coefficients (summarized with response curves in Fig. 4) could be interpreted to suggest a niche shift; they describe the variables that differentiate between the native range offset and the occurrence points in the introduced region. The responses to maximum temperature and isothermality have opposite directions in the model that includes the native range data compared with the model that excludes those data (Fig. 4a,c). This *suggests* a niche shift toward lower isothermality and higher temperature in the introduced range, compared with the native range.

The connection between Maxent and PPPMs suggests another way to approach this example. If one were to fit Bayesian PPPMs, the posterior distributions of coefficients obtained in the native range could be directly used as Bayesian priors on the same coefficients in the introduced range. It is unnecessary to project the native range model into geographical space as an intermediate step. Of course, incorporating Bayesian priors on coefficients is not possible within the Maxent software package, so our projection approach is necessary if working with Maxent. We also note that it is useful to project a native range model into geographical space if it uses a different response variable, making it impossible to compare the coefficients between the native and invaded ranges. For example, higher-quality data might be available in the native range where one might fit a presence–absence model. This model can be projected on to the introduced range and normalized to obtain an ROR to use as the offset. This is discussed further in Appendix S5.

## OTHER POSSIBLE APPLICATIONS

Extending Maxent/PPPMs to incorporate spatial information naturally leads to a wide range of possible applications. In this section we briefly highlight a few possible avenues, while noting that detailed examples are included in the Supporting Information.

Rare and narrowly distributed species pose a significant modelling challenge because they are often associated with small samples of presence observations. Different types of spatial information may help to increase precision or reduce bias. For example, it may be valuable to borrow strength across phylogenetic relatives (assuming niche conservatism, cf. Wiens *et al.*, 2010) or members of a functional group to reflect our best guess about a species' climatic niche. This approach may be particularly valuable for biodiversity studies, where the challenge of obtaining large sample sizes for many species is most apparent. To implement this approach, one can build a model for the offset from occurrences of similar species. Notably, such a model allows the use of different suites of covariates to describe occurrence of the clade/functional group, which may help to account for differences between the factors shaping the niche or distribution of the focal species compared with its clade/functional group. Such differences are expected based on niche differentiation or the differential influence of spatial processes. A worked example is provided in Appendix S6(d).

Experts have estimated ranges for many species on the basis of their extensive field experience (e.g. Little, 1971). Typically, the maps produced have relatively coarse resolution, skim over local habitat variation and are most informative about where a species is unlikely to be found (Jetz *et al.*, 2012). As such they provide a valuable complement to presence-only data. Because expert maps represent a spatial model (as opposed to a niche model), they are helpful for predicting the realized distribution. Expert maps could be valuable in a number of scenarios.

**1.** Expert maps can improve predictions of ROR when occurrence data are sparse or biased (the latter may be unknown).

**2.** If the expert map is potentially biased, the occurrence data can be used to 'update' the expert map or identify the environmental conditions where the expert map is lacking to guide further data collection. When expert maps comprise large polygons that are meant to broadly identify presences (rather than absences inside the polygons), combining them with occurrence data can help to determine the 'holes' in the distribution.

**3.** Expert maps can constrain predictions when factors other than those included in the covariates shape the distribution (e.g. biotic interactions or dispersal limitation).

A worked example is provided in Appendix S6(c).

Minxent models can generally be used to combine data sets (Application 3 was one specific example). Many different data sources may be available for some species, each coming with different caveats. For example, a small-scale survey may contain unbiased presence/absence data and these absence data should not be discarded in order to use a presence-only model (Guillera-Arroita *et al.*, 2014). A regional survey may contain unbiased presence data, but these data may be sparse if the species is cryptic or has recently invaded a region. Biased opportunistic data may be available at continental scales, such as that available from large online databases or citizen science initiatives. Different types of sampling bias may exist in each of these sources. Simply aggregating these data sources while ignoring their differences both omits valuable information (absences) and introduces new sources of bias. That is, bias may be absent from high-resolution, small-scale studies, but including these data in regional studies implies that some areas have been sampled more intensely than others. To incorporate different data sources with Minxent, one can sequentially update models with each data type. A worked example is provided in Appendix S6(a).

## DISCUSSION

We have developed a novel extension of Maxent and PPPMs for species distribution modelling that takes advantage of researchers' prior knowledge of species' distributions, hence avoiding the typical assertion of prior ignorance. To achieve this, we have described how to construct spatially explicit models that represent existing knowledge of species' ranges to improve predictions for a variety of applications. These applications were focused on addressing biological scenarios where range modelling is extremely challenging – non-equilibrium species distributions and rare and narrowly distributed species – due to data limitations. When data are limited, we are typically forced to make informal assumptions or lean on predictions of other models in order to obtain useful predictions. Our applications of Minxent provide a formal way of describing these assumptions and connections to other models via spatial offsets.

The connection between Maxent and PPPMs allows some additional flexibility for model checking, model selection and a number of modelling extensions (Fithian & Hastie, 2013; Renner & Warton, 2013; Fithian *et al*., 2014; Warton *et al*., 2014). Notably, Maxent models do not explicitly include any sampling uncertainty – uncertainty estimates are usually obtained with some form of bootstrapping. In contrast, PPPMs are readily formulated in a Bayesian framework (Chakraborty *et al*., 2011), allowing uncertainty to be incorporated through all stages of model construction. Importantly, this can allow one to distinguish between Bayesian prior distributions on coefficients and offsets that are based on purely spatial constraints. The models we present here split analysis into two parts: first we build a model for the offset and then we build a model for ROR. This omits any uncertainty associated with the offset. By working with PPPMs, one can incorporate this uncertainty using a hierarchical model (Warton *et al*., 2014). Despite the advantages of PPPMs for capturing uncertainty, machine learning approaches are also valuable for exploring data sets with a flexibility that would be prohibitive in a parametric model. For a thorough discussion of the pros and cons of parametric and nonparametric SDMs see Merow *et al.* (2014).

The applications of Minxent discussed here provide efficient alternatives to a number of more complex modelling paradigms. Conceptually, Minxent models address similar problems to hierarchical Bayesian models (Latimer *et al*., 2006; Ibáñez *et al*., 2009; Chakraborty *et al*., 2010), aiming to borrow strength from different data sources. For example, a more formal Bayesian approach to combining native and invasive range data that better characterizes uncertainty has been developed (Ibáñez *et al*., 2009), but relatively more data will be necessary to obtain convergence. Dynamic spatio-temporal models that combine occurrence and dispersal data (cf. Hooten & Wikle, 2007; Pagel & Schurr, 2012) that jointly estimate all model components and provide an alternative to the dispersal models discussed here are also available. Models with latent states are useful for combining different data types while explicitly representing the sampling process (e.g. presence-only and presence–absence data; Fithian *et al*., 2014). Hierarchical models to borrow strength across species can simultaneously account for sampling bias and unequal sample sizes across species (Mcinerny & Purves, 2011; Dorazio, 2014; Fithian *et al*., 2014). All these Bayesian approaches provide valuable model-based estimates of uncertainty. However, Minxent models are much less data hungry, less computationally intensive and require less technical expertise, and are therefore efficient for building a large number of models to study many species or model scenarios.

Some potential caveats when using Minxent are worth noting. Foremost, predictions are only improved to the extent that the offset provides accurate information on niches or distribution. A poor dispersal model or expert map will critically bias predictions. Incorporating sampling bias is perhaps the most crucial way to improve presence-only models; however, it can be difficult to reasonably account for sampling bias when locations receive low sampling effort. Appendix S2 highlights an example where low sampling effort can introduce extreme bias in range predictions (models cannot determine much about a range when potentially suitable habitat has a disproportionately lower sampling effort). It is also worth considering whether splitting analysis into two steps – one to build the offset and one to build the niche/range model – is necessary. Such a split can enable the identification of parameters when in fact fitting both models jointly (occupancy models; cf. Royle & Dorazio, 2008) would highlight problems with identifiability. For example, this is why sampling bias models cannot contain covariates that are correlated with those used to describe the niche/distribution (intuitively, the model cannot disentangle whether a higher presence density derives from better habitat or greater sampling effort when the same covariate describes both).

Minxent provides a useful way to explore the implications of different prior assumptions on predictions, with the same precautions as Maxent (Elith *et al*., 2011; Yackulic *et al*., 2012; Halvorsen, 2013; Merow *et al*., 2013; Merow & Silander, 2014). One can only stand to gain from exploring alternative assumptions in the face of sparse data, or by formalizing the role of those assumptions in models. For example, one might categorically reject any models that predict distributions that are far too large or small compared with researchers' expectations. But the expectations that drive these decisions may often be informally based on the types of information that we propose to incorporate via spatial offsets: the expectation might derive from expert experience with the species or its relatives, results of a previously published study or knowledge of the species' traits (informing dispersal or performance in a native range). Minxent provides a framework to include these types of expectations in predictions and helps to understand the spectrum of possibilities consistent with different assumptions. Insights may not always result from finding a single best prediction, but rather finding categorical differences among predictions deriving from different viable assumptions (and how they interact with the data) and generating hypotheses about these differences to guide further study.

## ACKNOWLEDGEMENTS

C.M. acknowledges funding from NSF grant 1046328 and NSF grant 1137366. J.M.A., M.A.-L. and J.A.S. acknowledge funding from NSF grant 1046328 and USDA-NRI (grant 2008-35615-19014). J.M.A. also acknowledges funding from NSF grant EAPSI-0812970. We thank Noboyuki Tanaka for access to the PRDB data and Tony Rebelo for access to the Protea Atlas data. We thank Robert P. Anderson and two anonymous referees for helpful comments that improved the manuscript.

## REFERENCES

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

Anderson, R.P. (2013) A framework for using niche models to estimate impacts of climate change on species distributions. *Annals of the New York Academy of Sciences*, **1297**, 8–28.

Anderson, R. & Raza, A. (2010) The effect of the extent of the study region on GIS models of species geographic distributions and estimates of niche evolution: preliminary tests with montane rodents (genus *Nephelomys*) in Venezuela. *Journal of Biogeography*, **37**, 1378–1393.

Bocedi, G., Palmer, S.C.F., Pe'er, G., Heikkinen, R.K., Matsinos, Y.G., Watts, K. & Travis, J. (2014) RangeShifter: a platform for modelling spatial eco-evolutionary dynamics and species' responses to environmental changes. *Methods in Ecology and Evolution*, **5**, 388–396.

Bois, S.T., Silander, J.A. & Mehrhoff, L.J. (2011) Invasive plant atlas of New England: the role of citizens in the science of invasive alien species detection. *BioScience*, **61**, 763–770.

Broennimann, O., Treier, U.A., Muller-Scharer, H., Thuiller, W., Peterson, A.T. & Guisan, A. (2007) Evidence of climatic niche shift during biological invasion. *Ecology Letters*, **10**, 701–709.

Buckley, L., Urban, M., Angilletta, M., Crozier, L., Rissler, L. & Sears, M. (2010) Can mechanism inform species' distribution models? *Ecology Letters*, **13**, 1041–1054.

Caticha, A. (2008) *Lectures on probability, entropy, and statistical physics*. http://arxiv.org/abs/0808.0012

Chakraborty, A., Gelfand, A.E., Wilson, A.M., Latimer, A.M. & Silander, J.A., Jr (2010) Modeling large scale species abundance with latent spatial processes. *Annals of Applied Statistics*, **4**, 1403–1429.

Chakraborty, A., Gelfand, A., Wilson, A., Latimer, A.M. & Silander, J.A. (2011) Point pattern modelling for degraded presence only data over large regions. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **60**, 757–776.

Dorazio, R. (2014) Accounting for imperfect detection and survey bias in statistical analysis of presence-only data. *Global Ecology and Biogeography*, **23**, 1472–1484.

Early, R. & Sax, D.F. (2014) Climatic niche shifts between species' native and naturalized ranges raise concern for ecological forecasts during invasions and climate change. *Global Ecology and Biogeography*, **23**, 1356–1365.

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

Elith, J., Phillips, S., Hastie, T., Dudik, M., Chee, Y. & Yates, C. (2011) A statistical explanation of MaxEnt for ecologists. *Diversity and Distributions*, **17**, 43–57.

Engler, R. & Guisan, A. (2009) MigClim: predicting plant distribution and dispersal in a changing climate. *Diversity and Distributions*, **15**, 590–601.

Fithian, W. & Hastie, T. (2013) Finite-sample equivalence of several statistical models for presence-only data. *Annals of Applied Statistics*, **7**, 1917–1939.

Fithian, W., Elith, J., Hastie, T. & Keith, D.A. (2014) Bias correction in species distribution models: pooling survey and collection data for multiple species. *Methods in Ecology and Evolution*, **6**, 424–438.

Franklin, J. (2010) Moving beyond static species distribution models in support of conservation biogeography. *Diversity and Distributions*, **16**, 321–330.

Gallien, L., Münkemüller, T., Albert, C.H., Boulangeat, I. & Thuiller, W. (2010) Predicting potential distributions of invasive species: where to go from here? *Diversity and Distributions*, **16**, 331–342.

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

Guillera-Arroita, G., Lahoz-Monfort, J.J. & Elith, J. (2014) Maxent is not a presence–absence method: a comment on Thibaud *et al. Methods in Ecology and Evolution*, **5**, 1192–1197.

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

Halvorsen, R. (2013) A strict maximum likelihood explanation of MaxEnt, and some implications for distribution modelling. *Sommerfeltia*, **36**, 1–132.

Hastings, A., Cuddington, K., Davies, K.F., Dugaw, C.J., Elmendorf, S., Freestone, A., Harrison, S., Holland, M., Lambrinos, J. & Malvadkar, U. (2005) The spatial spread of invasions: new developments in theory and evidence. *Ecology Letters*, **8**, 91–101.

Higgins, S.I. & Richardson, D.M. (1996) A review of models of alien plant spread. *Ecological Modelling*, **87**, 249–265.

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.

Holmes, E.E. (1993) Are diffusion models too simple? A comparison with telegraph models of invasion. *The American Naturalist*, **142**, 779–795.

Hooten, M.B. & Wikle, C.K. (2007) A hierarchical Bayesian non-linear spatio-temporal model for the spread of invasive species with application to the Eurasian Collared-Dove. *Environmental and Ecological Statistics*, **15**, 59–70.

Ibáñez, I., Silander, J.A., Wilson, A.M., LaFleur, N., Tanaka, N. & Tsuyama, I. (2009) Multivariate forecasts of potential distributions of invasive plant species. *Ecological Applications*, **19**, 359–375.

Jetz, W., McPherson, J.M. & Guralnick, R.P. (2012) Integrating biodiversity distribution knowledge: toward a global map of life. *Trends in Ecology and Evolution*, **27**, 151–159.

Kapur, J.N. & Kesavan, H.K. (1990) The inverse MaxEnt and MinxEnt principles and their applications. *Maximum entropy and Bayesian methods* (ed. by Paul F. Fougere), pp 433–450. Springer, Dordrecht, The Netherlands.

Kearney, M. & Porter, W. (2009) Mechanistic niche modelling: combining physiological and spatial data to predict species' ranges. *Ecology Letters*, **12**, 334–350.

Kesavan, H. & Kapur, J.N. (1989) The generalized maximum entropy principle. *IEEE Transactions on Systems, Man and Cybernetics*, **19**, 1042–1052.

Kullback, S. (1959) *Information theory and statistics*. New York, Dover Press.

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.

Little, J.E. (1971) *Atlas of United States trees, vol. 1. Conifers and important hardwoods*. US Department of Agriculture Miscellaneous Publication 1146.

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.

Merow, C. & Silander, J.A., Jr (2014) A comparison of Maxlike and Maxent for modelling species distributions. *Methods in Ecology and Evolution*, **5**, 215–225.

Merow, C., LaFleur, N., Silander, J.A., Wilson, A.M. & Rubega, M. (2011) Developing dynamic mechanistic species distribution models: predicting bird-mediated spread of invasive plants across northeastern North America. *The American Naturalist*, **178**, 30–43.

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**, 1–12.

Merow, C., Smith, M.J., Edwards, T.C., Jr, Guisan, A., Mcmahon, S.M., Normand, S., Thuiller, W., Wüest, R.O., Zimmermann, N.E. & Elith, J. (2014) What do we gain from simplicity versus complexity in species distribution models? *Ecography*, **37**, 1267–1281.

Okubo, A. & Levin, S.A. (2001) *Diffusion and ecological problems: modern perspectives*. Springer Science & Business Media, New York.

Pagel, J. & Schurr, F.M. (2012) Forecasting species ranges by statistical estimation of ecological niches and spatial population dynamics. *Global Ecology and Biogeography*, **21**, 293–304.

Phillips, S. (2008) Transferability, sample selection bias and background data in presence only modelling: a response to Peterson *et al.* (2007). *Ecography*, **31**, 272–278.

Phillips, S. & Dudik, M. (2008) Modeling of species distributions with Maxent: new extensions and a comprehensive evaluation. *Ecography*, **31**, 161–175.

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

Phillips, S., Dudik, M., Elith, J., Graham, C., 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.

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.

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

Renner, I.W. & Warton, D.I. (2013) Equivalence of Maxent and Poisson point process models for species distribution modeling in ecology. *Biometrics*, **69**, 274–281.

Renner, I.W., Elith, J., Baddeley, A., Fithian, W., Hastie, T., Phillips, S.J., Popovic, G. & Warton, D.I. (2015) Point process models for presence-only analysis. *Methods in Ecology and Evolution*, **6**, 366–379.

Royle, J. & Dorazio, R.M. (2008) *Hierarchical modelling and inference in ecology*. Academic Press, San Diego, CA.

Royle, J.A., Chandler, R.B., Yackulic, C. & Nichols, J.D. (2012) Likelihood analysis of species occurrence probability from presence only data for modelling species distributions. *Methods in Ecology and Evolution*, **3**, 545–554.

Sastre, P. & Lobo, J.M. (2009) Taxonomist survey biases and the unveiling of biodiversity patterns. *Biological Conservation*, **142**, 462–467.

Václavík, T. & Meentemeyer, R. (2009) Invasive species distribution modeling (iSDM): are absence data and dispersal constraints needed to predict actual distributions? *Ecological Modelling*, **220**, 3248–3258.

VanDerWal, J., Shoo, L P, Graham, C. & Williams, 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. & Shepherd, L.C. (2010) Poisson point process models solve the 'pseudo-absence problem' for presence-only data in ecology. *Annals of Applied Statistics*, **4**, 1383–1402.

Warton, D.I., Shipley, B. & Hastie, T. (2014) CATS regression: a model based approach to studying trait based community assembly. *Methods in Ecology and Evolution*, **6**, 389–398.

Wiens, J.J., Ackerly, D.D., Allen, A.P., Anacker, B.L., Buckley, L.B., Cornell, H.V., Damschen, E.I., Jonathan Davies, T., Grytnes, J.-A., Harrison, S.P., Hawkins, B.A., Holt, R.D., Mccain, C.M. & Stephens, P.R. (2010) Niche conservatism as an emerging principle in ecology and conservation biology. *Ecology Letters*, **13**, 1310–1324.

Wisz, M.S. & Guisan, A. (2009) Do pseudo-absence selection strategies influence species distribution models and their predictions? An information-theoretic approach based on simulated data. *BMC Ecology*, **9**, 8.

With, K.A. (2002) The landscape ecology of invasive spread. *Conservation Biology*, **16**, 1192–1203.

Yackulic, C.B., Chandler, R., Zipkin, E.F., Royle, J.A., Nichols, J.D., Campbell Grant, E.H. & Veran, S. (2012) Presence only modelling using Maxent: when can we trust the inferences? *Methods in Ecology and Evolution*, **4**, 236–243.

## SUPPORTING INFORMATION

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

**Appendix S1** Data.
**Appendix S2** Maxent details.
**Appendix S3** Application 1: sampling bias.
**Appendix S4** Application 2: dispersal information.
**Appendix S5** Application 3: using native range data to forecast invasions.
**Appendix S6** Other applications: (a) combining data sets more generally; (b) data: *Protea* in South Africa; (c) using expert range maps; (d) higher-order taxonomic information.
**Figure S1** The sampling scheme can have strong impacts on Maxent predictions.
**Figure S2** Predictions for *Berberis thunbergii* derived from including sampling bias.
**Figure S3** (a) A null model for the distribution of *Berberis thunbergii*, based on a target group. (b) The Minxent prediction for a model including sampling bias, but not native range information. (c) An offset that includes both sampling bias and native range information. (d) The Minxent prediction for a model including both sampling bias and native range data.
**Figure S4** Model response curves for each climate predictor in each range and with the combined offset for *Berberis thunbergii*. The range of climatic conditions in New England is contained within the spectrum of environments in the native range (Japan).
**Figure S5** Illustration of integrating two data sets of different sizes with different sampling bias for *Celastrus orbiculatus*). See Application 2 in the main text for a discussion of why the order of using data sets matters.
**Figure S6** Combining an expert map with presence data for *Protea punctata*.
**Figure S7** Using models for higher order taxa as an offset to predict the distribution of *Protea witzenbergiana*.
**Figure S8** *Protea scolym*.
**Figure S9** *Protea acuminate*.
**Figure S10** *Protea canaliculata*.
**Figure S11** *Protea nana*.
**Figure S12** *Protea pityphylla*.
**Figure S13** *Protea lactiflora*.
**Figure S14** *Protea punctate*.
**Figure S16** *Protea aurea* subsp. *aurea*.
**Figure S17** *Protea venusta*.
**Table S1** Comparison of prediction performance between Maxent and Minxent when using taxonomic offsets for various *Protea* species

## BIOSKETCH

**Cory Merow** is a quantitative ecologist interested in forecasting ecological response to global change.

Editor: Marie-Josée Fortin