TITLE:

A statistical explanation of MaxEnt for ecologists

AUTHORS:

Jane Elith[1], Steven Phillips[2], Trevor Hastie[3], Miroslav Dudík[4], Yung En Chee[1], Colin Yates[5].

ADDRESSES:

[1]School of Botany, The University of Melbourne, Parkville, VIC 3010 Australia

[2] AT&T Labs – Research, 180 Park Avenue, Florham Park, NJ 07932, U.S.A.

[3] Department of Statistics, Stanford University, CA 94305, USA

[4] Yahoo! Labs, 111 West 40th Street (17th Floor). New York, NY 10018

[5] Science Division, Western Australian Department of Environment and Conservation, LMB 104, Bentley Delivery Centre, WA6983, Australia.

KEYWORDS:

species distribution model, ecological niche, presence-only, absence, machine learning, river, banksias

SHORT RUNNING TITLE:

Statistical explanation of MaxEnt

In press. Do not circulate.

ABSTRACT:

MaxEnt is a program for modelling species distributions from presence-only species records. This paper is written for ecologists and describes the MaxEnt model from a statistical perspective, making explicit links between the structure of the model, decisions required in producing a modelled distribution, and knowledge about the species and the data that might affect those decisions. To begin we discuss the characteristics of presence-only data, highlighting implications for modelling distributions. We particularly focus on the problems of sample bias and lack of information on species prevalence. The keystone of the paper is a new statistical explanation of MaxEnt which shows that the model minimizes the relative entropy between two probability densities (one estimated from the presence data and one, from the landscape) defined in covariate space. For many users, this viewpoint is likely to be a more accessible way to understand the model than previous ones that rely on machine learning concepts. We then step through a detailed explanation of MaxEnt describing key components (e.g. covariates and features, and definition of the landscape extent), the mechanics of model fitting (e.g. feature selection, constraints and regularization) and outputs. Using case studies for a *Banksia* species native to south-west Australia and a riverine fish, we fit models and interpret them, exploring why certain choices affect the result and what this means. The fish example illustrates use of the model with vector data for linear river segments rather than raster (gridded) data. Appropriate treatments for survey bias, unprojected data, locally restricted species, and predicting to environments outside the range of the training data are demonstrated, and new capabilities discussed. Online appendices include additional details of the model and the mathematical links between previous explanations and this one, example code and data, and further information on the case studies.

INTRODUCTION

Species distribution models (SDMs) estimate the relationship between species records at sites and the environmental and/or spatial characteristics of those sites (Franklin, 2009). They are widely used for many purposes in biogeography, conservation biology and ecology (Elith & Leathwick, 2009a; Table 1). In the last two decades there have been many developments in the field of species distribution modelling, and multiple methods are now available. A major distinction among methods is the kind of species data they use. Where species data have been collected systematically – for instance, in formal biological surveys in which a set of sites are surveyed and the presence/absence or abundance of species at each site are recorded - regression methods familiar to most ecologists (e.g. generalized linear or additive models, GLMs or GAMs; or ensembles of regression trees: random forests or boosted regression trees, BRT) are used.

However for most regions systematic biological survey data tend to be sparse and/or limited in coverage. Species records are available, though, in the form of presence-only records in herbarium and museum databases. Many of these databases represent well over a century of public and private investment in biological science and are a hugely important resource of species occurrence data. The desire to maximize the utility of such resources has spawned an array of SDM methods for modelling presence-only data. MaxEnt (Phillips *et al.*, 2006; Phillips & Dudík, 2008) is one such method and is the focus of this paper.

MaxEnt's predictive performance is consistently competitive with the highest performing methods (Elith *et al.*, 2006). Since becoming available in 2004, it has been utilized extensively for modelling species distributions. Published examples cover diverse aims (finding correlates of species occurrences, mapping current distributions, and predicting to new times and places) across many ecological, evolutionary, conservation and biosecurity applications (Table 1). Government and non-government organisations have also adopted MaxEnt for large-scale, real-world biodiversity mapping applications, including the Point Reyes Bird Observatory online application (http://www.prbo.org/) and the Atlas of Living Australia (http://www.ala.org.au/). JE and SJP's involvement in such programs identified a need for an ecologically-accessible explanation of

In press. Do not circulate.

MaxEnt. Existing descriptions include concepts from machine learning that tend to be outside the common experience of many ecologists.

In this paper we explain the MaxEnt modelling method with emphases on a statistical explanation of the method, on what it assumes, and on the impacts of choices made in the modelling process. We use two case studies to examine the effects of background selection and model settings, and to illustrate the applicability of the model for exploring ecological relationships with fine-scale, vector-based environmental data. Our aim is to promote understanding of the method and recommend useful approaches to data preparation and model fitting and interpretation.

PREAMBLE: WHAT IS SPECIAL ABOUT THE PRESENCE-ONLY CASE?

Expanding use of presence-only data for modelling species distributions has prompted wide discussion about the sorts of distributions (e.g., potential vs realized) that can be modelled with presence-only data in contrast to presence-absence data (e.g., Soberón & Peterson, 2005; Chefaoui & Lobo, 2007; Hirzel & Le Lay, 2008; Jiménez-Valverde *et al.*, 2008; Soberón & Nakamura, 2009; Lobo *et al.*, 2010). As mentioned in several of these papers, the subject is complex due to the interplay of data quality (amount and accuracy of species data; ecological relevance of predictor variables; availability of information on disturbances, dispersal limitations and biotic interactions), modelling method and scale of analysis. A comprehensive review of the issues would be useful, but here we restrict ourselves to key points important for this paper.

Some of the published discussion suggests that presence-only data in some sense release us from the problems of unreliable absence records (e.g., Jiménez-Valverde *et al.*, 2008), particularly emphasising that absences bear such strong imprints of biotic interactions, dispersal constraints and disturbances that they may preclude modelling of potential distributions (sensu Svenning & Skov, 2004). However, the presence records are also imprinted by many of the factors affecting absences. If a species is absent from an environmentally suitable area because, say, past disturbances have caused local extinctions, the signal of that absence will be found in the distribution of presence records: there will be no presence records in the disturbed area. Regardless of whether absences are used in modelling, the pattern in the presence records will suggest the area is unsuitable, and the model will be affected by this patterning. Similarly, if the detectability of a particular species varies from site to site, then not only does this result in some false absences in presence-absence data, it also affects the pattern of presences in presence-only data. This leads naturally to the conclusion that dispensing with absences does not address the limitations often attributed to absence data, such as the fact that species are not perfectly detectable and may not occupy all suitable habitat. This thinking means that we will approach the description of the presence-only modelling problem as one that is trying to model the same quantity that is modelled with presence-absence data, that is: the probability of presence of a species (to be defined more carefully below).

From here on, we assume that the data available to the modeler are presence-only, i.e. a set of locations within L, the landscape of interest, where the species has been observed. Let y=1 denote presence, y=0 denote absence, $\boldsymbol{z}$ denote a vector of environmental covariates, and background be defined as all locations within L (or a random sample thereof). Assume the environmental variables or covariates $\mathbf{z}$ (representing environmental conditions) are available landscape wide. Define $f(\boldsymbol{z})$ to be the probability density of covariates across L, $f_1(\boldsymbol{z})$ to be the probability density of covariates across locations within L where the species is present, and similarly, $f_0(\boldsymbol{z})$ where the species is absent. (Densities – or probability density functions - describe the relative likelihood of random variables over their range, and can be univariate or multivariate). The quantity that we wish to estimate is, as with presence-absence data, the probability of presence of the species, conditioned on environment: $\Pr(y=1|\boldsymbol{z})$. Strictly presence-only data only allows us to model $f_1(\boldsymbol{z})$, which on its own cannot approximate probability of presence. Presence/background data allows us to model both $f_1(\boldsymbol{z})$ and $f(\boldsymbol{z})$, and this gets to within a constant of $\Pr(y=1|\boldsymbol{z})$, because Bayes' rule gives:

In press. Do not circulate.

$$\Pr(y=1|\mathbf{z}) = f_1(\boldsymbol{z})\Pr(y=1) \,/\, f(\boldsymbol{z}). \qquad \text{.......................... (1)}$$

The only quantity that is lacking is the second term, Pr(y=1), i.e. the prevalence of the species (proportion of occupied sites) in the landscape. Formally, we say that prevalence is not identifiable from presence-only data (Ward *et al.* 2009). This means that it cannot be exactly determined, regardless of the sample size; this is a fundamental limitation of presence-only data. As an aside we note, however, that absence data are plagued by issues of detection probability (Wintle *et al.*, 2004; MacKenzie, 2005) so that even presence-absence data may not yield a good estimate of prevalence. A second fundamental limitation of presence-only data is that sample selection bias (whereby some areas in the landscape are sampled more intensively than others) has a much stronger effect on presence-only models than on presence-absence models (Phillips *et al.*, 2009). Imagine that $f_1(\boldsymbol{z})$ is contaminated by a sample selection bias $s(\boldsymbol{z})$. This bias will most commonly occur in geographic space (e.g. close to roads) but could be environmentally based (e.g. visiting wet gullies) but, regardless, will map into covariate space. Under biased sampling, a presence-only model gives an estimate of $f_1(\boldsymbol{z})s(\boldsymbol{z})$ rather than $f_1(\boldsymbol{z})$. That is, we get a model that combines the species distribution with the distribution of sampling effort (Soberón & Nakamura, 2009). In contrast, for presence-absence models, sample selection bias affects both presence and absence records, and the effect of the bias cancels out (under reasonable assumptions, see Zadrozny, 2004).

So far we have treated presence or absence as a binary event, but in reality defining the response variable is not straightforward, and in this regard presence-only data are quite different from presence-absence data (Pearce & Boyce, 2006). Presence or absence of a species is dependent on the time frame and spatial scale -- for example, a vagile species (such as a bird) may be present at some times but not others, while a plant species will be more likely to be found in a large plot with given environmental conditions than in a small plot with the same conditions. Absence of a plant species from a 1km^2 quadrat around a point implies absence in a 1m^2 quadrat around that point, but not vice versa. With presence-absence data, it is not hard to incorporate these complexities in the formulation of the response variable (i.e., the specification of what constitutes a sample), or via sampling covariates in the model, provided survey details are available (Leathwick, 1998; MacKenzie & Royle, 2005; Schulman *et al.*, 2007; Ward, 2007b). However, with presence-only data, we typically have occurrence data that do not have any associated temporal or spatial scale. The record is usually simply a record of the species at a location, with no information on search area or time.

With presence-absence data, the definition of the response variable should naturally be consistent with the sampling method: for example, if the available data are surveys of 1m^2 quadrats, then y=1 should correspond to the species being present in a 1m^2 quadrat. With presence-only data, the available data do not usually describe the survey method, so the modeller has considerable leeway in defining the response variable. A common approach is to implicitly assume a sampling unit of size equal to the grain size of available environmental data (see Elith & Leathwick 2009a for discussion of grain).

To summarize, we posit that with presence and background data, we can model the same quantity as with presence-absence data, up to the constant Pr(y=1). However, if presence-absence survey data are available, we believe it is generally advisable to use a presence-absence modelling method, since in that case the models are less susceptible to problems of sample selection bias, the survey method will often be known and can be used to appropriately define the response variable for modelling, and we take advantage of all information in the data. In particular, presence-absence data give us much better information about prevalence than presence-only, because – even though there may be some difficulties due to imperfect detection - they solve the major problem of non-identifiability. We will come back to this when we discuss the logistic output of MaxEnt.

In press. Do not circulate.

EXPLANATION OF MAXENT

Here for the first time we describe MaxEnt using statistical terminology and notation, providing a break from the machine learning terminology in previous papers. As we describe the model we will highlight possibilities for – and implications of - modelling choices and defaults, and consider how MaxEnt addresses the limitations of presence-only data identified above. We relegate the more technical considerations to boxes and online appendices, to avoid interrupting the flow of the explanation.

COVARIATES AND FEATURES

Most ecologists, following the statistical literature, call the independent variables in a model the covariates, predictors or inputs. In SDMs these include environmental factors that are relevant to habitat suitability (e.g. estimates of climate, topography, and soil for plants; temperature, salinity and prey abundance for marine fishes). Since species' responses to these tend to be complex, it is usually desirable to fit non-linear functions (Austin, 2002). In regression, this can be achieved by applying transformations to the covariates – for instance, creating basis functions for polynomials and splines, including piecewise linear functions. Complex models are fitted as linear combinations of these basis functions in methods including GLMs and GAMs (Hastie *et al.* 2009, Chapter 5). In machine learning, basis functions and other transformations of available data are termed features – i.e., features are an expanded set of transformations of the original covariates.

In MaxEnt selected features are formed “behind the scenes”, in the same way as in regression, where the model matrix is augmented by terms specified in the model (e.g. polynomials, interactions). The MaxEnt fitted function is usually defined over many features, meaning that in most models there will be more features than covariates. MaxEnt currently has six feature classes: linear, product, quadratic, hinge, threshold and categorical (further details in Online Appendix 1). Products are products of all possible pair-wise combinations of covariates, allowing simple interactions to be fitted. Threshold features allow a "step" in the fitted function; hinge features are similar except they allow a change in gradient of the response. Many threshold or hinge features can be fitted for one covariate, giving a potentially complex function. Hinge features (which are basis functions for piecewise linear splines), if used alone, allow a model rather like a generalized additive model (GAM): an additive model, with non-linear fitted functions of varying complexity but without the sudden steps of the threshold features. MaxEnt’s default is to allow all feature types (conditional on sufficient species data being available), but it is worth considering simpler models, as discussed later under implications for modelling.

THE MAXENT MODEL – A SHORT OVERVIEW

Previous papers have described MaxEnt as estimating a distribution across geographic space (Phillips *et al.*, 2006; Phillips & Dudík, 2008). Here we give a different (but equivalent) characterization that focuses on comparing probability densities in covariate space (Figure 1). In doing so we rely strongly on the PhD research of TH's past student, Gill Ward (Ward, 2007b), and acknowledge her contribution. Equation 1 shows that if we know the conditional density of the covariates at the presence sites, $f_1(\boldsymbol{z})$, and the marginal (i.e. unconditional) density of covariates across the study area $f(\boldsymbol{z})$, we then only need knowledge of the prevalence Pr(y=1), to calculate conditional probability of occurrence. MaxEnt first makes an estimate of the ratio $f_1(\boldsymbol{z})/ f(\boldsymbol{z})$, referred to as MaxEnt’s “raw” output. This is the core of the MaxEnt model output, giving insight about what features are important, and estimating the relative suitability of one place vs another. Because the required information on prevalence is not available for calculating conditional probability of occurrence, a work-around has been implemented (termed MaxEnt’s “logistic” output). This treats the log of the output --- $\eta(\mathbf{z})=\log(f_1(\mathbf{z})/f(\mathbf{z}))$ --- as a logit score, and calibrates the intercept so that the implied probability of presence at sites with "typical" conditions for the species (i.e. where $\eta(\mathbf{z})$ = the average value of $\eta(\mathbf{z})$ under f_1) is a parameter τ. Knowledge of τ would solve the non-identifiability of prevalence, and in the absence of that knowledge MaxEnt arbitrarily

In press. Do not circulate.

sets τ to equal 0.5. This logistic transformation is monotone (order preserving) with the raw output. We work through each part of the MaxEnt model in the following sections, showing how the choice of landscape, species data, and selected settings influence the results.

THE LANDSCAPE AND SPECIES RECORDS

The landscape of interest (L) is a geographic area suggested by the problem and defined by the ecologist. It might, for instance, be limited by geographic boundaries, or by an understanding of how far the focal species could have dispersed. We then define L_1 as the subset of L where the species is present.

The distribution of covariates in the landscape is conveyed by a finite sample – a collection of points from L with associated covariates, typically called a background sample. These data may be supplied in the form of grids of covariates covering a pixelation of the landscape; as a default MaxEnt randomly samples 10,000 background locations from covariate grids, but the background data points can also be specified (see Yates *et al.*, 2010, and case studies below) and grids are not essential (case study 2). Note that the background sample does not take any account of the presence locations – it is simply a sample of L, and could by chance include presence locations. Using a random background sample implies a belief that the sample of presence records is also a random sample from L_1. We deal later with the case of biased samples.

DESCRIPTION OF THE MODEL

MaxEnt uses the covariate data from the occurrence records and the background sample to estimate the ratio $f_1(\boldsymbol{z})/f(\boldsymbol{z})$. It does this by making an estimate of $f_1(\boldsymbol{z})$ that is consistent with the occurrence data; many such distributions are possible, but it chooses the one that is closest to $f(\boldsymbol{z})$. Minimizing distance from $f(\boldsymbol{z})$ is sensible, because $f(\boldsymbol{z})$ is a null model for $f_1(\boldsymbol{z})$: without any occurrence data, we would have no reason to expect the species to prefer any particular environmental conditions over any others, so we could do no better than predict that the species occupies environmental conditions proportionally to their availability in the landscape. In MaxEnt, this distance from $f(\boldsymbol{z})$ is taken to be the relative entropy of $f_1(\boldsymbol{z})$ with respect to $f(\boldsymbol{z})$ (also known as the Kullback-Leibler divergence).
Using background data informs the model about $f(\boldsymbol{z})$, the density of covariates in the region, and provides the basis for comparison with the density of covariates occupied by the species – i.e. $f_1(\boldsymbol{z})$. Constraints are imposed so that the solution is one that reflects information from the presence records. For example, if one covariate is summer rainfall, then constraints ensure that the mean summer rainfall for the estimate of $f_1(\boldsymbol{z})$ is close to its mean across the locations with observed presences. The species' distribution is thus estimated by minimizing the distance between $f_1(\boldsymbol{z})$ and $f(\boldsymbol{z})$ subject to constraining the mean summer rainfall estimated by f_1 (and the means of other covariates) to be close to the mean across presence locations.

We note that previous papers describing MaxEnt focused on a location-based definition over a finite landscape (typically a grid of pixels). We will call this a definition based in geographic space, and compare it with our new description, which focuses on environmental (covariate) space. Note, though, that we are not implying by this wording that in either definition there is any consideration of the geographic proximity of locations unless geographic predictors are used. In the original definition (Phillips *et al.* 2006), the target was $\pi(x) = p(x|y=1)$, which was a probability distribution over pixels (or locations) x. This was called the "raw" distribution (Phillips *et al.* 2006), and gave the probability, given the species is present, that it is found at pixel x. Maximising the entropy of the raw distribution is equivalent to minimizing the relative entropy of $f_1(\boldsymbol{z})$ relative to $f(\boldsymbol{z})$, so the two formulations are equivalent (see online Appendix 2 for equations showing the transition from the geographic to environmental definitions). The null model for the raw distribution was the uniform distribution over the landscape, since without any data we would have no reason to think the species would prefer any location to any other. In environmental space the equivalent null model for $\boldsymbol{z}$ is $f(\boldsymbol{z})$, since without any data, we have no reason to think the species prefers any particular

In press. Do not circulate.

environmental conditions, and therefore occupies environmental conditions in proportion to how prevalent they are in the landscape.

Constraints were described above in reference to covariates, but – as explained in the section on covariates and features - MaxEnt actually fits the model on features that are transformations of the covariates. These allow potentially complex relationships to be modelled. The constraints are extended from being constraints on the means of covariates to being constraints on the means of the features. We will call the vector of features $h(\boldsymbol{z})$ and the vector of coefficients β (note, this notation is different to previous papers: Table 2). As explained in Phillips *et al.* (2006), minimizing relative entropy results in a Gibbs distribution (Della Pietra *et al.*, 1997) which is an exponential-family model:

$$f_1(\boldsymbol{z}) = f(\boldsymbol{z})\, e^{\eta(z)} \qquad \text{..............................} \ (2)$$

where $\eta(\boldsymbol{z}) = \alpha + \beta \cdot h(\boldsymbol{z})$
and α is a normalizing constant that ensures that $f_1(\boldsymbol{z})$ integrates (sums) to 1

From this it is clear that the target of a MaxEnt model is $\eta(\boldsymbol{z})$, which estimates the ratio $f_1(\boldsymbol{z})/f(\boldsymbol{z})$. It is a log-linear model, similar in form to a GLM, and depends on both the presence samples and the background samples that are used in forming the estimate. Hence the definition of the landscape is intimately linked to the solution that is given.

MECHANICS OF THE SOLUTION

In coming to a solution MaxEnt needs to find coefficients (betas) that will result in the constraints being satisfied but not match them so closely that it overfits and produces a model with limited generalization. MaxEnt handles the issue by setting an error bound, or maximum allowed deviation from the sample (empirical) feature means. MaxEnt first automatically rescales all features to have the range 0 to 1. Then an error bound (λ_j in equation 3) is calculated for each feature (again note the change in notation from previous papers, Table 2). It will reflect the variation in sample values for that feature, adjusted by a tuned (pre-set) parameter for the feature class (Phillips and Dudík 2008, and equation 3). MaxEnt *could* estimate feature error bounds only from the data, for example using cross-validation, but to simplify model fitting and because the data are often biased, it uses feature class-specific tuned parameters based on a large international data set (Phillips & Dudík, 2008). That dataset covers 226 species, 6 regions of the world, sample sizes ranging from 2 to 5822, and 11-13 predictors per region (Elith *et al.* 2006). It is possible that the tuning may not work well for very different datasets – e.g. if there are many more predictors. The tuned parameters can be changed by the user if desired. The pre-tuning also includes restrictions to the set of feature classes that will be considered for small samples.

$$\lambda_j = \lambda \sqrt{\frac{s^2[h_j]}{m}} \qquad \text{.............} \ (3)$$

where λ_j is the regularization parameter for feature h_j. This feature's variance is s^2 over the m presence sites, and its feature class has a tuning parameter λ. Conceptually λ_j corresponds to the width of the confidence interval, therefore it takes the form of the standard error (the square root expression) multiplied by the parameter λ according to the desired confidence level.

The lambdas in equation 3 allow regularization – i.e. smoothing the distribution, making it more regular. These error bounds are a specific form of regularization called L1-regularization (Tibshirani, 1996) that gives sparse solutions (ones with many zeros, i.e. many features removed). Regularization is not specific to MaxEnt; it is a common modern approach to model selection. It can

In press. Do not circulate.

be thought of as a way of shrinking the coefficients (the betas) – i.e. penalizing them - to values that balance fit and complexity, allowing both accurate prediction and generality. In MaxEnt, the fit of the model is measured at the occurrence sites, using a log likelihood (Box 1). A highly complex model will have a high log likelihood, but may not generalize well. The aim of regularization is to trade off model fit (the first term in equation 4 below) and model complexity (the second term in equation 4). In this sense MaxEnt fits a penalized maximum likelihood model (Phillips and Dudík 2008; equation 4) closely related to other penalties for complexity such as Akaike's Information Criterion (AIC, Akaike, 1974). Maximising the penalized log likelihood is equivalent to minimising the relative entropy subject to the error bound constraints.

$$\max_{\alpha,\beta} \frac{1}{m}\sum_{i=1}^{m} \ln\left(f(z_i)e^{\eta(z_i)}\right) - \sum_{j=1}^{n} \lambda_j|\beta_j|$$

.... equation 4

subject to $\int_L f(z)e^{\eta(z)}dz = 1$

Where z_i is the feature vector for occurrence point i of m sites,
and for j = 1 … n features

------------ Box 1 - Log likelihood -------------

In statistics a log likelihood describes the log of the probability of an observed outcome. It varies from 0 [ln(1)] to negative infinity [ln(0)]. If the space of outcomes is continuous, we measure the probability density at the observed outcome, rather than probability. With presence-only data the only known outcomes are presences, so when measuring likelihoods, the calculation is simply done at the presence sites (in comparison to logistic regression where they are calculated at presence and absence sites). For a set of observations the average log likelihood is estimated. When fitting a MaxEnt model from the software interface, a gain bar is shown that reports the improvement in penalized average log likelihood compared to a null model.

--------------- end of Box 1 --------------------

MAXENT'S LOGISTIC OUTPUT

MaxEnt (from version 3 onwards) gives a logistic output as its default. It is an attempt to get as close as we can to an estimate of the probability that the species is present, given the environment, $\Pr(y = 1|\boldsymbol{z})$. This is a post-transformation of the MaxEnt raw output that makes certain assumptions about prevalence and sampling effort (Box 2 and Online Appendix 3). These two output types of MaxEnt (raw and logistic) are monotonically related, so if the purpose of a study is to rank sites according to suitability, it does not matter which type is used – both will yield identical ranking and hence identical rank-based measures (e.g., AUC values). MaxEnt's logistic transformation is not a commonly used statistical procedure, so here we explain the background and the issues.

From Eqn. 1 we see that a simple approach to estimate $\Pr(y=1|\boldsymbol{z})$ would be to simply multiply $e^{\eta(\mathbf{z})}$ by a constant that estimates prevalence; this approach has the disadvantage that $e^{\eta(\mathbf{z})}$ can be arbitrarily large, which implies that we may get an estimate of $\Pr(y = 1|\mathbf{z})$ that exceeds 1 (Keating & Cherry, 2004; Ward, 2007b). Exponential models can be especially badly behaved when applied to new data, for instance, when extrapolating to new environments. To avoid these problems, and to side-step the non-identifiability of the species prevalence, Pr(y=1), MaxEnt's logistic output transforms the model from an exponential family model (Eqn. 2) to a logistic model:

In press. Do not circulate.

$$\Pr(y = 1|\mathbf{z}) = \tau e^{\eta(\mathbf{z})-r} / (1-\tau + \tau e^{\eta(\mathbf{z})-r}) \qquad \text{... Equation 5}$$

where $\eta(\mathbf{z})$ is the linear score from Eqn. 2, r is the relative entropy of MaxEnt's estimate of $f_1(\mathbf{z})$ from $f(\mathbf{z})$, and τ is the probability of presence at sites with "typical" conditions for the species (i.e. where $\eta(\mathbf{z})$ = the average value of $\eta(\mathbf{z})$ under f_1). The default value for τ is arbitrarily set at 0.5. Equation 5 is derived using a "minimax" or robust Bayes approach (details in Online Appendix 3). In unsuitable areas, the logistic output's denominator is close to 1-τ, so the result is just a linear scaling of raw output. For more suitable areas, the effect of the denominator is mainly to bound model output below 1. The logistic output with τ = 0.5 empirically gives a better calibrated estimate of Pr(y = 1|**z**) than the untransformed raw values (Phillips and Dudík, 2008).

Because the species prevalence, Pr(y=1), is not identifiable from occurrence data, the prevalence Pr(y=1) implied by the logistic output will not converge to the true prevalence, even given ample occurrence data. On the other hand, the true prevalence depends on the definition of the response variable y, which itself depends on the sampling method - often unknown for presence-only data (see Preamble). Further, if additional information is available that could be used to estimate τ, prevalence will be identifiable. We therefore offer guidance for interpretation of MaxEnt's logistic output in relation to sampling effort and τ (Box 2).

------- Box 2: Consider the jaguar: reconciling logistic output and sampling effort -----------

The jaguar (*Panthera onca*) and the collared peccary (*Pecari tajacu*) have very similar ranges in South and Central America, and MaxEnt models for the two species would therefore be similar using the default τ. However, the jaguar is much rarer than the peccary, so how can the outputs be compared? The answer is that probability of presence is only defined relative to a given definition of presence/absence (i.e., the temporal and spatial scale of a sample; see Preamble). For instance, for a rare species like the jaguar a presence record is likely to derive from sampling over a longer time and/or larger area (e.g. using camera traps over months) than it would for the peccary, which is fairly common and easier to observe. Since with presence-only data there is usually no information on sampling effort, this elasticity in definition is largely conceptual – it explains how to think about the meaning of the probabilities across species, when τ is 0.5. When τ is 0.5 typical presence sites will have a logistic output near 0.5. This is reasonable as long as we can interpret logistic output as corresponding to a temporal and spatial scale of sampling that results in a 50% chance of the species being present in suitable areas. See Online Appendix 3 for more information.

Alternatively, if the value of τ is available for a given level of sampling effort, it could be used instead of the default, and then the predictions for the two species would be directly comparable. Tau measures a form of rarity (Rabinowitz *et al.* 1986). The jaguar has very low local abundance even in suitable areas within its range, so a very small value τ is appropriate for all but the most intensive sampling schemes. The estimate of τ could come from expert knowledge or targeted surveys. While τ is determined by prevalence, and vice versa, τ is arguably more ecologically intuitive, as it is more a property of the species while prevalence strongly depends on the choice of study area.

------------------------- end of Box 2 ----------------------

IMPLICATIONS FOR MODELLING

These properties of the MaxEnt model have several implications for how it should be used. MaxEnt relies on an unbiased sample (as do all species modelling methods), so efforts in collecting a comprehensive set of presence records (cleaned for duplicates and errors) and dealing with biases are critical (Newbold, 2010). Methods are implemented for dealing with biased species data (see case study 1, and Dudík *et al.*, 2006; Phillips *et al.*, 2009; Elith *et al.*, 2010 in press). The main alternatives are to provide background data with similar biases to those in the presence data (e.g. by using sites surveyed for other species in the same biological group), or to use a bias grid that

In press. Do not circulate.

indicates the biases in the survey data (see tutorial provided with MaxEnt for an example). All the values in this grid should be positive (or specified as no data), and should be scaled to represent relative survey effort across the landscape *L*. There is one additional important consideration. If the covariate grids are unprojected (i.e. latitude and longitude in degrees, for instance WorldClim data - http://www.worldclim.org/), any region covering a non-trivial range in latitude (say, more than 200km, especially away from the equator) will have grid cells of varying area. For instance, in Australia cells in the north are approximately 1.3 times the area of cells in the south. MaxEnt randomly samples cells, implicitly assuming equal area cells. Solutions are to project the grids to an equal area projection, create a grid showing the variations in cell area, that can then be used as a bias grid, or create your own background sample with appropriate sampling weights (case study 1).

The MaxEnt solution is affected by the landscape (region) used for the background sample, as demonstrated by VanDerWal *et al.* (2009). Conceptually, that landscape should include the full environmental range of the species, and exclude areas that definitely have not been searched (unless the reason for no searching is that there is unambiguous knowledge that the species does not occur there). A local endemic that is, for instance, likely to be geographically restricted due to barriers to dispersal, should be modelled with background selected from areas into which it might have dispersed. Cleared areas that would not be surveyed because there is no remaining habitat for the species should be excluded. Excluding areas from the background sample can be achieved through use of masks, as explained in the online tutorial for MaxEnt (and see Table 2). Predictions can still be made to excluded areas, if required, by using the projection facilities. We will discuss some caveats to these general concepts for background selection in the first case study.

MaxEnt includes a range of feature types, and subsets of these can be used to simplify the solution. By default, the program restricts the model to simple features if few samples are available (linear is always used; quadratic with at least 10 samples; hinge with at least 15; threshold and product with at least 80) because – as for any modelling method – few samples provide limited information for determining the relationships between the species and its environment (Barry & Elith, 2006; Pearson *et al.*, 2007). In such cases, it is also a good idea to first reduce the candidate predictor set using ecological understanding of the species (Elith & Leathwick, 2009b). Hinge features tend to make linear and threshold features redundant, and one way to form a model with relatively smooth fitted functions, more like a GAM, is to use only hinge features (e.g. Elith *et al.*, 2010 in press, and case study 1). Excluding product features creates an additive model that is easier to interpret, though less able to model complex interactions.

MaxEnt has an inbuilt method for regularization (L1-regularization) that is reliable and known to perform well (Hastie *et al.*, 2009). It implicitly deals with feature selection (relegating some coefficients to zero) and is unlikely to be improved - and more likely, degraded - by procedures that use other modelling methods to pre-select variables (e.g. Wollan *et al.*, 2008). In particular, it is more stable in the face of correlated variables than stepwise regression, so there is less need to remove correlated variables (unless some of them are known to be ecologically irrelevant), or preprocess covariates by using PCA and selecting a few dominant axes. Note, though, that since there are often many variables available, some expert pre-selection of a candidate set is often a good idea; Elith and Leathwick 2009b. Selecting proximal variables is likely to be particularly important when models are to be used in different regions or climates. If smoother models are required, regularization parameters can be increased by the user (e.g., see Elith *et al.* 2010 in press).

If comparing models for different species some care is needed in use of the logistic outputs because probability of presence is only defined relative to a given level of sampling effort, which as a default is assumed to be one that results in a 50% chance of observing the species in suitable areas (Box 2). The implied sampling effort therefore depends on the species. This presents some challenges for cross-species comparisons of habitable areas, but these are a direct result of using presence-only data, and is not a unique problem to MaxEnt. Some users may in fact see the species-specific scaling

In press. Do not circulate.

as an opportunity, since the literature on favourability functions (e.g., Real *et al.*, 2006) claims that probability of presence is itself hard to work with.

USING MAXENT

CASE STUDY 1: MODELLING CURRENT AND FUTURE DISTRIBUTIONS OF A PLANT

This analysis predicts the current distribution of *Banksia prionotes*, then uses the model to identify where suitable environments for the species are likely to occur under climate change. In it we highlight the importance of choice of landscape and dealing with survey bias, debiasing background samples from unprojected grids, use of a reduced set of feature types for a smoother model, and tools for assessing the environments in new times or places.

B. prionotes is a woody shrub to small tree native to south west Western Australia (WA). It is widely distributed across its range, and shows a preference for deep sandy soils. Often a dominant plant in scrubland and low woodlands, it is an important nectar source for honeyeaters, and an outstanding ornamental species for cut flowers.

<u>Methods</u>: Here we use species data from the Banksia Atlas (Taylor & Hopper, 1988; Yates *et al.*, 2010), with 361 records for *B. prionotes* from the 4631 sites across the South West Australia Floristic Region (SWAFR) that were surveyed for *Banksia* and for which we had complete environmental data. The atlas is the result of a community science project, and records could either be interpreted as presence-only or presence-absence data, depending on what assumptions are made about the search patterns of contributors. Here we treat them as presence-only data, but use the full set of locations as one "background" treatment. To demonstrate the effect of this choice, two alternative backgrounds (i.e. landscape definitions) were evaluated: a sample of 10000 sites within the SWAFR (Yates *et al.* 2010, and Figure 2), and a sample of 20000 sites across the whole of Australia. The larger number of sites across Australia was used to ensure good representation of all environments, based on previous tests of the effects of background sample size on model structure for these predictors (Elith unpubl.). Because the covariate data for this study are unprojected, these samples were weighted according to cell area (see methods in Online Appendix 4) but otherwise random.

Using random sites within the floristic region implies that the presence records are a random sample from all locations where the species is present in the region which is unlikely because records were from extant vegetation patches in likely suitable environments (the region has been extensively cleared for agriculture, and some of the more inland areas are too arid for many *Banksia* species). Using random sites across Australia implies the species could have dispersed anywhere across the continent, and the whole continent considered available for sampling. This is questionable because the desert areas to the north and east of the inhabited area are likely barriers to dispersal. We will come back to implications of this later.

Yates *et al.* (2010) identified important climatic drivers for plants of southwest Western Australia. We base our candidate set of predictors on their study, but use a different data source so we can train and predict over the whole of Australia. Described in online Appendix 4, our covariates (all unprojected, at 0.01 degree or ~ 1km grid resolution) included five climate variables: isothermality (ISOTHERM), mean temperature of the wettest quarter (TEMPWETQ), mean temperature of the warmest quarter (TEMPWARMQ), annual precipitation (RAIN) and precipitation of the driest quarter (RAINDRYQ), and an estimate of the solum plant-available water holding capacity (solwhc). We present this as a demonstration study only, and recognize that, for rigorous application in this region, better soils data and predictors representing land transformation are needed for more precise predictions (Yates *et al.* 2010). The future environment was represented by changes predicted under the A1FI scenario for 2070 estimated over the ensemble of 23 GCMs in IPCC AR4 (Solomon et al. 2007); the solwhc was assumed to remain as it is now.

In press. Do not circulate.

Models were fitted and projected to both current and future climates (Figure 3) using only hinge features, with default regularization parameters (see Appendix 5 for model details, and for a comparison with models fitted with all feature types). We fitted all models on the full data sets but also used 10-fold cross validation to estimate errors around fitted functions and predictive performance on held-out data. The latter is a good test for each model but – given the different backgrounds – not comparable across models. Note also that the AUC in this case is calculated on presence vs background data (Phillips *et al.* 2006). We also divided the atlas data into training and testing sets for a manual 5-fold cross-validation, testing each model on identical withheld data via two test statistics (area under the receiver operating characteristic curve (AUC), and correlation, COR; details in online Appendix 4). Example code for running such analyses are available online (Appendix 4).

<u>Results:</u> Atlas background (model 1) produced a mapped distribution in the inhabited region with more of an eastward emphasis compared with other background treatments (Figure 3). The coastward (westerly) bias in the distribution of survey sites (Figure 2) affected the distributions predicted by models 2 and 3 (random background across SWAFR or Australia) but was factored out by using atlas background (model 1). The more easterly distribution is more consistent with the known ecology of the species, and with the observed distribution (Taylor and Hopper 1988). Variable importance varies with data set, with TEMPWETQ being much more prominent when using an all-Australia background than when restricted to the south-west. Similarly, shapes of fitted functions vary across data sets (Appendix 5). This is to be expected, because each data set implies a different modelling question (e.g. the all-Australia background asks: why is this species only in environments occurring in the southwest?).

An increasing number of SDM applications involve prediction to new environments (e.g. to new places or times; Elith & Leathwick, 2009a). These are contentious applications, making strong assumptions (Dormann, 2007) and usually requiring prediction to environments not sampled by the training data. MaxEnt has been extended to include new capabilities to inform users about predicting to novel environments (Elith *et al.*, 2010 in press). MaxEnt already provides mapped information on the effect of model "clamping" – i.e. the process by which features are constrained to remain within the range of values in the training data. This identifies locations where predictions are uncertain due to the method of extrapolation, by showing where clamping substantially affects the predicted value. We feel that extreme care should be taken whenever extrapolating outside the training, so new calculations ("MESS maps", i.e. multivariate environmental similarity surfaces) display differences between the training and prediction environments (Figure 3). In this case they show that, compared with environments at the atlas sites, the northern parts of the SWAFR will experience novel climates in 2070 (Figure 3 model #1). Models based on random background across SWAFR or the continent (models 2 and 3) require less extrapolation (because wider sampling of background points brings with it wider sampling of environments) but, given the problems with the realism of these treatments, we do not view the result as a necessary advantage for future predictions.

Online Appendices 5 and 6 include further information on how these models predict across the continent, for both current and future climates. They provide interesting insights into model variation across scales, regions, and datasets, and emphasize the importance of choice of background (see commentary, Appendix 5). In particular, it is interesting that model 3 restricts predictions to the correct general area, and has the highest 10-fold cross-validated AUC (Table 3), yet has the poorest ecological justification for its choice of background and is least likely to be useful for managing the species locally. The advantage of limiting background to local, reachable areas (models 1 and 2) is that contrasts between occupied and unoccupied environments in the local area are the model focus, and – particularly with fine-scale environmental data – differentiation useful at the management scale might be achievable. It is also likely to be the most ecologically realistic choice for many locally restricted species. On the other hand, if models are to be projected well outside the local geographic area, use of local backgrounds brings with it the

In press. Do not circulate.

penalty that prediction to other areas is likely to involve considerable extrapolation. Some trade-off is clearly required.

Case study 2: Modelling the distributions of fish in rivers.

This analysis predicts the current distribution of *Gadopsis bispinosus,* the two-spined blackfish, in rivers of south-eastern Australia. In the preamble we make a case that with presence and background data, we can model the same quantity as with presence-absence data, up to the constant Pr(y=1). One implication of that is that we should be able to use the same types of data, including fine-scale, detailed information, to model ecological relationships – i.e. we need not be restricted to coarse grid cells and basic climate variables. Here we use detailed ecological information at the river segment scale to model the distribution of a native fish species. To our knowledge it is the first example using MaxEnt with vector (river segment) data.

G. bispinosus is a native freshwater fish endemic to south-eastern Australia. It occurs in cool, clear upland or montane streams with abundant in-stream cover. It is most common in medium to large streams that are deep enough for reduced stream velocities, and in forested catchments with relatively small sediment inputs (Lintermans 2000).

<u>Methods:</u>
The species data are from surveys (described further in online Appendix 7) of the inland-draining rivers of northwest Victoria, Australia. In this area there are ten major river systems grouped into four regions that start in hilly to mountainous terrain and drain northwards. *G. bispinosus* was recorded at 255 sites. We use covariate data from the 255 capture sites as our sample of L_1 and a random sample of 10000 of the ~240000 river segments for our sample of L , the background data.

The candidate predictor set comprised 20 variables summarizing information across three hierarchically nested spatial scales (segment, immediate watershed and entire upstream catchment area) and also downstream to the large river system draining to the ocean. The environmental variables estimate climate, river slope, riparian vegetation and catchment characteristics (Online Appendix Table S7.1). River system was also included to quantify spatial variation in land characteristics and disturbances not covered by the environmental predictor set.

These segment-based (non-gridded) data are modelled using the SWD (samples-with-data) format in MaxEnt – this involves presenting spreadsheet-like summaries of environments at both presence and background sites. All environmental variables were continuous except the categorical river system covariate. Default settings for features and regularization were used for model training, and 10-fold cross-validation used to obtain out-of-sample estimates of predictive performance and estimates of uncertainty around fitted functions. For mapping, the model was projected to a selected area in the Goulburn-Broken catchment. Technically, this was achieved by projecting to SWD format data, then linking the predictions to the relevant river segments in a GIS. Online Appendix 8 includes data and code for replicating this case study, including information on how to run MaxEnt from batch files.

<u>Results:</u> Consistent with ecological knowledge about the species, the model predicts *G. bispinosus* will most frequently occur in the larger streams of montane areas (Figure 4). These locations are identified as those whose upstream catchments have relatively more precipitation in the warmest quarter and steeper maximum stream slopes. Amongst these, emphasis on segments with warmer summer maximum temperatures served to exclude the higher elevation cold streams (Figure 5). Jackknife tests of variable importance help to identify those with important individual effects; the three most important single predictors were the summed length of all upstream links (TOTLENGTH_UCA), the upstream maximum slope (US_MAXSLOPE) and the amount of riparian tree cover upstream (UC_RIP_TRECOV); and the predictor with the most information not present in the other variables is the segment-based maximum temperature of the warmest month

In press. Do not circulate.

(MAXWARMP_TEMP). Many predictors had small to minimal impacts in the final model. The model shows strong discrimination on held out data, with a cross-validated AUC of 0.97.

<u>Extensions / alternatives</u>: Since records on one river system might share a more similar environment than those on different systems, an alternative approach to cross-validation would be to test the predictions iteratively on held-out rivers. We chose not to do it in this case, because presence records were concentrated in relatively few river systems, so the training sets would be substantially reduced, and the test sets, relatively few.

CONCLUSIONS

Here we have described MaxEnt from a statistical viewpoint, showing that the model minimizes the relative entropy between two probability densities defined in feature space. An understanding of the model leads naturally to recommendations for implementation, and ours included the importance of providing appropriate background samples, of dealing with sample biases, and of tuning the model – through feature type selection and regularization settings - to suit the data and application. Presence-only data are a valuable resource and potentially can be used to model the same ecological relationships as with presence-absence data, provided that biases can be dealt with and except for the non-identifiability of prevalence.

MaxEnt is regularly updated, usually to include new capabilities to suit the expanding applications, and also sometimes to change the program defaults to those most often used in practice. Recent new capabilities include the cross-validation and MESS maps (i.e. estimates of how the environmental space in predicted times and places compares with that of the training data) demonstrated in case study 1. In addition, new clickable maps allow users to interrogate predictions spatially, providing information for any grid cell on the components of the prediction (i.e. what contributes to its particular value) and where the environmental conditions "sit" on the fitted functions. Maps of limiting factors show the variable most influencing the prediction for every grid cell. For further details see Elith *et al.* (2010 in press) and the most recent online tutorial (http://www.cs.princeton.edu/~schapire/maxent/). SDMs can provide useful information for exploring and predicting species distributions, and we are keen to see their continued development and use for learning about and conserving the world's biodiversity.

ACKNOWLEDGEMENTS

JE was supported by an Australian Research Council grant, FT0991640 and by an early consultancy that raised the question of how to explain MaxEnt to end-users (Jeff Tranter, Environmental Resources Information Network, Canberra, AUS). TH was partially supported by grant DMS-1007719 from the U.S. National Science Foundation. Simon Ferrier, John Baumgartner and Tord Snäll provided useful feedback on ideas and/or the manuscript. Robert Hijmans provided the method for taking samples proportional to area. Thanks to the three reviewers including Mark Robertson and Janet Franklin for generous and constructive comments and good ideas.

In press. Do not circulate.

Table 1 – Examples of published studies using MaxEnt, showing variation in purpose, scale and organism

Primary purpose	Scales	Organisms	Refs
Predict current distributions as input for conservation planning, risk assessments or IUCN listing, or new surveys	Andes Global	Humming-birds Stony corals seamounts	Tinoco *et al.* (2009) Tittensor *et al.* (2009)
Understand environmental correlates of species occurrences, groups of species, or other	Norway Portugal	Macrofungi European wildcat	Wollan *et al.* (2008) Monterroso *et al.* (2009)
Predict potential distributions for invasive species, or explore expanding distributions	New Zealand China	Ants Nematode	Ward (2007a) Wang *et al.* (2007)
Predict species richness or diversity	California Brazil	Amphibians and reptiles Myrtaceae 19 species	Graham and Hijmans (2006) Murray-Smith *et al.* (2009)
Predict current distributions for understanding morphological / genetic diversity ("phylogeography", "phyloclimatic studies"), endemism and evolutionary niche dynamics	Global Andes Madagascar	Seaweeds Birds Bats	Verbruggen *et al.* (2009) Young *et al.* (2009) Lamb *et al.* (2008)
Hindcast distributions to understand patterns of endemism, vicariance etc	NW Europe Brazilian coast	Pond snails Forests	Cordelier and Pfenninger (2009) Carnaval and Moritz (2008)
Forecast distributions to understand changes with climate change / land transformation; includes retrospective studies	Mediterr'n + surrounds Regional W. Australia Canada	Cyclamen Banksia Butterflies	Yesson and Culham (2006) Yates *et al.* (2010) Kharouba *et al.* (2009)
Test model performance against other methods	Patagonia Local region in California Regional to national	Insects Rare plants Many species	Tognelli *et al.* (2009) Williams *et al.* (2009) Elith *et al.* (2006)

In press. Do not circulate.

Table 2: Terminology used in this paper

Item / concept	**Definition**	**Notation**
Background	A sample of points from the landscape	
Entropy	A measure of disperdness. Previous papers[1] described the model as maximizing entropy in geographic space; this paper focuses on minimizing relative entropy in covariate space.	
Features	An expanded set of transformations of the original covariates	
Mask	A gridded layer of 1 / no data used to indicate areas to be included in background sampling (=1) and those to be excluded (=no data). To be included as a predictor. For projecting to the whole region, a grid called mask, but containing any values – say, 1 across the whole region of interest – should be supplied along with all other covariate grids.	
MESS map	Multivariate Environmental Similarity Surface –measures the similarity of any given point to a reference set of points, with respect to the chosen predictor variables. It reports the closeness of the point to the distribution of reference points, gives negative values for dissimilar points and maps these values across the whole prediction region (Elith *et al.* 2010 in press)	
Prevalence is not identifiable	Prevalence cannot be exactly determined from presence-only data in isolation, regardless of the sample size. This is a fundamental limitation of presence-only data.	
Probability density functions	Describe the relative likelihood of random variables over their range; can be univariate or multivariate.	
Regularization (tuning) parameters	Regularization refers to smoothing the model, making it more regular, so as to avoid fitting too complex a model. In MaxEnt the regularization parameters can be changed if required.	β in previous papers[1], λ in this paper
Sampling bias	Some areas in the landscape are sampled more intensively than others. Usually occurs in geographic space but could be environmentally based.	$s(z)$
Weights or coefficients	These are the parameters of the model that weight the contribution of each feature.	λ in previous papers[1], β in this paper

[1] Phillips *et al.* (2006), Phillips & Dudík (2008)

In press. Do not circulate.

<u>Table 3. Variable importance and evaluation statistics for case study 1</u>. Variable names and abbreviations for evaluation statistics are consistent with the text.

Model (background)	Variable importance						AUC (10fold CV but varying data sets)	AUC; COR (5fold CV on atlas data)
	RAIN DRYQ	RAIN	TEMP-WARMQ	TEMP-WETQ	ISO-THERM	SOL-PWHC		
1 (atlas)	57.9	30.7	7.9	0.4	1.1	2.0	0.92	0.96; 0.62
2 (southwest)	45.3	35.4	4.7	3.4	9.9	1.4	0.90	0.93; 0.52
3 (Australia)	19.7	17.7	5.3	54.0	3.0	0.3	0.99	0.91; 0.45

In press. Do not circulate.

FIGURES

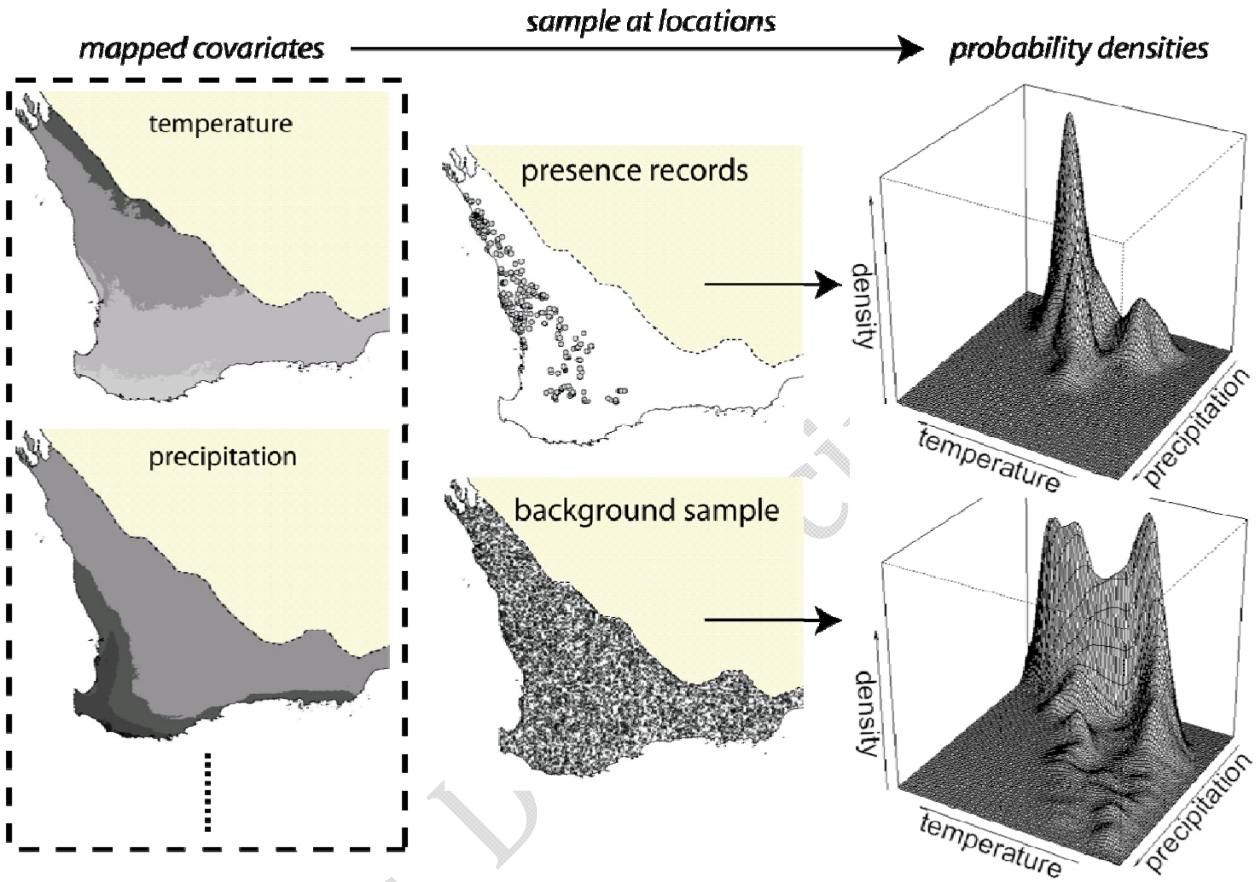


Figure 1 – A diagrammatic representation of the probability densities relevant to our statistical explanation, using data presented in case study 1. The maps on the left are two example mapped covariates (temperature and precipitation). In the centre are the locations of the presence and background samples. The density estimates on the right are not in geographic (map) space, but show the distributions of values in covariate space for the presence (top right) and background (bottom right) samples. These could represent the densities *f1(**z**)* and *f(**z**)* for a simple model with linear features.

In press. L

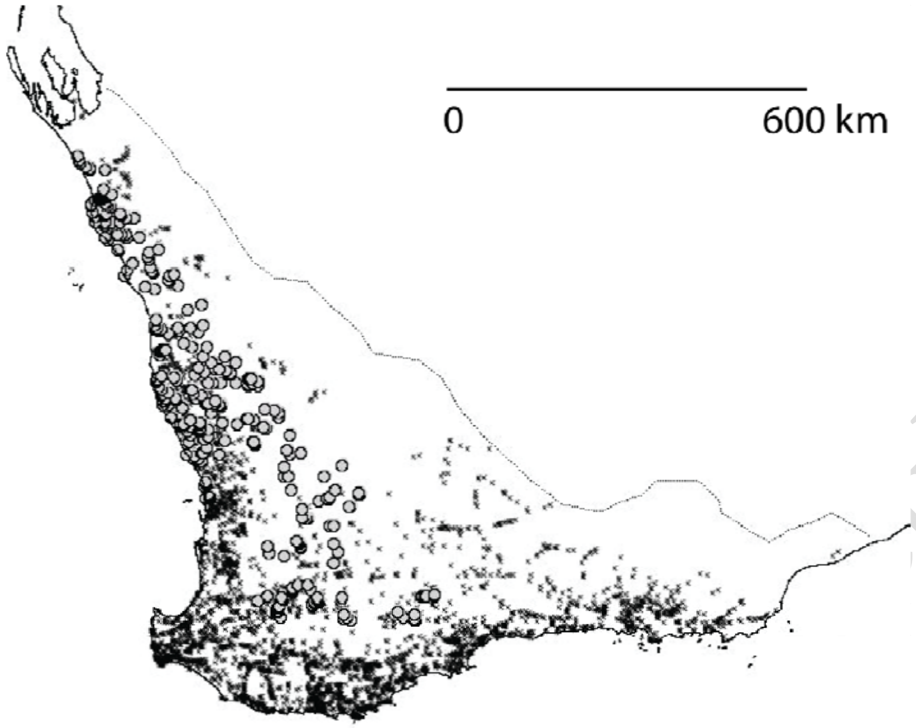


Figure 2: All banksia atlas sites (black) with occurrences of *B. prionotes* in grey circles.

In press. Do not circulate.

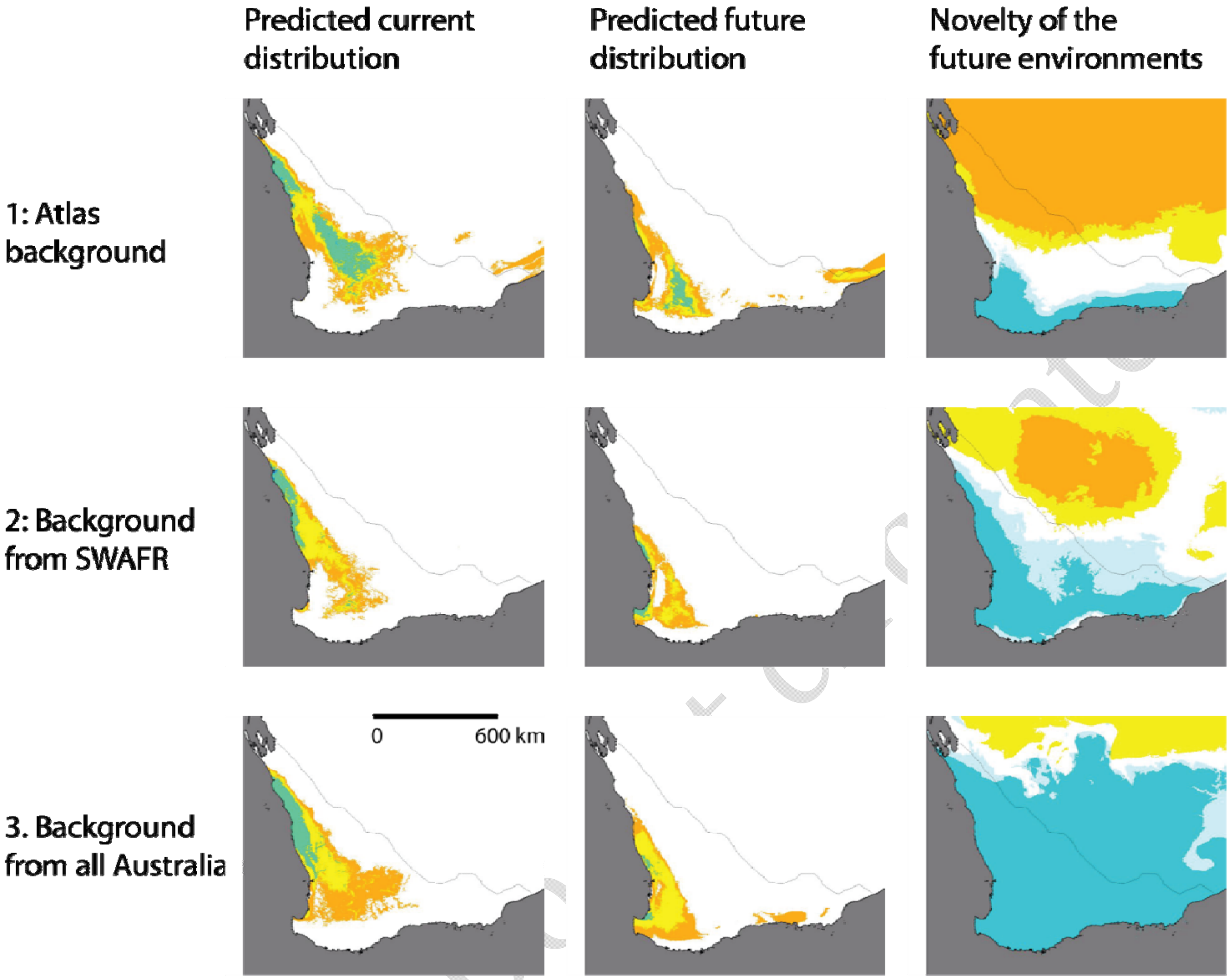


Figure 3. Model results for case study 1, showing for the three data sets (in rows): predicted current and future distributions, and extent of extrapolation compared with the training data. Predicted distributions are logistic outputs, from low values (white, 0 to 0.2) through orange, yellow, green to blue (0.8 to 1.0). For extrapolation maps, warm colours indicate extrapolation is occurring, with orange the most extreme. Grey indicates the ocean.

In press. Do not cite

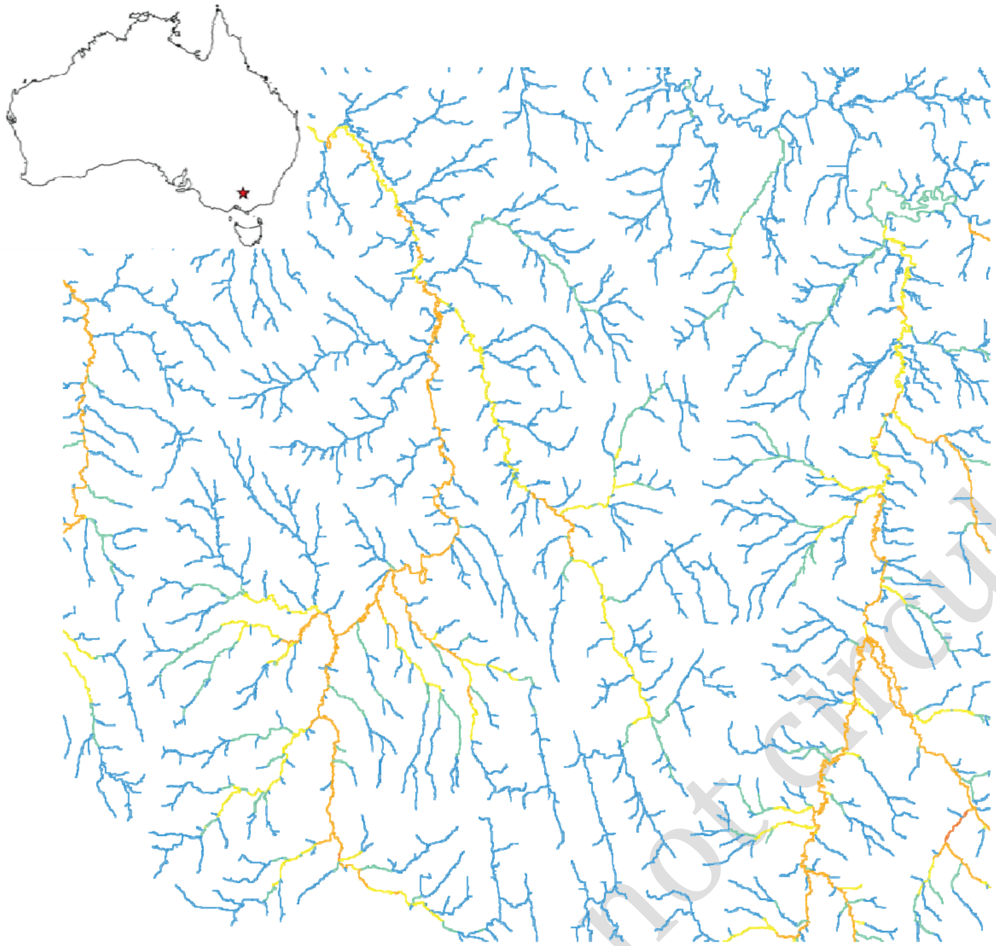
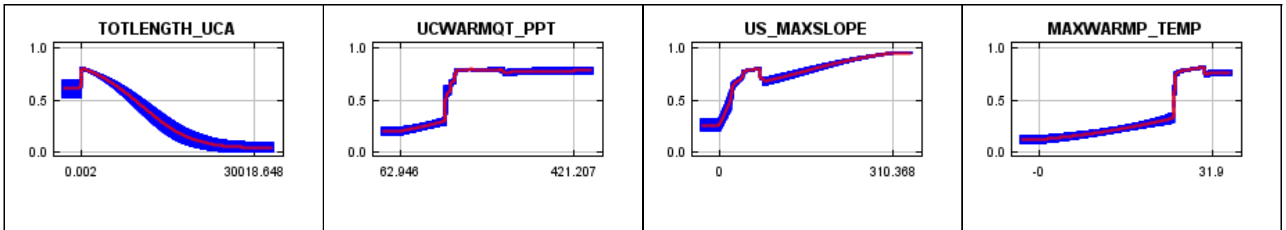


Figure 4 – Predicted distribution of *Gadopsis bispinosus,* showing logistic output predictions from MaxEnt. Legend: predictions in equal intervals from 0 to 1, from blue (low) through green – yellow –orange (high). Scale: east to west the rivers map spans 45km. The star on the inset shows location.

In press. Do not circulate.



45%	18%	8%	6%

Figure 5: Partial dependence plots showing the marginal response of *Gadopsis bispinosus* to the four most important variables (i.e., for constant values of the other variables), with variable importance below each graph. The y axes indicates logistic output.

In press. Do not circulate.

References

Akaike, H. (1974) A new look at statistical model identification. *IEEE Transactions on Automatic Control*, **AU-19**, 716-722.

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

Barry, S.C. & Elith, J. (2006) Error and uncertainty in habitat models. *Journal of Applied Ecology*, **43**, 413-423.

Carnaval, A.C. & Moritz, C. (2008) Historical climate modelling predicts patterns of current biodiversity in the Brazilian Atlantic forest. *Journal of Biogeography*, **35**, 1187-1201.

Chefaoui, R.M. & Lobo, J.M. (2007) Assessing the effects of pseudo-absences on predictive distribution model performance. *Ecological Modelling*, **210**, 478-486.

Cordellier, M. & Pfenninger, M. (2009) Inferring the past to predict the future: climate modelling predictions and phylogeography for the freshwater gastropod Radix balthica (Pulmonata, Basommatophora). *Molecular Ecology*, **18**, 534-544.

Della Pietra, S., Della Pietra, V. & Lafferty, J. (1997) Inducing features of random fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **19**, 1-13.

Dormann, C.F. (2007) Promising the future? Global change projections of species distributions. *Basic and Applied Ecology*, **8** 387-397.

Dudík, M. & Phillips, S.J. (2008) Generative and discriminative learning with unknown labeling bias. *Advances in Neural Information Processing Systems*, **21**

Dudík, M., Schapire, R.E. & Phillips, S.J. (2006) Correcting sample selection bias in maximum entropy density estimation. *Advances in Neural Information Processing Systems 18: Proceedings of the 2005 Conference* pp. 323-330. Cambridge, MA.

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

Elith, J. & Leathwick, J.R. (2009b) Conservation prioritization using species distribution models. Chapter 6 in: *Spatial conservation prioritization: quantitative methods and computational tools* (ed. by A. Moilanen, K.A. Wilson & H.P. Possingham). Oxford University Press.

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

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

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

Graham, C.H. & Hijmans, R.J. (2006) A comparison of methods for mapping species ranges and species richness. *Global Ecology & Biogeography*, **15**, 578.

Hastie, T., Tibshirani, R. & Friedman, J.H. (2009) *The elements of statistical learning: data mining, inference, and prediction, second edition*, 2nd edn. Springer-Verlag, New York.

Hirzel, A.H. & Le Lay, G. (2008) Habitat suitability modelling and niche theory. *Journal of Applied Ecology*, **45**, 1372-1381.

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.

Kharouba, H.M., Algar, A.C. & Kerr, J.T. (2009) Historically calibrated predictions of butterfly species' range shift using global change as a pseudo-experiment. *Ecology*, **90**, 2213-2222.

Lamb, J.M., Ralph, T.M.C., Goodman, S.M., Bogdanowicz, W., Fahr, J., Gajewska, M., Bates, P.J.J., Eger, J., Benda, P. & Taylor, P.J. (2008) Phylogeography and predicted distribution of African-Arabian and Malagasy populations of giant mastiff bats, Otomops spp. (Chiroptera : Molossidae). *Acta Chiropterologica*, **10**, 21-40.

In press. Do not circulate.

Leathwick, J.R. (1998) Are New Zealand's *Nothofagus* species in equilibrium with their environment? *Journal of Vegetation Science*, **9**, 719-732.

Lintermans, M. (2000) The Status of Fish in the Australian Capital Territory: a Review of Current Knowledge and Management Requirements. Technical Report No. 15, Environment ACT, Canberra.

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.

MacKenzie, D.I. (2005) Was it there? Dealing with imperfect detection for species presence/absence data. *Australia and New Zealand Journal of Statistics*, **47**, 65-74.

MacKenzie, D.I. & Royle, J.A. (2005) Designing efficient occupancy studies: general advice and tips on allocation of survey effort. *Journal of Applied Ecology*, **42**, 1105-1114.

Monterroso, P., Brito, J.C., Ferreras, P. & Alves, P.C. (2009) Spatial ecology of the European wildcat in a Mediterranean ecosystem: dealing with small radio-tracking datasets in species conservation. *Journal of Zoology*, **279**, 27-35.

Murray-Smith, C., Brummitt, N.A., Oliveira-Filho, A.T., Bachman, S., Moat, J., Lughadha, E.M.N. & Lucas, E.J. (2009) Plant Diversity Hotspots in the Atlantic Coastal Forests of Brazil. *Conservation Biology*, **23**, 151-163.

Newbold, T. (2010) Applications and limitations of museum data for conservation and ecology, with particular attention to species distribution models. *Progress in Physical Geography*, **34**, 3-22.

Pearce, J.L. & Boyce, M.S. (2006) Modelling distribution and abundance with presence-only data. *Journal of Applied Ecology*, **43**, 405-412.

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

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

Rabinowitz, D., Cairns, S. & Dillon, T. (1986) Seven forms of rarity and their frequency in the flora of the British Isles. *Conservation Biology: the Science of Scarcity and Diversity* (ed. by M.E. Soulé), pp. 182-204. Sinauer Associates, Sunderland, Massachusetts, USA.

Real, R., Barbosa, A.M. & Vargas, J.M. (2006) Obtaining environmental favourability functions from logistic regression. *Environmental and Ecological Statistics*, **13**, 237-245.

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.

Soberón, J. & Nakamura, M. (2009) Niches and distributional areas: Concepts, methods, and assumptions. *Proceedings of the National Academy of Sciences*, **106**, 19644-19650.

Soberón, J.M. & Peterson, A.T. (2005) Interpretation of models of fundamental ecological niches and species' distributional areas. *Biodiversity Informatics*, **2**, 1-10.

Solomon, S., Qin, D., Manning, M., Chen, Z., Marquis, M., Averyt, K.D., Tignor, M. & Miller, H.L. (eds) (2007) *Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*.

Svenning, J.C. & Skov, F. (2004) Limited filling of the potential range in European tree species. *Ecology Letters*, **7**, 565-573.

Taylor, A. & Hopper, S.D. (1988) The Banksia Atlas. AGPS, Canberra, ACT.

Tibshirani, R. (1996) Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B*, **58**, 267-288.

Tinoco, B.A., Astudillo, P.X., Latta, S.C. & Graham, C.H. (2009) Distribution, ecology and conservation of an endangered Andean hummingbird: the Violet-throated Metaltail (Metallura baroni). *Bird Conservation International*, **19**, 63-76.

In press. Do not circulate.

Tittensor, D.P., Baco, A.R., Brewin, P.E., Clark, M.R., Consalvey, M., Hall-Spencer, J., Rowden, A.A., Schlacher, T., Stocks, K.I. & Rogers, A.D. (2009) Predicting global habitat suitability for stony corals on seamounts. *Journal of Biogeography*, **36**, 1111-1128.

Tognelli, M.F., Roig-Junent, S.A., Marvaldi, A.E., Flores, G.E. & Lobo, J.M. (2009) An evaluation of methods for modelling distribution of Patagonian insects. *Revista Chilena De Historia Natural*, **82**, 347-360.

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.

Verbruggen, H., Tyberghein, L., Pauly, K., Vlaeminck, C., Van Nieuwenhuyze, K., Kooistra, W., Leliaert, F. & De Clerck, O. (2009) Macroecology meets macroevolution: evolutionary niche dynamics in the seaweed Halimeda. *Global Ecology and Biogeography*, **18**, 393-405.

Ward, D. (2007a) Modelling the potential geographic distribution of invasive ant species in New Zealand. *Biological Invasions*, **9**, 723-735.

Ward, G. (2007b) *Statistics in ecological modeling; presence-only data and boosted mars*. Stanford University, Palo Alto.

Williams, J.N., Seo, C.W., Thorne, J., Nelson, J.K., Erwin, S., O'Brien, J.M. & Schwartz, M.W. (2009) Using species distribution models to predict new occurrences for rare plants. *Diversity and Distributions*, **15**, 565-576.

Wintle, B.A., McCarthy, M.A., Parris, K.M. & Burgman, M.A. (2004) Precision and bias of methods for estimating point survey detection probabilities. *Ecological Applications*, **14**, 703-712.

Wollan, A.K., Bakkestuen, V., Kauserud, H., Gulden, G. & Halvorsen, R. (2008) Modelling and predicting fungal distribution patterns using herbarium data. *Journal of Biogeography*, **35**, 2298-2310.

Yates, C., McNeill, A., Elith, J. & Midgley, G. (2010) Assessing the impacts of climate change and land transformation on Banksia in the South West Australian Floristic Region. *Diversity and Distributions*, **16**, 187-201.

Yesson, C. & Culham, A. (2006) A phyloclimatic study of Cyclamen. *BMC Evolutionary Biology*, **6,** 72-95

Young, B.F., Franke, I., Hernandez, P.A., Herzog, S.K., Paniagua, L., Tovar, C. & Valqui, T. (2009) Using spatial models to predict areas of endemism and gaps in the protection of Andean slope birds. *Auk*, **126**, 554-565.

YunSheng, W., BingYan, X., FangHao, W., QiMing, X. & LiangYing, D. (2007) The potential geographic distribution of Radopholus similis in China. *Agricultural Sciences in China*, **6**, 1444-1449.

Zadrozny, B. (2004) Learning and evaluating classifiers under sample selection bias. p114 in *Proceedings of the Twenty-First International Conference on Machine Learning. Association of Machine Learning, New York, New York, USA.*

In press. Do not circulate.