**RESEARCH REVIEW**

# Is my species distribution model fit for purpose? Matching data and models to applications

Gurutzeta Guillera-Arroita[1]*, José J. Lahoz-Monfort[1], Jane Elith[1], Ascelin Gordon[2], Heini Kujala[1], Pia E. Lentini[1], Michael A. McCarthy[1], Reid Tingley[1] and Brendan A. Wintle[1]

[1]*School of Botany, The University of Melbourne, Parkville 3010, Vic., Australia,* [2]*School of Global, Urban and Social Studies, RMIT University, Melbourne 3001, Vic., Australia*

## ABSTRACT

Species distribution models (SDMs) are used to inform a range of ecological, biogeographical and conservation applications. However, users often underestimate the strong links between data type, model output and suitability for end-use. We synthesize current knowledge and provide a simple framework that summarizes how interactions between data type and the sampling process (i.e. imperfect detection and sampling bias) determine the quantity that is estimated by a SDM. We then draw upon the published literature and simulations to illustrate and evaluate the information needs of the most common ecological, biogeographical and conserva-tion applications of SDM outputs. We find that, while predictions of models fitted to the most commonly available observational data (presence records) suffice for some applications, others require estimates of occurrence probabilities, which are unattainable without reliable absence records. Our literature review and simula-tions reveal that, while converting continuous SDM outputs into categories of assumed presence or absence is common practice, it is seldom clearly justified by the application's objective and it usually degrades inference. Matching SDMs to the needs of particular applications is critical to avoid poor scientific inference and management outcomes. This paper aims to help modellers and users assess whether their intended SDM outputs are indeed fit for purpose.

*Correspondence: Gurutzeta Guillera-Arroita, School of Botany, University of Melbourne, Parkville 3011, Vic., Australia.
E-mail: gurutzeta.guillera@unimelb.edu.au*

## INTRODUCTION

Models play a critical role in conservation decision-making and ecological or biogeographical inference, but can lead to subop-timal conservation outcomes and misguided theory if the underlying data do not suit the intended application. Building models with unsuitable data can waste valuable resources and deliver outputs that do not solve the problem at hand.

Species distribution models (SDMs) have become a funda-mental tool in ecology, biogeography, biodiversity conservation and natural resource management (Guisan & Thuiller, 2005; Newbold, 2010; Franklin, 2013; Guisan *et al.*, 2013). SDMs typi-cally correlate the presence (or presence/absence) of species at multiple locations with relevant environmental covariates to estimate habitat preferences or predict distributions; these outputs are commonly used to inform ecological and biogeo-graphical theory as well as conservation decisions (e.g. Akcakaya *et al.*, 1995; Pearce & Lindenmayer, 1998; Ferrier *et al.*, 2002; Bekessy *et al.*, 2009; Keith *et al.*, 2014). A review of current use (see Appendix S1 in Supporting Information) reveals that SDMs are applied broadly, including for the management of threat-ened species (16% of papers), to control threatening processes (8%), to predict impacts of climate change (13%), to under-stand phylogeographic patterns (9%) and to manage landscapes (8%) and biological invasions (7%). A variety of diverse mod-elling algorithms are used (34 methods in 100 randomly selected papers).

SDMs can also be built with different types of species data, and these fundamentally affect the meaning of the quantity that is estimated. SDMs are therefore particularly prone to problems arising from a mismatch between data type and intended purpose. For example, some SDM applications require that the

probability of a species occurring at a site (e.g. Hauser & McCarthy, 2009) or the total area occupied (e.g. Keith *et al.*, 2008) be known. Other applications only require knowledge about relative site suitability so that the best (or worst) sites can be identified (e.g. Moilanen *et al.*, 2005). Although recent studies have mentioned such variation in required information (e.g. Lawson *et al.*, 2014), a comprehensive evaluation of how different sampling processes, data types and modelling approaches influence the utility of SDMs across applications is lacking. Here, we aim to raise awareness of this often disregarded, but critical, issue.

We distinguish three types of data that are typically used in correlative SDMs, which we label as *presence-background, presence–absence* and *occupancy-detection* (hereafter PB, PA and DET, respectively). With PB data (used in 53% of reviewed papers), only presence records (a non-exhaustive sample of all true presences) and environmental information are available. These types of data are often referred to as *presence-only* data, but we avoid this term, as discussed below. Many methods exist for modelling species distributions based on PB data, including the widely used Maxent (Phillips & Dudík, 2008; used in 41% of papers), various implementations of regression methods (e.g. Elith *et al.*, 2006) and the broader class of spatial point-process models (PPMs) which have received recent attention (Warton & Shepherd, 2010; Renner & Warton, 2013). For simplicity here we will discuss the data first with reference to sites (since this is how ecologists often think about data) and later clarify how this links with data structures for PPMs.

PB methods estimate habitat preferences by comparing the environmental characteristics at sites where the species has been recorded (P) with those throughout the region modelled, which we refer to as the 'background' (B). In contrast, PA data (47% of papers) provide information on whether a species was detected or not detected at a set of sampling sites. Logistic regression is commonly used to analyse PA data, but so are other statistical and machine learning techniques (e.g. see the review in Elith & Franklin, 2013). PA methods estimate the probability of observing a species at a site by comparing the environmental characteristics at sites where the species was detected with those at sites where it was not.

DET data (5% of papers) also consist of detection and non-detection records, but these are collected in such a way that the detection (or observation) process can be explicitly modelled within the SDM (Lahoz-Monfort *et al.*, 2014). For instance, sampling may involve collecting data from repeat visits to surveyed sites (MacKenzie *et al.*, 2002; Stauffer *et al.*, 2002; Tyre *et al.*, 2003) or recording times to detection during a single visit (Garrard *et al.*, 2008; Guillera-Arroita *et al.*, 2011). This provides information about the probability of detecting the species given that it is present at a site, and how that probability may vary from site to site or visit to visit (e.g. Wintle *et al.*, 2005b), allowing models to account for imperfect detection in the estimation of species occupancy probability.

As noted above, PB data are also sometimes referred to as presence-only data, but here we avoid this term to prevent confusion with true presence-only methods (PO; used in 11% of papers) (e.g. climatic envelopes like BIOCLIM; Busby, 1991; Booth *et al.*, 2014) which utilize information about sites where the species was detected without considering the environmental conditions in the rest of the landscape. We have chosen not to discuss PO methods here because they do not discriminate between environmental suitability and landscape characteristics (i.e. availability; Elith *et al.*, 2011). We note that using PO data limits SDMs by at least the same degree as working with PB data. Using PO methods does not solve issues related to working with PB methods; hence the issues we highlight regarding the limitations of SDMs fitted to PB data also apply to PO approaches.

For a given species, the probability of a site being recorded as a detection in a dataset is determined by three probabilities: the probability that the species occupies the site, the probability that the site is sampled and the probability that the species is detected given it is present at the site (Yackulic *et al.*, 2013). The first of these, the probability of occupancy, is the most common object of inference in a SDM. The second, the probability of a site being surveyed, can be affected by sampling bias (e.g. when surveys are only conducted near populated areas or reserves), whereas the third, detectability, may vary with environmental variables. As we explain later, our ability to deal with these three probabilities depends critically on the type of data that are used to build SDMs, and this in turn determines what SDMs can estimate. DET are the richest type of data in terms of information content, followed by PA and finally PB. However, this ordering is reversed when we consider data availability, and quite often PB are the only type of data that can be obtained (e.g. records from museum or herbarium specimens; Newbold, 2010). Approximately half of recently published papers involving SDMs rely on PB data (Appendix S1). Hence it is crucial to understand whether, given the available data, the output of a proposed SDM will be appropriate for the intended application.

The suitability of a SDM for a given application also relies critically on how the raw output from the SDM is used. Our review (Appendix S1) revealed that 54% of recent papers converted the continuous output produced by SDMs into binary (discretized) predictions of 'presence/absence' or 'habitat/non-habitat'. There appears to be a belief that this step is required by many ecological, biogeographical or conservation applications (e.g. Jiménez-Valverde & Lobo, 2007; Li & Guo, 2013; Liu *et al.*, 2013). However, attention is rarely paid to whether a binary output is indeed required, or whether this may lead to an unnecessary loss of information and hence be detrimental in the context of the intended application (but see Merow *et al.*, 2013; Calabrese *et al.*, 2014; Lawson *et al.*, 2014). Given the prevalence of binary conversion of predictions and the paucity of critical appraisal of this issue in the literature, we explore the need for and potential impacts of discretizing SDM outputs in the context of specific ecological, biogeographical and conservation applications.

In summary, despite being an essential consideration in modelling, confusion remains about the limitations of different types of data for building SDMs and about the uses various outputs can justifiably support. This confusion results in insufficient evaluation of whether SDM outputs are appropriate for

proposed applications; a failing we aim to redress here. Our specific aims are twofold. First, we clarify and synthesize current knowledge regarding the inherent properties and limitations of correlative SDMs with respect to the type of data that are used to build them. To do this, we provide a framework that clarifies how the interaction between the sampling process and the type of species data determines the quantity that is ultimately estimated from a SDM. Our second aim is to evaluate the information needs of a range of ecological, biogeographical and conservation applications that typically use SDMs as inputs, including whether the conversion of continuous SDM outputs into binary predictions is necessary or appropriate. Our paper provides modellers and managers with a means to assess whether the species data that they have available or are planning to collect are appropriate for their intended use. Since the focus of this paper is to identify how types of species data limit the uses of SDMs, we assume that other relevant aspects of the model-building process are well covered. These include having access to a dataset that is sufficiently large and appropriately collected, that all the relevant environmental covariates are considered and that modelling techniques and associated software programs are correctly used. Apart from particular biases induced by certain data types, we do not consider these issues further.

## THE INTERPLAY BETWEEN DATA TYPES AND BIASES IN SDMS

There are three fundamental issues to consider with respect to the type of species data used to build a SDM: (1) whether prevalence (the proportion of sites occupied) can be estimated; (2) the impact of imperfect detection; and (3) the impact of environmental sampling bias (Fig. 1a). These factors, discussed in more detail below, dictate the types of application that a SDM can reliably support given that they determine whether the resulting model estimates: (1) the probability of species occurrence, (2) a relative likelihood of species occurrence (which is proportional to the actual probability), (3) a correct ranking of the sites in terms of occurrence probability; or (4) whether it provides a biased estimation of species occurrence in which not even the correct ranking of sites in terms of true occurrence probability is achieved.

In ecology, biogeography and conservation most users want to know the probability that a species *occupies* a given environment, which when mapped represents an estimate of its geographic *distribution*. This might be substantially different from where the species is most likely to be *observed*, which is what the least informative models estimate (Fig. 1a, column 1). Column 2 (yellow in Fig. 1a) represents models that correctly rank the suitability of locations for the species, although the predicted suitability is not proportional to the actual probability of occurrence (Fig. 1b). Models in columns 3 and 4 naturally provide such a ranking as well, but importantly also capture the shape of the environmental relationships explaining species occurrence probability, either directly or with a constant scaling.

SDMs that achieve any of the three top information content categories (columns 2–4) have good discrimination abilities (i.e. they distinguish between occupied and unoccupied sites better than random; Pearce & Ferrier, 2000; Phillips *et al.*, 2006). However, only those that estimate probabilities can achieve good calibration (i.e. agreement between predicted probabilities and observed proportions of sites occupied; Pearce & Ferrier, 2000). Some of these categories of information content can be attained with different data types, given particular conditions of sampling bias and imperfect detection (Fig. 1), as discussed below. What ultimately matters from the point of view of applications is whether a SDM estimates the quantity required for a given use, be that a probability of occurrence, a relative likelihood or a correct ranking of sites. Our diagram in Fig. 1(a) distinguishes between the information requirements of a particular application and the type of species data (PB, PA or DET) used to construct the SDM. Such decoupling allows evaluation of the requirements of different applications without repeated reference to data types or specific issues like bias or imperfect detection: the key is to identify the minimum information required for a given use.

Although for simplicity we present our categorization and discussion only from the point of view of estimating *probabilities* of species occurrence at sites, we note that the issues we cover apply equally well to cases where the focus of estimation is the *intensity* of points (observations) in the landscape. This is the quantity estimated by PPMs (Warton & Shepherd, 2010; Renner & Warton, 2013), which are not intrinsically based on dividing space into discrete sites and allow probabilities to be derived at any spatial resolution if needed. Regardless of whether intensities or probabilities are obtained, PPMs for PB data present the same limitations as other PB methods in terms of their ability to estimate prevalence (and hence absolute intensities/probabilities of occurrence) and are similarly affected by issues of sampling bias and imperfect detection.

### Estimation of prevalence

PB methods do not estimate actual probabilities, but relative likelihoods of species occurrence (or observation). From a set of presence records alone (PB data) it is not possible to distinguish whether a species is rare and well surveyed or common but under-surveyed. Distinguishing between these two possibilities requires PA or DET data. Despite being openly acknowledged by developers of PB methods (e.g. Ferrier & Watson, 1997; Elith *et al.*, 2006) this limitation is not always fully appreciated. Confusion remains among users, as indicated for instance by frequent assumptions that the output of Maxent is a probability of occurrence (a fact reported, e.g., by Yackulic *et al.*, 2013 and Guillera-Arroita *et al.*, 2014c).

Some authors note that, under certain strong conditions, prevalence can be estimated (identified) from PB data (Lancaster & Imbens, 1996; Lele & Keim, 2006; Royle *et al.*, 2012). However, these conditions involve very specific restrictions about how species occurrence varies with explanatory
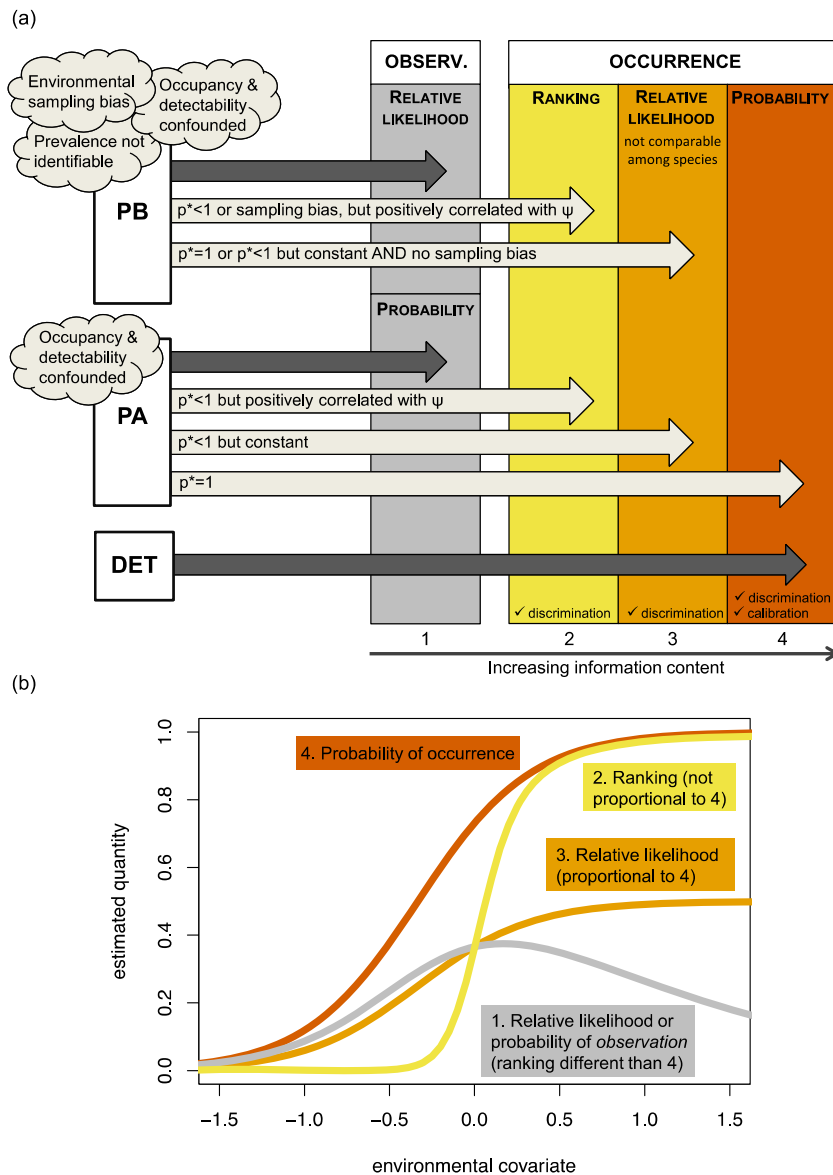
**Figure 1** (a) Synthesis of how the type of survey data interacts with sampling bias and imperfect detection to determine what a correlative species distribution model can estimate. Dark arrows denote the *default* level of information that can be achieved with each type of survey data (PA, presence–absence; PB, presence-background; DET, occupancy-detection). Light arrows indicate under which conditions higher levels of information can be achieved from those data types. $\psi$ denotes the probability of species occurrence at a site, and $p^*$ the probability of detecting the species at a site where present (given all the survey effort applied per site). Column 1 refers to the probability/likelihood of *observing* the species. Columns 2–4 refer to information about species *occurrence* (i.e. distribution). What we represent for convenience as distinct quantities (columns) correspond in practice to a gradation (e.g. a bias may be negligibly small). (b) Graphical example of the different levels of information represented by each column in panel (a), assuming a single environmental covariate for simplicity. Taking the probability of occurrence in (4) as a reference, the relative likelihood of occurrence (3) is proportional to it, while (2) is not proportional but would rank sites in the same order. The relative likelihood or probability of *observation* (1) may not even provide a good ranking in terms of occurrence. Although for simplicity we present this figure from the point of view of the estimation of *probabilities* of species occurrence at sites, the general ideas also apply when the focus of estimation is the *intensity* of points (observations) in the landscape (i.e. from point-process models). Only the terminology is slightly different; for instance, instead of probability or relative likelihood of occurrence, we would refer to absolute/relative intensity of species occurrence.

variables. Whether prevalence can be identified largely relies on the *true* relationship conforming *exactly* to an assumed parametric form, which is highly unlikely to occur in reality (Hastie & Fithian, 2013; Phillips & Elith, 2013). Mild departures from the assumed functional form can cause profound errors (Fig. S2.1 in Appendix S2), and even if the assumption is perfectly met estimates remain very imprecise (Merow & Silander, 2014). Hence, for practical purposes it can be argued that

prevalence is not obtainable from PB data (Hastie & Fithian, 2013; Lele *et al*., 2013; Phillips & Elith, 2013).

## Imperfect detection

It is widely recognized that surveys often fail to detect the species present at a site (Yoccoz *et al*., 2001; Kéry, 2002), even for sessile species such as plants (Garrard *et al*., 2008; Chen *et al*., 2013). If detection is imperfect and this is not accounted for, occupancy and detection processes are confounded and SDMs might estimate where the species is more likely to be observed rather than where it does occur (Gu & Swihart, 2004; Kéry, 2011; Lahoz-Monfort *et al*., 2014). Estimated distributions can be substantially different depending on whether detectability is accounted for or not (e.g. Kéry *et al*., 2010; Kéry *et al*., 2013).

The impact of ignoring imperfect detection is greatest when detectability is a function of environmental variables, because this distorts our understanding of even the general shape of the environmental relationships explaining the distribution of the species (see figures 1 and 2 in Lahoz-Monfort *et al*. 2014). Despite some past confusion in the literature, it is now clearly established that imperfect detection will affect both PA and PB methods in this situation (Elith *et al*., 2011; Dorazio, 2012; Yackulic *et al*., 2013; Lahoz-Monfort *et al*., 2014), i.e. using only presence records does not circumvent the problem of imperfect detection. In the case of PA data, imperfect detection can lead to false absence records, while the effect on PB methods stems from the fact that presence records do not provide a random sample of locations in which the species is actually present. Although not always recognized, imperfect detection is also an issue when detection and occurrence depend on different and uncorrelated covariates. Bias can still be induced in the estimation of the species distribution as the modelling process can mistakenly identify detection covariates as relevant predictors of occurrence probability (Lahoz-Monfort *et al*., 2014).

Acknowledging that detection can be imperfect implies accepting that PA and PB methods might provide a biased estimation of species distributions (column 1 in Fig. 1a). It is only under particular conditions that a more meaningful output can be obtained. If detectability is constant across sites (and there is no sampling bias), PB methods can reliably estimate the relative likelihood of species occurrence despite imperfect detection (column 3). Constant detectability might occur if the same observer carried out all surveys under similar conditions and in comparable habitats. The relative likelihood of species occurrence is the maximum amount of information that a PB method can yield, even when detection is perfect. Likewise, when PA data are available and detection is imperfect but constant, SDMs can estimate only a relative likelihood of species occurrence (column 3). However, with PA data there is also the potential to obtain an unbiased estimation of occurrence probabilities *when detection is perfect* (column 4). When detectability depends on environmental covariates in the same way that species occurrence does (i.e. occupancy and detectability are positively correlated), the general shape of the estimated environmental relationship may be affected but the correct ranking of the suitability of sites is maintained (column 2).

When appropriate data are available (i.e. DET data), models that estimate species occupancy while accounting for detectability separate detection from occurrence (MacKenzie *et al*., 2002; Stauffer *et al*., 2002; Tyre *et al*., 2003; Guillera-Arroita *et al*., 2011, 2014b), enabling unbiased estimation of occupancy probabilities even if detectability varies (provided the detection process is well characterized in the model). The lower the detectability of a species, the more data are required to precisely estimate its distribution (MacKenzie & Royle, 2005; Guillera-Arroita & Lahoz-Monfort, 2012).

## Sampling bias

PB methods work under the assumption that sampling is unbiased (Phillips *et al*., 2009; Elith *et al*., 2011). However, PB data seldom arise from random sampling and are often collected in an opportunistic and spatially biased manner. Common examples include datasets derived from herbaria and museum specimens (Newbold, 2010). In such datasets, sampling is often biased towards accessible locations near roads or towns. The problem for modelling species distributions is not the spatial bias in itself but a bias in how the available environmental conditions are sampled. For example, areas close to cities may tend to be at lower elevations and have higher soil fertility. In PB SDMs, sampling bias causes biased estimation of environmental relationships, with suitability being over-estimated for environments that have been sampled more intensively and underestimated for those surveyed less frequently. Some methods have been proposed to mitigate the problem of sampling bias in PB SDMs. However, none of these can completely resolve the problem as they rely on indirectly inferring the sampling process (e.g. by considering records of other species as a proxy; Phillips *et al*., 2009) or on the ability to model the sampling process separately, which requires having good predictors of sampling effort and independence between those and the predictors of species occurrence (Chakraborty *et al*., 2011; Fithian & Hastie, 2013; Warton *et al*., 2013; Fithian *et al*., 2014).

For PB SDMs, sampling bias is analogous to imperfect detection when detectability depends on environmental covariates. In both cases the species is less likely to be detected in certain environments, either because it is truly more difficult to detect when those sites are surveyed ('true' detectability issue) or because those sites are less likely to be visited (environmental sampling bias). It is only when there is no sampling bias that PB methods are able to estimate the relative likelihood of species occupancy (column 3 in Fig. 1a); when there is environmental sampling bias but it is positively correlated with occupancy, site ranking is nevertheless maintained (column 2). Sampling bias is not only an issue when sampling effort is correlated with covariates determining species occurrence. Sampling effort that varies with factors that are not correlated with species occurrence can still bias the estimation of the relative likelihood of occurrence, as those factors can be incorrectly identified as relevant predictors in the model.

In contrast to PB models, the effect of sampling bias on PA or DET SDMs is very different and has less critical consequences, as it does not introduce bias in the estimation (Phillips *et al.*, 2009). Instead, sampling bias implies reduced estimation precision for sites in the environmental space that are less intensively sampled.

## SDM OUTPUTS IN ECOLOGICAL, BIOGEOGRAPHICAL AND CONSERVATION APPLICATIONS

SDM outputs need to match the objectives of the applications that they aim to inform. In Appendix S3 we review extensively the information required from SDMs for ecological and conservation applications, organized into four tables that represent the following domains: (1) management of invasive species, (2) management of threatened species, (3) spatial planning, and (4) ecological and biogeographical inference. For each application, we specify how SDM outputs are used, describe the potential consequences of using a biased estimation of occupancy and discuss whether a relative likelihood or correct ranking is a sufficient input. Readers can search for their application in the first column of the tables, find the minimum level of information content required and use Fig. 1 to determine the circumstances in which each data type will produce a suitable SDM output for that application. We also assess whether particular applications actually require a binary input (generated by applying a threshold to the SDM output) or whether working with the continuous SDM output is more appropriate.

In the following subsections, we explore five simulation-based case studies that characterize common uses of SDMs in ecology, biogeography and conservation (summarized in Table 1), and use these to illustrate key points regarding SDM data needs. For each case study we assess the consequences of using SDMs that estimate probabilities or relative likelihoods, considering both direct use of continuous SDM outputs and use after conversion to binary maps. For the latter, we tested a set of commonly used thresholds (Appendix S1; Liu *et al.*, 2005): (1) the minimum estimate corresponding to a training presence ('MTP'); (2) the 10th percentile estimate for training presences ('10TP'); (3) the threshold that resulted in equal sensitivity and specificity ('SS = SP'); and (4) the threshold that maximized the sum of sensitivity and specificity ('max (SS + SP)'). For PB data, specificity cannot be defined in the usual way given that absence records are not available, so we followed the practice in Maxent that uses the background data instead, and is therefore related to predicted area.

### Case study 1: wildlife monitoring

The area (or number of sites) occupied by a species (sometimes abbreviated AOO for 'area of occupancy') is the state variable of interest in many large-scale monitoring programmes which aim to track variation in AOO over time (MacKenzie *et al.*, 2006, pp. 41–44). Both AOO and changes in AOO are important elements

of IUCN Red List assessments (criteria A, B2 and D2; IUCN, 2012). AOO can be estimated by summing the site occupancy probabilities obtained with a SDM over the landscape of interest. It is clear that this cannot be done based on relative likelihoods (since by definition AOO is the prevalence multiplied by the total area). Here we use a simulation study to show the extent to which it is possible to track temporal changes in AOO by *fitting* models that provide lower levels of information than occupancy probabilities. We also explore whether converting the output of SDMs into binary maps assists in estimating AOO.

We simulated a temporal decline in the AOO of a species (representing the 'true' change) and the sampling of large PA and PB datasets for three time steps, which, when analysed led to estimated probabilities and relative likelihoods of occupancy, respectively (details in Appendix S4). To estimate the change in AOO, SDM estimates were summed over the landscape, either directly as a continuous output or after applying the four thresholds listed above to convert them to binary maps.

The simulations demonstrate that relative likelihoods produced by a PB-based SDM are not comparable across time: prevalence is proportional to the declining AOO but cannot be estimated, and thus the resulting estimates are clearly unable to track declines in AOO (Fig. 2). PB data cannot determine whether an apparent decrease in the number of detections for a species through time reflects an actual reduction in AOO or declining survey effort. Binary conversion of the SDM output prior to calculation of AOO does not solve the problem of working with relative likelihoods of occurrence, because a binary categorization does not fix the fact that prevalence cannot be estimated without absence data. Furthermore, binary conversion is detrimental compared with using the actual probabilities of occurrence when available. This is because a binary categorization represents a coarse interpretation of species occurrence probabilities and reduces the information content compared with using the full range of values provided by the SDM. In our simulations, detection was assumed to be perfect, hence the PA-based SDM produced reliable estimates of probabilities of occurrence. However, imperfect detection can cause PA methods to miss trends or detect spurious ones. An exception to that rule occurs if detectability is constant across space and time (a relatively strong assumption). Then, PA methods estimate relative likelihoods of occupancy that track changes in AOO.

### Case study 2: invasive species priority lists

Biosecurity resources are limited, so it is common for government agencies to prioritize exotic species for management intervention. The potential distribution of an exotic species is a key indicator of its ability to spread, and is therefore frequently considered in this prioritization process. In Australia, for example, weeds of national significance are in part determined by examining the potential distribution of a range of candidate species under current and future climates (Lizzio *et al.*, 2009). This listing process influences the allocation of potentially millions of dollars worth of biosecurity resources.

**Table 1** Adequacy of SDM data type and impact of binary conversion for the set of applications chosen as case studies. For a detailed assessment of a wider range of applications see Appendix S3.

| Application | Domain | S/M | Consequences of biased estimation | Is relative likelihood/ranking ok? | Has binary classification been used? | Consequences of using a binary classification |
|---|---|---|---|---|---|---|
| Calculate AOO or compare temporal changes in AOO, to monitor or assess status of a species (case study 1) | Management of invasive or threatened species | S | Species status may be incorrectly assessed due to biased estimation of AOO or its temporal trends (real trend missed or spurious trend detected) | Relative likelihoods and rankings not able to estimate AOO, and not sufficient input to track temporal trends (as not comparable across time; except for PA data when detectability constant in space and time) | Yes | Degrades AOO estimation, even when probabilities are estimated; It does not solve the inadequacy of relative likelihoods or rankings |
| Compare potential AOO for a list of species to prioritize management resources across species (case study 2) | Management of invasive species | M | Management effort may be allocated toward species that actually have low potential for spread and vice versa | Relative likelihoods and rankings are not sufficient inputs because they are not comparable across species | Yes | Degrades AOO estimation, even when probabilities are estimated; It does not solve the inadequacy of relative likelihoods or rankings |
| Determine optimal surveillance to minimize monitoring plus management costs (case study 3) | Management of invasive species | S | Overall costs will increase, as the level of surveillance guiding early management is not optimal | Computing expected costs requires probabilities; can result in suboptimal solution otherwise | No | Not valid. Method based on acknowledging uncertainty in species presence to evaluate trade-offs in expected costs |
| Estimate species richness (by stacking single-species SDMs) to identify biodiversity hotspots for prioritization, or for ecological or biogeographical inference (case study 4) | Spatial planning; ecological/ biogeographic inference | M | Areas where the likelihood of *observing* species is highest (rather than the best habitats) may be incorrectly identified as biodiversity hotspots | Relative likelihood not sufficient to estimate absolute richness. Relative values may lead to robust site rankings in terms of richness in *some* scenarios when comparing sites within the same modelled region | Yes | Degrades richness estimation, even when probabilities are estimated, and does not solve the inadequacy of relative likelihoods or rankings |
| Spatial prioritization: SDM as input layer for evaluating the constraint in a 'minimum set' approach, e.g. Marxan (case study 5) | Spatial planning | M | Can lead to an underperforming prioritization, with some targets not met and/or solution involving greater cost than needed. Prioritization will favour sites where the likelihood of *observing* species is highest, rather than the best sites | Relative likelihood is fine if targets are set in terms of the percentage of distribution to be retained, but not if set as absolute areas. A ranking that does not capture the correct shape can lead to an underperforming prioritization | Yes | Resulting information loss may lead to an underperforming prioritization. Effect more pronounced when targets are set as absolute areas (as opposed to percentage of distributions) |

AOO, area of occupancy; SDM, species distribution model; S/M, single/multiple species (this refers to whether the calculations involve SDMs from several species, e.g. by combining or prioritizing them; we do not consider as multiple species those applications where, although several species may be considered, calculations are independent for each of them).
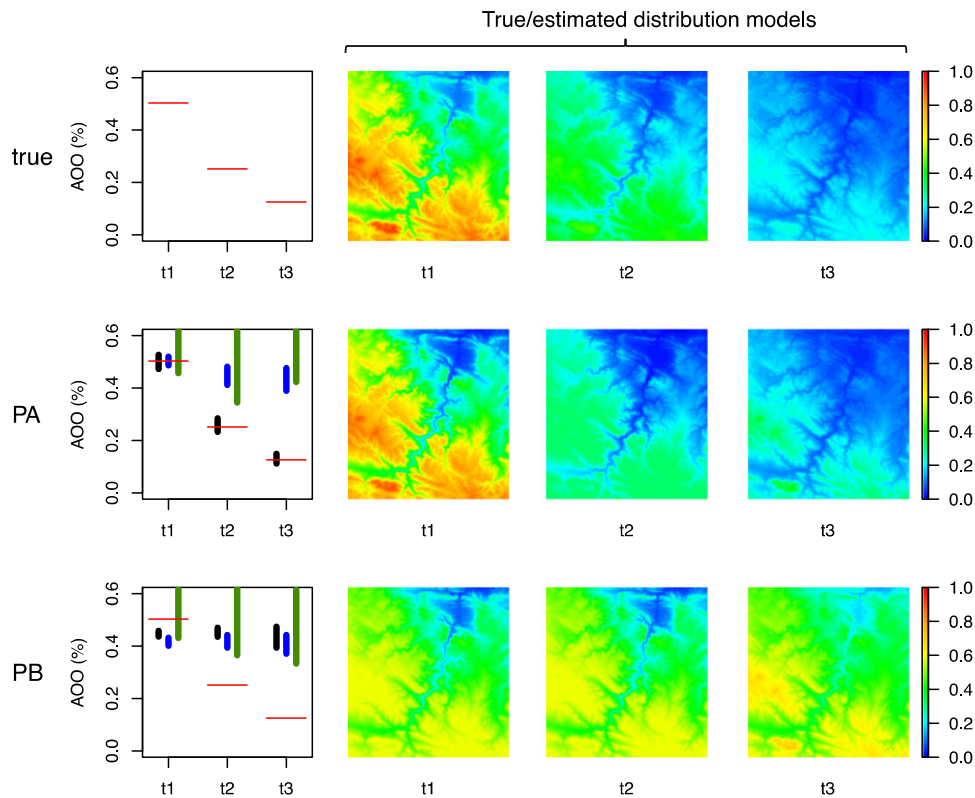
**Figure 2** Wildlife monitoring case study. The rows show the true and estimated area of occupancy (AOO, %) for a virtual species as a function of time: row 1, true AOO; row 2, AOO derived from a presence–absence (PA) model; row 3, AOO derived from a presence-background (PB) model. In column 1 the red horizontal lines report true AOO, and each bar summarizes the result of 20 simulations (10th to 90th percentiles), with colours representing how the species distribution model (SDM) output was used (black, no threshold; blue, 'SS = SP' threshold; green, 'max (SS + SP)' threshold; see threshold definitions in the main text and similar results for other thresholds in Appendix S4). Columns 2–4 show the estimated distributions for one of the simulations for the three time steps (t1, t2 and t3), as well as the 'true' distribution model. PB SDMs are not able to track the species decline. Using thresholds does not solve the problem, and is detrimental for estimation based on PA data.

Here we show that estimates of the relative likelihood of occupancy are not suitable for prioritizing species according to their potential AOO. For this, we simulated a set of 25 species, sampled their distributions and built SDMs based on PA and PB datasets (details in Appendix S4). For each species, we computed the sum of the SDM estimates across the whole landscape, using the continuous output as well as the binary outputs obtained after applying the thresholds.

In statistical terms, the sum of estimated occupancy probabilities across the region is the expected value of the AOO, and hence it is a good estimator for AOO, leading to the correct prioritization of species (PA method; Fig. 3). However, if the output of the SDM is a relative likelihood of species occupancy, the AOO cannot be estimated, as we saw in case study 1. Crucially, the quantities obtained are not comparable across species, and hence species cannot be prioritized based on these data (PB method; Fig. 3). Applying a binary conversion to the SDM output does not solve the problem for the same reasons as described above. In this case study, detection was assumed to be perfect, hence the PA-based SDMs produced estimates of actual probabilities. Imperfect detection can also lead to incorrect

prioritization when PA data are used, because even if detectability is constant across space it is likely to be species specific.

### Case study 3: optimal surveillance of invasive species

SDMs can be used to inform invasive species surveillance. For instance, Hauser & McCarthy (2009) proposed a spatial detection and treatment model, and identified the level of surveillance that minimizes total expected costs while taking into account ease of detection and control as well as the probability of species occurrence at the site. This approach was subsequently used to design surveys for invasive orange hawkweed *Pilosella aurantiaca*, in south-eastern Australia (Herbert *et al.*, 2013). The method considers the following scenario: a site is surveyed and, if the species is detected, the site is managed for eradication. If not detected, the species is assumed absent and the site is not managed. However, the species could have been present and not observed, in which case the lack of early management may lead to a wider infestation that ultimately incurs higher management costs. There is thus a trade-off between the
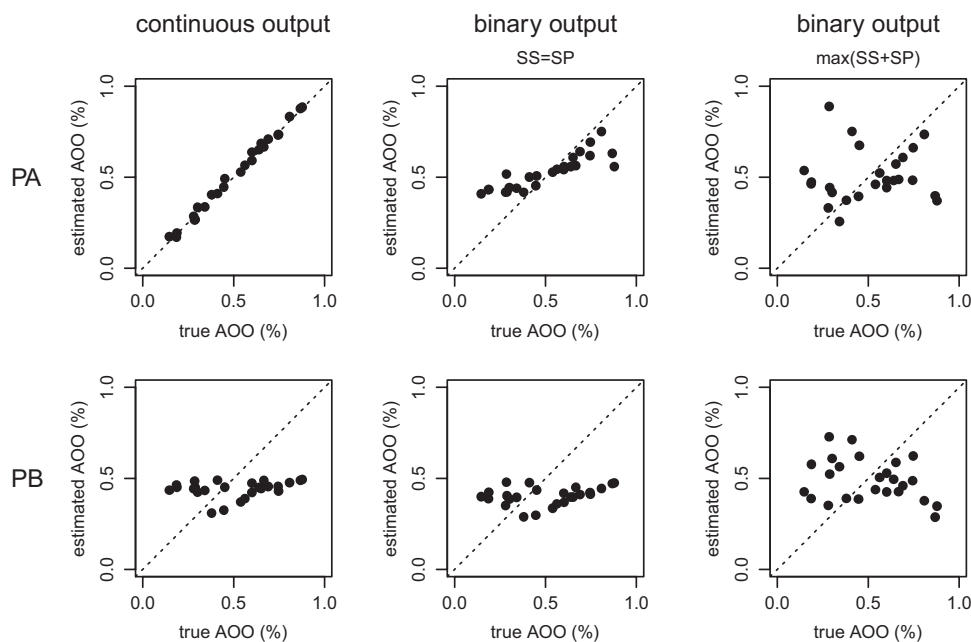
**Figure 3** Invasive species prioritization case study. Estimated area of occupancy (AOO, %) for 25 virtual species as a function of their true AOO, based on the output of presence–absence (PA) and presence-background (PB) species distribution models (SDMs; first and second row, respectively). In the first column, the SDM output is used without a threshold. In the second and third columns the SDM output is first converted to a binary output using a threshold ('SS = SP' or 'max (SS + SP)'; see threshold definitions in the main text and similar results for other thresholds in Appendix S4). The PB SDM does not prioritize species correctly. Binary conversion did not solve the problem, and was detrimental for estimation based on PA data.

amount of resources spent on surveying and early management on infested sites and the resources ultimately spent in delayed management at sites where the species was present but undetected.

For this case study, we sampled the distribution of a simulated invasive species, built SDMs based on PA and PB datasets and then applied the approach by Hauser & McCarthy (2009) to determine the optimal level of surveillance without imposing budget restrictions, using the SDM outputs as estimates of species occurrence probabilities (details in Appendix S4). Only when the SDM values used for the optimization represent actual probabilities of occurrence (PA method) can the potential savings from surveying be fully realized (Fig. 4; but see Guillera-Arroita *et al.* (2014a) for cases where relative probabilities may suffice). In our simulated example, applying the truly optimal level of surveillance reduced costs by 26% compared with a situation without surveillance and early management. In contrast, if the level of surveillance was determined by wrongly interpreting the output of a PB method as probabilities, total costs increased by 22%.

### Case study 4: estimation of species richness

SDMs can be combined to model biodiversity at the community level following a 'predict first, assemble later' strategy (Ferrier & Guisan, 2006), whereby separate SDMs for each species are stacked using some form of aggregation. One example of this approach is where SDMs are summed site-by-

site to derive predictions of species richness (i.e. the number of species present at each site), which can then be used to explore further ecological or biogeographical questions or to identify biodiversity hotspots for conservation (e.g. Parviainen *et al.*, 2009).

The sum of species occupancy probabilities at a site is equal to the expected number of species present, and hence is a good estimator of species richness (Calabrese *et al.*, 2014). One would not expect richness to be reliably estimated if the individual SDMs only estimate relative likelihoods of species occurrence. Yet, a number of studies have applied this methodology using PB SDMs (e.g. Pineda & Lobo, 2009; Schmidt-Lebuhn *et al.*, 2012). There are also claims in the literature that when calculating species richness it is best to transform continuous SDM predictions into binary predictions prior to summing them (e.g. Newbold *et al.*, 2009; Pineda & Lobo, 2009), or even that such transformation is simply 'inevitable' (Trotta-Moreu & Lobo, 2010). We test whether these seemingly counterintuitive assertions are justified in the simulations described below.

We simulated the distribution of a set of 100 species, sampled them, fitted SDMs using PA and PB methods and then aggregated the resulting SDMs to estimate species richness (details in Appendix S4). Here we report the results of a scenario where half of the species depend on one covariate and tend to have low prevalence, while the second half depend on a different covariate and tend to have high prevalence (results for other scenarios are presented in Table S4.2 in Appendix S4).
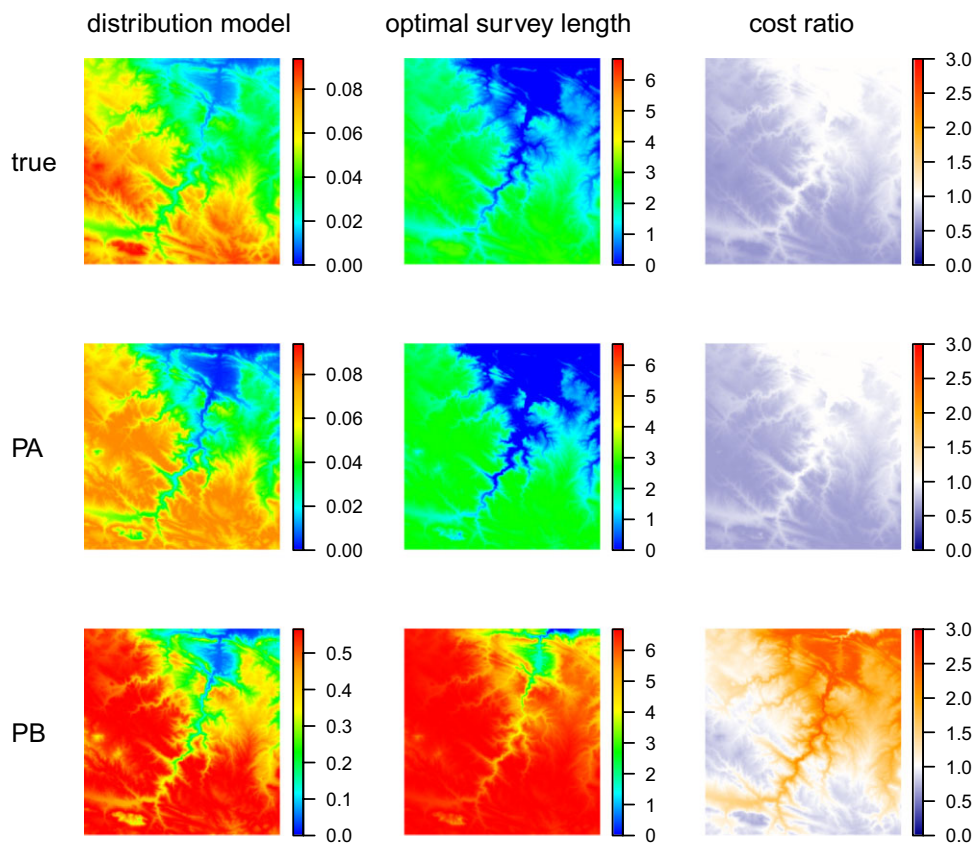
**Figure 4** Optimal surveillance of invasive species case study. Rows correspond to three scenarios of species occurrence information used to determine the optimal amount of surveillance: row 1, 'true' occurrence probabilities; row 2, species distribution model (SDM) obtained from presence-absence (PA) data; row 3, SDM obtained from presence-background (PB) data. The first column displays the SDM (true/estimated occurrence probability or estimated likelihood of occurrence; note the difference in scale). The second column displays the optimal survey length at each site, determined by taking the maps in the first row as true occurrence probabilities. The third column displays the ratio of total costs (survey + management) between the case with surveys as determined above and a case with no surveys. Blue indicates that surveys lead to total cost savings, while red indicates that surveying increases total costs. Applying the optimal level of surveillance in this example would reduce total costs by 26%. Applying the level of surveillance obtained by wrongly interpreting the output of the PB SDM as a probability would lead to an increase in total costs (22%).

Only when the stacked SDMs predict true probabilities is the estimation of species richness unbiased (Fig. 5). Attempts to estimate richness based on SDM outputs that only represent relative likelihoods of species occupancy are arbitrarily biased (in our example, richness is overestimated). Binary conversion of SDM outputs does not solve the problems associated with relative likelihoods and it is clearly detrimental when applied to SDMs that estimate probabilities (Fig. 5), as recently shown mathematically by Calabrese *et al.* (2014). Our simulated examples suggest that continuous (i.e. uncategorized) relative likelihoods may broadly capture *relative* patterns of species richness in *some* scenarios, even if they are unable to estimate absolute numbers of species (i.e. actual and estimated richness correlated). This finding is compatible with results presented in the literature (e.g. Aranda & Lobo, 2011), although further research is needed before establishing how widely this pattern applies in real systems.

## Case study 5: spatial prioritization

SDMs are frequently used to prioritize areas for conservation actions such as reserve establishment or restoration. To date, multiple methods and software tools have been applied to spatial conservation planning, including sophisticated approaches based on algorithms that can solve highly complex spatial problems (Moilanen *et al.*, 2009). The two most frequently used modes of planning are: (1) the minimum set approach (Cocks & Baird, 1989), in which areas are selected in order to meet species-specific (or feature-specific) targets while minimizing the cost of the solution, and (2) the maximum coverage/utility approach (Hof & Raphael, 1993; Camm *et al.*, 1996), in which biodiversity benefits are maximized within given budgetary or area constraints. Here we explore the effect of SDM output scaling and binary conversion on reserve selection when using a target-based minimum set approach. We use a
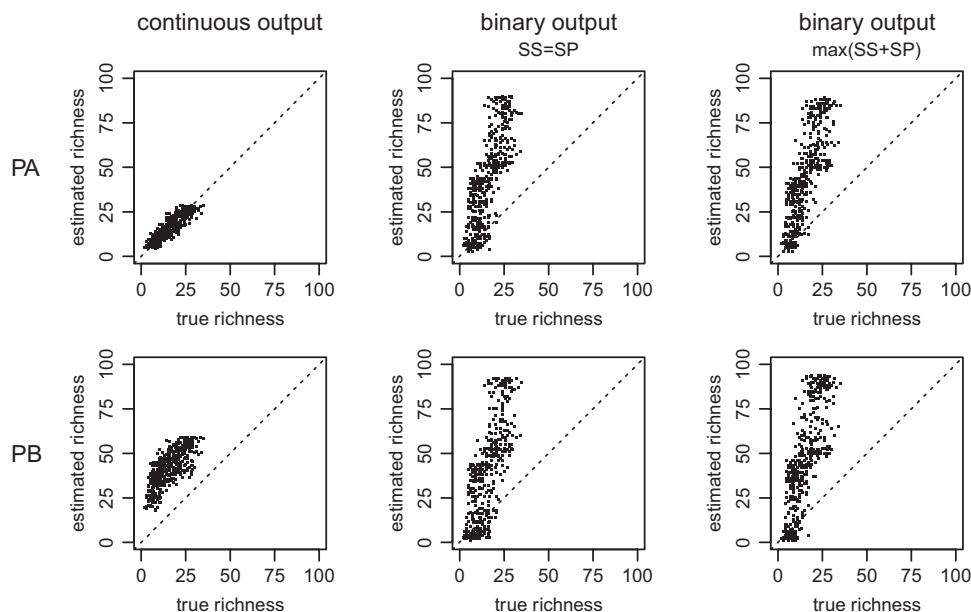
**Figure 5** Species richness case study. Plots display estimated versus true richness for 500 sites, based on stacking presence–absence (PA) and presence-background (PB) species distribution models (SDMs; first and second row, respectively). A total of 100 species were simulated. In the first column, estimated richness is obtained as the sum of the continuous output of the SDMs, without a threshold. In the second and third columns the SDM output is first converted to a binary output using a threshold ('SS = SP' or 'max (SS + SP)'; see threshold definitions in the main text and similar results for other thresholds in Appendix S4). The sum of continuous outputs of PA SDMs provides a good estimation of species richness; PB SDMs cannot estimate richness well given that their output does not represent a probability of presence. Applying thresholds does not solve the problem and is detrimental when using PA data. Despite not being able to estimate absolute richness, in this example PB SDMs provide a relatively robust *ranking* of sites in terms of richness (although worse than PA SDMs).

widely adopted planning tool, Marxan (Ball *et al.* 2009), as an example but the conclusions derived are equally valid for other tools within this group. In Appendix S4 we present the analogous analysis using Zonation (Moilanen *et al.*, 2012) as an example of maximum coverage/utility approaches.

We used SDMs based on real PA data for seven species from the Lower Hunter region in Australia (Wintle *et al.*, 2005a), rescaled them to mimic PB methods that produce estimates of relative likelihoods (scaling between 0.3 and 0.9) and applied the thresholds above to obtain binary SDMs (details in Appendix S4). For the Marxan analysis, the amount of each conservation feature (species) contained in each cell was set to the prediction of the corresponding SDM. Two types of target were assessed: 'proportional' (conserve a set proportion of the area that each species occupies) and 'absolute' (conserve a set area occupied by each species). The absolute areas were chosen by multiplying the proportional targets by the total area occupied by each species according to the original PA-based SDMs. The planning solutions produced after applying scaling and binary conversion to the SDMs were compared with results obtained using the original SDMs, in terms of the number of 'true' species distributions protected in each solution (the 'true' distributions were simulated based on the original SDM probabilities).

We found that planning solutions based on relative likelihoods are comparable to those based on true occurrence prob-

abilities *provided* proportional targets are used (Fig. 6b). However, the minimum set approach over- or under-shoots true absolute targets when based on SDMs that estimate relative likelihoods (Fig. 6a). When SDMs underestimate prevalence (scaling < 1), targets are met but solutions are more costly (e.g. the cost of achieving the targets was double in Fig. 6a when prevalence was underestimated; open circles). Conversely, if SDMs overestimate prevalence, the algorithm interprets that it is conserving more than it actually does, and conservation targets are missed. Binary conversion of SDM outputs leads to a reserve system that fails to meet true targets and/or is more costly, even if relative targets are used (Fig. 6c, d): the information loss ultimately results in expending excessive resources on some species while under-representing others. Targets and achieved levels of conservation disagree more when absolute targets are used.

Although we have shown that the use of relative likelihoods does not affect the planning solution when targets (or algorithms) are based on proportions of species distributions, it is important to realize that the true area occupied by species in reserves will remain unknown. This limitation may have implications in ecological, biogeographical and conservation applications. For example, one will not be able to judge whether the occupied area within a reserve is large enough to represent a viable population of the species.

**Figure 6** Spatial prioritization case study. Scatterplots represent the proportion of occupied area selected by Marxan (*y*-axis) for the seven considered species with respect to the requested target (*x*-axis) in different scenarios. In all panels the same targets are applied, although they are expressed in two ways: as absolute areas (left panels) or as a proportion of area (right panels). Closed circles represent solutions produced for the original SDM probabilities, open circles for scaled probabilities, and other symbols represent solutions where thresholds were applied to the latter to create binary SDM outputs: 'MTP' (✗), '10% TP' (◇), 'SS = SP' (△), 'max (SS + SP)' (▽) (see threshold definitions in the main text). The values displayed in the legend of each plot represent the total cost for each prioritization solution.
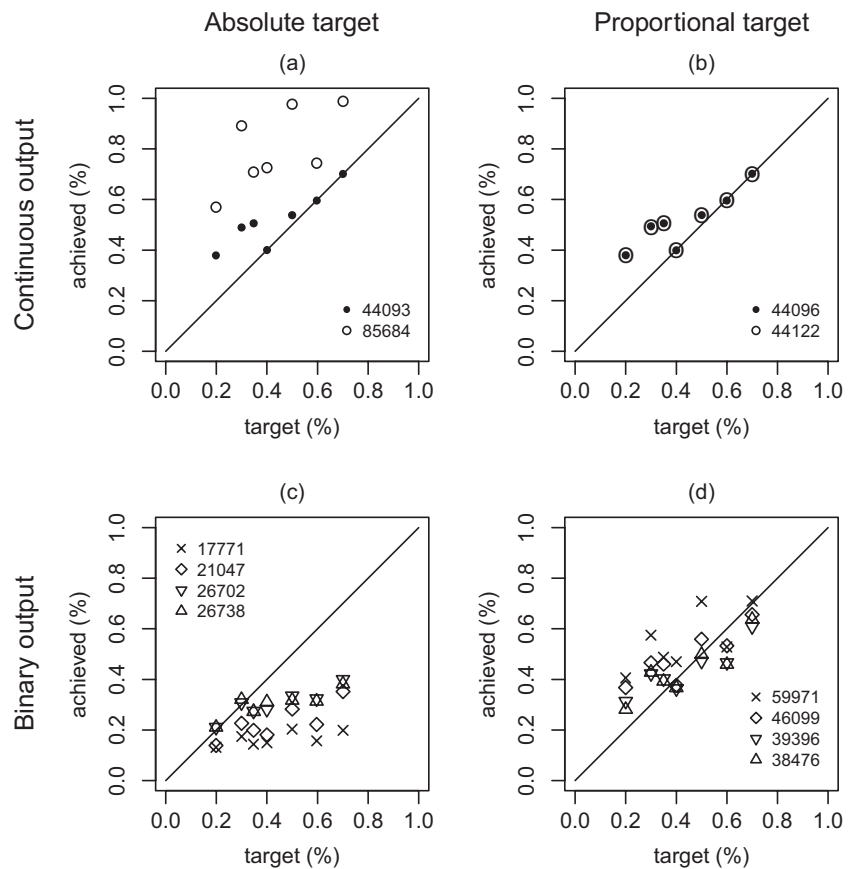
## DISCUSSION

We have shown how the type of survey data, sampling bias and imperfect detection interact to determine the quantity that is ultimately estimated by a SDM (Fig. 1), and the implications of reducing SDM outputs to a binary categorization. Each of these issues have been discussed in the literature (e.g. Phillips *et al.*, 2009; Phillips & Elith, 2013; Calabrese *et al.*, 2014; Lahoz-Monfort *et al.*, 2014), but have generally been addressed as separate problems, with little or no evaluation of their implications for particular applications. We have integrated these issues into a single framework to aid interpretation and explicit consideration of the implications for the most common applications of SDMs. Our simulations demonstrate that these issues really matter and that misguided or inefficient management and inference may result if they are ignored.

### SDMs, data and decision-theory

It has recently been advocated that the modelling of species distributions should be carried out within a structured and transparent decision-making framework (Guisan *et al.*, 2013). Formulating a clear objective is the starting point for such a process. To ensure that a SDM is ultimately fit for its purpose one must first ask 'what is my SDM to be used for?'. Here we argue that a critical consideration is whether the type of information demanded by the application in question is available or

could be obtained. In practice, this requires evaluation of whether the SDM output needs to provide estimates of actual occupancy probabilities or whether relative likelihoods, or the ranking of sites, will suffice. Our assessment shows that while learning about the relative likelihood of species occupancy in different sites is sufficient for some applications (e.g. spatial prioritization when targets are expressed as proportion of distributions, Fig. 6; or cost-sharing agreements about invasive species management, Appendix S3), for many other applications information about the actual probability that a site contains a species is crucial (e.g. species prioritization according to the area occupied, Fig. 6; or estimation of species richness, Fig. 5). One should also evaluate whether the data (existing or to be collected) can be assumed to be free of considerable biases or, otherwise, whether the type of data allows unbiased estimation when used in conjunction with appropriate modelling methods. Thinking about the type of output that a SDM is expected to provide for a given application is also important from the point of view of model evaluation, as it directly relates to the metrics that should be used to assess its performance (Lawson *et al.*, 2014).

Where an application appears to require only knowledge about relative likelihoods, it is essential to consider carefully whether SDM predictions are indeed comparable within that specific context. For instance, assuming that detectability and sampling effort do not vary spatially, predictions from a PB model represent relative likelihoods and these are comparable

across space. However, comparisons *across time* or *between species* are not valid because prevalence will differ across SDMs and cannot be estimated from PB data: the predicted SDM values in a given location are relative to other locations modelled but are not comparable *across models* fitted for different species or over different time periods. Thus, even if relative likelihood predictions for two species happen to have the same value at a given location, it does not follow that both species have the same probability of occurrence on those locations. The same reasoning applies to temporal comparisons. In general, this rules out the use of SDMs fitted to PB data in monitoring temporal trends in species AOO (Fig. 2) or in the prioritization of a set of species according to their AOOs (Fig. 3). Trends in AOO can be explored with PB data when *projecting* a SDM to changing conditions but not when separate models are *fitted* to capture changing relationships with the predictors.

### Binary conversion of SDM predictions

Our case studies show how the conversion of SDM predictions into discrete categories diminishes the value of SDMs for a range of applications such as spatial prioritization (Fig. 6), species richness estimation (Fig. 5) or applications that require estimating species AOO (Figs. 2 & 3). Our results support other recent studies that argue against the discretization of SDM outputs for specific applications (e.g. Calabrese *et al.*, 2014; Lawson *et al.*, 2014). In short, the continuous outputs of SDMs provide richer information than discrete representations, and therefore the discretization of SDM outputs is detrimental in most applications.

The commonly held belief that applications require binary SDM outputs may stem partly from unrealistic expectations about what a SDM can, or should, provide. In an ideal world, a SDM would provide perfect discrimination between sites at which species are present and absent, leading to a binary map of true occurrence. However, data limitations (a sample of observations) and the inherently stochastic nature of site occupancy (i.e. the fact that reality tends to be too complex to be categorically described by measurable covariates in a statistical model) mean that SDMs can only be expected to provide *probabilistic predictions* of species occurrence. Transforming such probabilistic outputs into binary maps using a threshold does not provide an estimate of site occupancy that is closer to the true presences and absences; on the contrary, discretization degrades the information available in SDM predictions (Fig. S2.2 in Appendix S2). Some studies have used binary conversion in an attempt to get around the problem of not being able to estimate prevalence in PB SDMs (e.g. Pineda & Lobo, 2009; Liu *et al.*, 2013). Our simulations confirm that simplifying PB-based SDM outputs into binary categorizations does not solve this problem (Figs 2, 3, 5 & 6).

Our review of the applications of SDMs indicates that, despite being a relatively widespread practice, SDM outputs should rarely be converted into binary maps (Appendix S3). In this regard, we again emphasize the need to use SDMs within a structured decision-making framework. Specifying exactly how SDM outputs are to be used for particular applications helps to clarify whether binary conversion is actually necessary or beneficial, and which threshold is most appropriate, if any. For instance, a management body may decide to identify 'critical habitat' for a species to grant relevant sites special protection. In this case, the objective involves classifying the habitat into two categories (protected or not), hence binary conversion is fully justified by the objective. Yet not all conversions would be equally acceptable: the objective should be articulated to provide the rules to select the appropriate threshold. For example, if the aim were to protect 90% of a species' distribution, the threshold that achieved a sensitivity of 0.9 should be used. However, if the aim were instead to protect the *n* best sites for the species dictated by the available budget, a different threshold would be applied.

### Future prospects and recent developments

Due to the availability of data PB methods are frequently used despite their limitations. Future efforts should be directed towards the collection of more informative survey data, which will lead to SDMs that can support a wider range of applications in ecology, biogeography and conservation. A first step is to ensure that non-detections are recorded. Wider incorporation of non-detections into databases that bring together observations from different sources will render them much more useful, as this addresses the problem of sampling bias (Feeley & Silman, 2011) and allows the estimation of species prevalence. A greater appreciation of the limitations of PB data should also lead to careful investigation of existing datasets, to assess whether absence data can actually be extracted from raw survey data. Where PB data do not deliver the information required, they might still be usefully integrated with limited PA data to estimate distributions better than simply using PA data alone (Dorazio, 2014; Fithian *et al.*, 2014).

Collecting data in ways that allow one to deal with imperfect detection should also be an important component in designing future sampling protocols, while available data sources should be carefully examined to evaluate whether they allow this (e.g. Kéry *et al.*, 2010). Statistical methods that explicitly model the observation process (i.e. state-space models) are valuable tools for analysing ecological data (King, 2014; Guillera-Arroita *et al.*, 2014b). Apart from the development of models of species distributions that account for possible *false absence* records, recent statistical advances include extensions to model range dynamics and multiple occupancy states (Bailey *et al.*, 2014). Methods have also been developed to deal with *false positive* records which, although less prevalent, can be relevant for some types of surveys (Miller *et al.*, 2011). In summary, to reliably infer the distribution of a species, steps need to be taken during survey design, data collection and analysis to minimize the effects of the sampling process.

### CONCLUSION

Many issues need to be considered when attempting to build useful ecological models. As we have shown, a fundamental

question in the context of SDMs is whether the type and quality of data available are suitable for producing the information required for a given application. Not all data types or SDM approaches are valid for all applications in ecology, biogeography and conservation, and using inadequate SDM outputs may lead to misguided decisions and suboptimal use of resources. For some applications, learning about which variables are good predictors of a species' distribution is sufficient; for many others, however, knowledge about the prevalence of the species is required. Our results demonstrate that binary conversion of SDM outputs does not solve this problem, and should only be carried out when it is clearly justified by the application's objective. Framing the use of SDMs within a structured decision-making context helps to ensure that the resulting model is fit for its intended purpose.

## ACKNOWLEDGEMENTS

## REFERENCES

Akcakaya, H.R., McCarthy, M.A. & Pearce, J.L. (1995) Linking landscape data with population viability analysis – management options for the helmeted honeyeater *Lichenostomus melanops cassidix*. *Biological Conservation*, **73**, 169–176.

Aranda, S.C. & Lobo, J.M. (2011) How well does presence-only-based species distribution modelling predict assemblage diversity? A case study of the Tenerife flora. *Ecography*, **34**, 31–38.

Bailey, L.L., MacKenzie, D.I. & Nichols, J.D. (2014) Advances and applications of occupancy models. *Methods in Ecology and Evolution*, in press. doi:10.1111/2041-210X.12100.

Ball, I.R., Possingham, H.P. & Watts, M.E. (2009) Marxan and relatives: Software for spatial conservation prioritisation. Pages 185–196 in A. Moilanen, K.A. Wilson, and H.P. Possingham, editors. *Spatial Conservation Prioritization: Quantitative Methods and Computational Tools*. Oxford University Press, New York City.

Bekessy, S.A., Wintle, B.A., Gordon, A., Fox, J.C., Chisholm, R., Brown, B., Regan, T., Mooney, N., Read, S.M. & Burgman, M.A. (2009) Modelling human impacts on the Tasmanian wedge-tailed eagle (*Aquila audax fleayi*). *Biological Conservation*, **142**, 2438–2448.

Booth, T.H., Nix, H.A., Busby, J.R. & Hutchinson, M.F. (2014) Bioclim: the first species distribution modelling package, its early applications and relevance to most current MaxEnt studies. *Diversity and Distributions*, **20**, 1–9.

Busby, J.R. (1991) BIOCLIM – a bioclimate analysis and prediction system. *Plant Protection Quaterly*, **6**, 8–9.

Calabrese, J.M., Certain, G., Kraan, C. & Dormann, C.F. (2014) Stacking species distribution models and adjusting bias by linking them to macroecological models. *Global Ecology and Biogeography*, **23**, 99–112.

Camm, J.D., Polasky, S., Solow, A. & Csuti, B. (1996) A note on optimal algorithms for reserve site selection. *Biological Conservation*, **78**, 353–355.

Chakraborty, A., Gelfand, A.E., Wilson, A.M., 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.

Chen, G., Kéry, M., Plattner, M., Ma, K. & Gardner, B. (2013) Imperfect detection is the rule rather than the exception in plant distribution studies. *Journal of Ecology*, **101**, 183–191.

Cocks, K.D. & Baird, I.A. (1989) Using mathematical-programming to address the multiple reserve selection problem – an example from the Eyre Peninsula, South Australia. *Biological Conservation*, **49**, 113–130.

Dorazio, R.M. (2012) Predicting the geographic distribution of a species from presence-only data subject to detection errors. *Biometrics*, **68**, 1303–1312.

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

Elith, J. & Franklin, J. (2013) Species distribution modeling. *Encyclopedia of biodiversity* (ed. by S.A. Levin), pp. 692–705. Academic Press, Waltham, MA.

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

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

Feeley, K.J. & Silman, M.R. (2011) Keep collecting: accurate species distribution modelling requires more collections than previously thought. *Diversity and Distributions*, **17**, 1132–1140.

Ferrier, S. & Guisan, A. (2006) Spatial modelling of biodiversity at the community level. *Journal of Applied Ecology*, **43**, 393–404.

Ferrier, S. & Watson, G. (1997) *An evaluation of the effectiveness of environmental surrogates and modelling techniques in predicting the distribution of biological diversity*. Environment Australia, Canberra.

Ferrier, S., Watson, G., Pearce, J. & Drielsma, M. (2002) Extended statistical approaches to modelling spatial pattern in biodiversity in northeast New South Wales. I. Species-level modelling. *Biodiversity and Conservation*, **11**, 2275–2307.

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

Fithian, W., Elith, J., Hastie, T. & Keith, D. (2014) Bias correction in species distribution models: pooling survey and collection

data for multiple species. *Methods in Ecology and Evolution*, in press. doi: 10.1111/2041-210X.12242

Franklin, J. (2013) Species distribution models in conservation biogeography: developments and challenges. *Diversity and Distributions*, **19**, 1217–1223.

Garrard, G.E., Bekessy, S.A., McCarthy, M.A. & Wintle, B.A. (2008) When have we looked hard enough? A novel method for setting minimum survey effort protocols for flora surveys. *Austral Ecology*, **33**, 986–998.

Gu, W. & Swihart, R.H. (2004) Absent or undetected? Effects of non-detection of species occurrence on wildlife–habitat models. *Biological Conservation*, **116**, 195–203.

Guillera-Arroita, G. & Lahoz-Monfort, J.J. (2012) Designing studies to detect differences in species occupancy: power analysis under imperfect detection. *Methods in Ecology and Evolution*, **3**, 860–869.

Guillera-Arroita, G., Morgan, B.J.T., Ridout, M.S. & Linkie, M. (2011) Species occupancy modeling for detection data collected along a transect. *Journal of Agricultural, Biological, and Environmental Statistics*, **16**, 301–317.

Guillera-Arroita, G., Hauser, C.E. & McCarthy, M.A. (2014a) Optimal surveillance strategy for invasive species management when surveys stop after detection. *Ecology and Evolution*, **4**, 1751–1760.

Guillera-Arroita, G., Lahoz-Monfort, J.J., MacKenzie, D.I., Wintle, B.A. & McCarthy, M.A. (2014b) Ignoring imperfect detection in biological surveys is dangerous: a response to 'Fitting and Interpreting Occupancy Models'. *PLoS ONE*, **9**, e99571.

Guillera-Arroita, G., Lahoz-Monfort, J.J. & Elith, J. (2014c) Maxent is not a presence–absence method: a comment on Thibaud *et al. Methods in Ecology and Evolution*, in press. doi: 10.1111/2041-210X.12252

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

Guisan, A., Tingley, R., Baumgartner, J.B. *et al.* (2013) Predicting species distributions for conservation decisions. *Ecology Letters*, **16**, 1424–1435.

Hastie, T. & Fithian, W. (2013) Inference from presence-only data; the ongoing controversy. *Ecography*, **36**, 864–867.

Hauser, C.E. & McCarthy, M.A. (2009) Streamlining 'search and destroy': cost-effective surveillance for invasive species management. *Ecology Letters*, **12**, 683–692.

Herbert, K., Curran, I., Hauser, C., Hester, S., Kirkwood, A., Derrick, B. & King, D. (2013) All eyes focussed on hawkweed eradication in the Victorian Alps – a partnerships approach. *Proceedings of the 17th NSW Weeds Conference. Corowa, Australia*, pp. 140–145. Available at: http://nswweedsoc.org.au/items/839/17NSWWC%20Proceedings%202013.pdf

Hof, J.G. & Raphael, M.G. (1993) Some mathematical-programming approaches for optimizing timber age-class distributions to meet multispecies wildlife population objectives. *Canadian Journal of Forest Research–Revue Canadienne De Recherche Forestiere*, **23**, 828–834.

IUCN (2012) *IUCN Red List Categories and Criteria: version 3.1*, 2nd edn. IUCN, Gland, Switzerland and Cambridge, UK.

Jiménez-Valverde, A. & Lobo, J.M. (2007) Threshold criteria for conversion of probability of species presence to either–or presence–absence. *Acta Oecologica*, **31**, 361–369.

Keith, D.A., Akcakaya, H.R., Thuiller, W., Midgley, G.F., Pearson, R.G., Phillips, S.J., Regan, H.M., Araujo, M.B. & Rebelo, T.G. (2008) Predicting extinction risks under climate change: coupling stochastic population models with dynamic bioclimatic habitat models. *Biology Letters*, **4**, 560–563.

Keith, D.A., Elith, J. & Simpson, C.C. (2014) Predicting distribution changes of a mire ecosystem under future climates. *Diversity and Distributions*, **20**, 440–454.

Kéry, M. (2002) Inferring the absence of a species: a case study of snakes. *Journal of Wildlife Management*, **66**, 330–338.

Kéry, M. (2011) Towards the modelling of true species distributions. *Journal of Biogeography*, **38**, 617–618.

Kéry, M., Gardner, B. & Monnerat, C. (2010) Predicting species distributions from checklist data using site-occupancy models. *Journal of Biogeography*, **37**, 1851–1862.

Kéry, M., Guillera-Arroita, G. & Lahoz-Monfort, J.J. (2013) Analysing and mapping species range dynamics using occupancy models. *Journal of Biogeography*, **40**, 1463–1474.

King, R. (2014) Statistical ecology. *Annual Review of Statistics and its Application*, **1**, 401–426.

Lahoz-Monfort, J.J., Guillera-Arroita, G. & Wintle, B.A. (2014) Imperfect detection impacts the performance of species distribution models. *Global Ecology and Biogeography*, **23**, 504–515.

Lancaster, T. & Imbens, G. (1996) Case–control studies with contaminated controls. *Journal of Econometrics*, **71**, 145–160.

Lawson, C.R., Hodgson, J.A., Wilson, R.J. & Richards, S.A. (2014) Prevalence, thresholds and the performance of presence–absence models. *Methods in Ecology and Evolution*, **5**, 54–64.

Lele, S.R. & Keim, J.L. (2006) Weighted distributions and estimation of resource selection probability functions. *Ecology*, **87**, 3021–3028.

Lele, S.R., Merrill, E.H., Keim, J. & Boyce, M.S. (2013) Selection, use, choice and occupancy: clarifying concepts in resource selection studies. *Journal of Animal Ecology*, **82**, 1183–1191.

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

Liu, C., Berry, P.M., Dawson, T.P. & Pearson, R.G. (2005) Selecting thresholds of occurrence in the prediction of species distributions. *Ecography*, **28**, 385–393.

Liu, C.R., White, M. & Newell, G. (2013) Selecting thresholds for the prediction of species occurrence with presence-only data. *Journal of Biogeography*, **40**, 778–789.

Lizzio, J., Richmond, L., Mewett, O., Hennecke, B., Baker, J. & Raphael, B. (2009) *Methodology to prioritise weeds of national significance (WONS) candidates*. Bureau of Rural Sciences, Canberra.

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

MacKenzie, D.I., Nichols, J.D., Lachman, G.B., Droege, S., Royle, J.A. & Langtimm, C.A. (2002) Estimating site occupancy rates when detection probabilities are less than one. *Ecology*, **83**, 2248–2255.

MacKenzie, D.I., Nichols, J.D., Royle, J.A., Pollock, K.H., Bailey, L.L. & Hines, J.E. (2006) *Occupancy estimation and modeling: inferring patterns and dynamics of species occurrence.* Academic Press, New York.

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

Merow, C., Smith, M.J. & Silander, J.A. (2013) A practical guide to MaxEnt for modeling species' distributions: what it does, and why inputs and settings matter. *Ecography*, **36**, 1058–1069.

Miller, D.A., Nichols, J.D., McClintock, B.T., Campbell Grant, E.H., Bailey, L.L. & Weir, L.A. (2011) Improving occupancy estimation when two types of observation error occur: non-detection and species misidentification. *Ecology*, **92**, 1422–1428.

Moilanen, A., Franco, A.M.A., Early, R.I., Fox, R., Wintle, B. & Thomas, C.D. (2005) Prioritizing multiple-use landscapes for conservation: methods for large multi-species planning problems. *Proceedings of the Royal Society B: Biological Sciences*, **272**, 1885–1891.

Moilanen, A., Wilson, K.A. & Possingham, H.P. (2009) *Spatial conservation prioritization. Quantitative methods and computational tools.* Oxford University Press, New York.

Moilanen, A., Meller, L., Leppänen, J., Montesino Pouzols, F., Arponen, A. & Kujala, H. (2012) Zonation: spatial conservation planning framework and software version 3.1 user manual. Helsingin Yliopisto, Helsinki.

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.

Newbold, T., Gilbert, F., Zalat, S., El-Gabbas, A. & Reader, T. (2009) Climate-based models of spatial patterns of species richness in Egypt's butterfly and mammal fauna. *Journal of Biogeography*, **36**, 2085–2095.

Parviainen, M., Marmion, M., Luoto, M., Thuiller, W. & Heikkinen, R.K. (2009) Using summed individual species models and state-of-the-art modelling techniques to identify threatened plant species hotspots. *Biological Conservation*, **142**, 2501–2509.

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

Pearce, J. & Lindenmayer, D. (1998) Bioclimatic analysis to enhance reintroduction biology of the endangered helmeted honeyeater (*Lichenostomus melanops cassidix*) in southeastern Australia. *Restoration Ecology*, **6**, 238–243.

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. & Elith, J. (2013) On estimating probability of presence from use-availability or presence-background data. *Ecology*, **94**, 1409–1419.

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

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

Pineda, E. & Lobo, J.M. (2009) Assessing the accuracy of species distribution models to predict amphibian species richness patterns. *Journal of Animal Ecology*, **78**, 182–190.

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.

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.

Schmidt-Lebuhn, A.N., Knerr, N.J. & Gonzalez-Orozco, C.E. (2012) Distorted perception of the spatial distribution of plant diversity through uneven collecting efforts: the example of Asteraceae in Australia. *Journal of Biogeography*, **39**, 2072–2080.

Stauffer, H.B., Ralph, C.J. & Miller, S.L. (2002) Incorporating detection uncertainty into presence–absence surveys for marbled murrelet. *Predicting species occurrences: issues of accuracy and scale* (ed. by J.M. Scott, P.J. Heglund, M.L. Morrison, J.B. Haufler, M.G. Raphael, W.A. Wall and F.B. Samson), pp. 357–365. Island Press, Washington, DC.

Trotta-Moreu, N. & Lobo, J.M. (2010) Deriving the species richness distribution of Geotrupinae (Coleoptera: Scarabaeoidea) in Mexico from the overlap of individual model predictions. *Environmental Entomology*, **39**, 42–49.

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

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., Renner, I.W. & Ramp, D. (2013) Model-based control of observer bias for the analysis of presence-only data in ecology. *Plos One*, **8**, e79168.

Wintle, B.A., Elith, J. & Potts, J.M. (2005a) Fauna habitat modelling and mapping: a review and case study in the Lower Hunter Central Coast region of NSW. *Austral Ecology*, **30**, 719–738.

Wintle, B.A., Kavanagh, R.P., McCarthy, M.A. & Burgman, M.A. (2005b) Estimating and dealing with detectability in

occupancy surveys for forest owls and arboreal marsupials. *Journal of Wildlife Management*, **69**, 905–917.

Yackulic, C.B., Chandler, R., Zipkin, E.F., Royle, J.A., Nichols, J.D., Grant, E.H.C. & Veran, S. (2013) Presence-only modelling using MAXENT: when can we trust the inferences? *Methods in Ecology and Evolution*, **4**, 236–243.

Yoccoz, N.G., Nichols, J.D. & Boulinier, T. (2001) Monitoring of biological diversity in space and time. *Trends in Ecology and Evolution*, **16**, 446–453.

Additional references to reviewed papers for the literature review and assessment of applications can be found at the end of Appendices S1 & S3 in the Supporting Information.

## SUPPORTING INFORMATION

Additional supporting information may be found in the online version of this article at the publisher's web-site.

**Appendix S1** Literature review.
**Appendix S2** Additional figures.
**Appendix S3** Assessment of output adequacy of species distribution models in ecological, biogeographical and conservation applications.
**Appendix S4** Details about case studies.

## BIOSKETCH

The authors belong to the ARC Centre of Excellence for Environmental Decisions (http://www.ceed.edu.au). They share an interest in the construction of species distribution models and their application to inform environmental decisions, as well as in research on methodological aspects of such models. Their collective experience covers a wide range of SDM applications such as wildlife monitoring, invasive species management, population viability analysis and spatial conservation planning, as well as environmental decision analysis more generally.

Editor: Niklaus Zimmermann