Methods in Ecology and Evolution

British Ecological Society

Methods in Ecology and Evolution 2014, **5,** 54–64 doi: 10.1111/2041-210X.12123

Prevalence, thresholds and the performance of presence–absence models

Callum R. Lawson[1,2]*, Jenny A. Hodgson[3], Robert J. Wilson[1] and Shane A. Richards[4]

[1]Centre for Ecology and Conservation, University of Exeter, Cornwall Campus, Penryn TR10 9EZ, UK; [2]Department of Animal Ecology, Netherlands Institute of Ecology (NIOO-KNAW), Wageningen 6700AB, The Netherlands; [3]Department of Evolution, Ecology and Behaviour, University of Liverpool, Biosciences Building, Crown Street, Liverpool L69 7ZB, UK; and [4]School of Biological and Biomedical Sciences, Durham University, South Road, Durham DH1 3LE, UK

Summary

1. The use of species distribution models to understand and predict species' distributions necessitates tests of fit to empirical data. Numerous performance metrics have been proposed, many of which require continuous occurrence probabilities to be converted to binary 'present or absent' predictions using threshold transformations. It is widely accepted that both continuous and binary performance metrics should be independent of prevalence (the proportion of locations that are occupied). However, because these metrics have been mostly assessed on a case-specific basis, there are few general guidelines for measuring performance.

2. Here, we develop a conceptual framework for classifying performance metrics, based on whether they are sensitive to prevalence, and whether they require binary predictions. We use this framework to investigate how these performance metric properties influence the predictions made by the models they select.

3. A literature survey reveals that binary metrics are widely employed and that prevalence-independent metrics are used more frequently than prevalence-dependent metrics. However, we show that prevalence-dependent metrics are essential to assess the numerical accuracy of model predictions and are more useful in applications that require occupancy estimates. Furthermore, we demonstrate that in comparison with continuous metrics, binary metrics often select models that have reduced ability to separate presences from absences, make predictions which over- or underestimate occupancy and give misleading estimates of uncertainty. Importantly, models selected using binary metrics will often be of reduced practical use even when applied to ecological problems that require binary decision-making.

4. We suggest that SDM performance should be assessed using prevalence-dependent performance metrics whenever the absolute values of occurrence predictions are important and that continuous metrics should be used instead of binary metrics whenever possible. We thus recommend the wider application of prevalence-dependent continuous metrics, particularly likelihood-based metrics such as Akaike's Information Criterion (AIC), to assess the performance of presence–absence models.

Key-words: Area under the curve (AUC), Cohen's kappa, concordance index, deviance, explanatory power, goodness-of-fit, sensitivity, specificity, true skill statistic (TSS), *R*-squared

Introduction

Species distribution models (SDMs) are useful tools for understanding species' environmental requirements and for predicting their responses to environmental change (Peterson *et al.* 2011). SDMs describe how species are distributed in geographical space and/or across different environments (Fig. 1; Elith & Leathwick 2009), based on one of three types of data: observations of presence or absence (more correctly, detection and non-detection; Yackulic *et al.* 2012), point locations of known presence ('presence-only' data, sometimes used in combination with simulated 'pseudo-absence' or 'background' points) or counts of the number of individuals (abundance or proportional cover) in a given area. In each case, the development of informative and predictive SDMs necessitates measuring their performance, that is, how well predictions match an observed data set of occupied (present) and unoccupied (absent) locations; this study focuses on how to measure the performance of presence–absence models (see Hirzel *et al.* 2006 and Phillips & Elith 2010 for performance assessment in presence-only models).

There are several complementary reasons for measuring SDM predictive performance, including the following: (i) to test which environmental variables have an important influence on the distribution of a species, (ii) to find the model (parameter set or modelling method) that best predicts a species' distribution, (iii) to assess the reliability of predictions, or (iv) to suggest areas for model improvement (model criticism;

*Correspondence author. E-mail: C.Lawson@nioo.knaw.nl

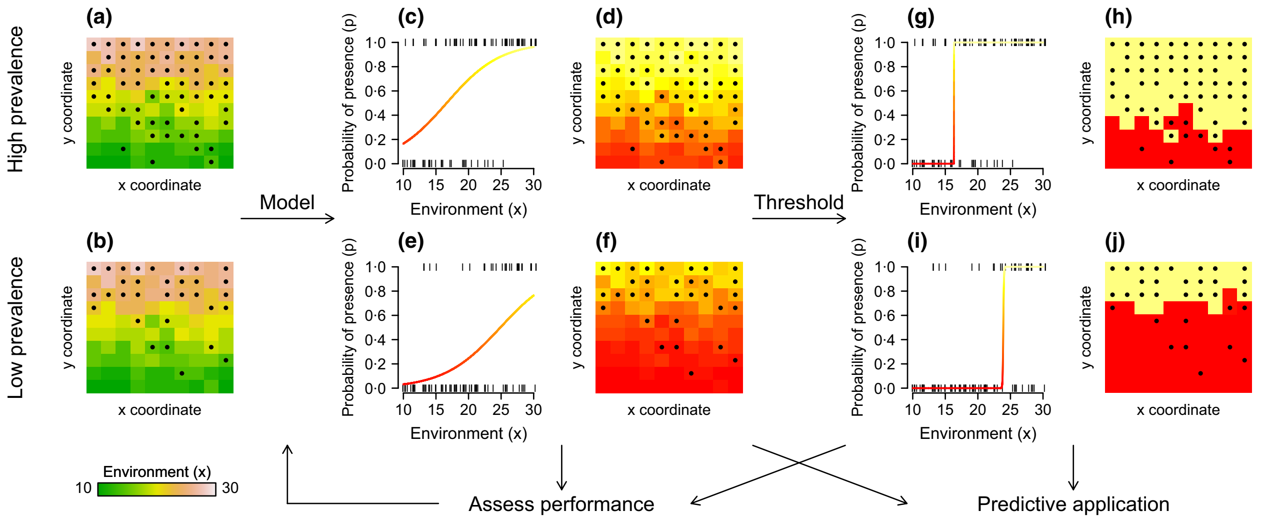

Fig. 1. Outline of a typical species distribution modelling process. A geographical area is surveyed for the presence (circles) and absence (blank squares) of a species across a discretized set of cells (squares) with a range of different environments (a–b). The species may be common (high prevalence; top row) or rare (low prevalence; bottom row) across the landscape. A species distribution model is constructed which predicts the probability of presence in different environments; both predictions and data may be viewed in environmental space (c, e, g, i) or geographical space (d, f, h, j). The initial probabilistic predictions (c–f) may be transformed into binary predictions by selecting an appropriate transformation threshold (g–j). Performance is assessed by comparing the predictions with the data (see Figure 2), following which the model may be further improved or used to predict occurrence in novel places or times, for example to facilitate conservation decision-making.

Guisan & Zimmermann 2000; Bolker 2008; Peterson *et al.* 2011). Performance assessment forms part of a cycle in which SDMs are formulated, tested and subsequently improved (Fig. 1), culminating in SDMs that are applied to practical problems such as the design of conservation strategies (Moilanen, Wilson & Possingham 2009) or forecasting species distributions in novel places or times (Fig. 1; Guisan & Zimmermann 2000; Peterson *et al.* 2011).

Species distribution model performance is frequently assessed using scalar performance metrics (Liu, White & Newell 2011). Different metrics assess different aspects of performance, which means that the choice of metric has an important influence on SDM parameterization and model selection, and thus which models are used for prediction (Fig. 1; see Pearce & Ferrier 2000; Vaughan & Ormerod 2005 for reviews). The properties of specific performance metrics have been explored using real or simulated data sets (e.g. McPherson, Jetz & Rogers 2004; Allouche, Tsoar & Kadmon 2006; Lobo, Jiménez-Valverde & Real 2008; Foody 2011), but there has been relatively little emphasis on developing a general theory that would reveal which metrics are best for a given application (Elith & Leathwick 2009; Peterson *et al.* 2011). This study develops a conceptual framework for SDM performance metrics, based on two 'axes' which define their properties, and explores the practical consequences of performance metric choice.

The first property we investigate is whether the metric is sensitive to the proportion of locations that are observed to be occupied, known as *prevalence* (Royle *et al.* 2012). Prevalence is generally lower for rarer species, for distributions surveyed at higher spatial resolutions (Kunin 1998), and when populations have yet to reach carrying capacity, such as when a species has recently colonized a new area (Fig. 1; Fielding 2007). The scores of many performance metrics are affected by prevalence; this effect is viewed as a statistical artefact (Pearce & Ferrier 2000; Vaughan & Ormerod 2005; Allouche, Tsoar & Kadmon 2006; Lobo, Jiménez-Valverde & Real 2008), and the application of performance metrics that eliminate the effects of prevalence has been widely advocated (McPherson, Jetz & Rogers 2004; Allouche, Tsoar & Kadmon 2006; Lobo, Jiménez-Valverde & Real 2008). In contrast, we will show that prevalence is essential to assess the numerical accuracy of SDM predictions.

The second property we investigate is whether the performance metric is 'threshold-dependent'; that is, whether the response variable is classified as present or absent. To use threshold-dependent performance metrics, continuous predictions – usually probability of presence estimates – are transformed to binary (present or absent) predictions by selecting a threshold probability above which predictions are classed as present and below which predictions are classified as absent (Fig. 1; Liu *et al.* 2005). Both the predictions and the performance of binary SDMs are altered by the choice of threshold, with fewer presences (lower prevalence) predicted when a higher threshold is chosen (Liu *et al.* 2005; Lobo, Jiménez-Valverde & Real 2008; Bean, Stafford & Brashares 2011). Although the use of threshold-dependent metrics has been widely encouraged (Liu *et al.* 2005; Lobo, Jiménez-Valverde & Real 2008; Bean, Stafford & Brashares 2011; but see Vaughan & Ormerod 2005), we will show why their use should generally be avoided in favour of threshold-independent metrics.

This paper is structured into four sections: section 1 develops a conceptual framework for SDM performance metrics; section 2 reviews the application of performance metrics in 100 recent SDM studies; section 3 explores differences in the predictions favoured by each type of performance metric; and section 4 considers the consequences of performance metric choice in practical applications of SDMs. We conclude by discussing the circumstances under which each type of performance metric should be employed.

Conceptual framework

CORE CONCEPTS

We first develop a conceptual framework for SDM performance metrics, drawing on work from Murphy & Winkler (1987); introduced to ecology by Pearce & Ferrier 2000). Presence–absence models are based on observations of species presence or absence ($y \in 1{,}0$) at spatial locations (grid cells, study sites, etc.; hereafter 'cells') $\boldsymbol{i} = 1,\ldots,n$ (Fig. 1, Fig. 2a). Each presence–absence observation y_{i} is associated with environmental conditions x_{i} (Fig. 1, Fig. 2a). SDMs use species' associations with environmental variables (the relationship between **y** and **x**) to parameterize a function $f(x)$ which describes the predicted probability of an individual of the species being present in each cell $\mathbf{p} = p_1,\ldots,p_n$ (Fig. 1, Fig. 2a; notation summarized in Table 1).

Performance metrics measure the extent to which the predictions **p** provide information about the data **y**, using the prediction-observation pairs or joint distribution (**p**,**y**) (Fig. 2b,c; Murphy & Winkler 1987). Viewing this joint distribution in different ways reveals different information about the predictions, data and the relationships between them. The marginal distribution Pr(**y**) shows the proportion of presences and absences in the data set, indicating the prevalence or spatially averaged probability of occurrence $\bar{y} = \Pr(y = 1)$ (Fig. 2d; Royle *et al.* 2012); Pr(**p**) shows how often each prediction was made (Fig. 2e,f; Murphy & Winkler 1987), indicating the predicted prevalence $\bar{p} = \Pr(p = 1)$. The joint distribution $\Pr(\mathbf{p}, \mathbf{y})$ can be used to assess the numerical match between the predicted and observed probabilities of occurrence, known as *calibration*. Calibration measures the extent to which the predictions can be taken at 'face value'; for example, if a species is present in 60 out of 100 cells in a given environment, then a perfectly calibrated model will predict a probability of presence of 60% in that environment (Murphy & Winkler 1987; Vaughan & Ormerod 2005).

Calibration can be contrasted with *discrimination ability*, which measures the ability to tell presences and absences apart based on model predictions (Pearce & Ferrier 2000; Vaughan & Ormerod 2005; Wilks 2011). Discrimination assessment is based on the conditional distributions (**p**|**y**), which show the predictions for occupied cells $\Pr(\mathbf{p}|\mathbf{y} = 1)$ and the predictions for unoccupied cells $\Pr(\mathbf{p}|\mathbf{y} = 0)$ (Fig. 2g,f). In a discriminating model, higher prediction values are associated with presences and lower values with absences: the distributions of $\Pr(\mathbf{p}|\mathbf{y} = 1)$ and $\Pr(\mathbf{p}|\mathbf{y} = 0)$ show little overlap (Fig. 2g,f). Crucially, when assessing the discrimination ability, the absolute values of the predictions are unimportant, so a discriminating model may be poorly calibrated (Pearce & Ferrier 2000; see 'Consequences of performance metric choice' for examples).

Table 1. Glossary of notation and technical terms used in this paper

Symbol/Name	Description
i	Index indicating grid cell (location)
n	Number of cells in data set
x	Vector of values of an environmental variable determining species presence
p	Vector of model-predicted probabilities of presence
y	Binary vector indicating the observed presence (1) or absence (0) of species
$\bar{p}$	Mean predicted probability of presence for data set
$\bar{y}$	Prevalence, representing the proportion of cells that were presences (or the chance that any given cell is occupied); also known as base rate
τ	Threshold for conversion of continuous predictions into binary predictions
$\hat{\mathbf{p}}$	Vector of binary predictions of species presence (1) or absence (0)
a	Number of presences predicted as present
b	Number of absences predicted as present
c	Number of presences predicted as absent
d	Number of absences predicted as absent
ψ	'True' underlying probability of presence
r	Proportion of cells protected
m	Number of protected cells
T	Target number of protected populations
Calibration	Numerical match between **p** (or $\hat{\mathbf{p}}$) and **y**
Discrimination	Ability to separate presences ($\mathbf{y} = 1$) and absences ($\mathbf{y} = 0$) based on **p** (or $\hat{\mathbf{p}}$)
Continuous	Refers to probabilistic predictions $0 \le p \le 1$, or metrics that utilize them; used in contrast to binary
Binary	Refers to 'present or absent' predictions $\hat{p} \in (0, 1)$, or metrics that require them
Confusion matrix	Contingency table showing a, b, c and d and/or their relative frequencies

Our first classification axis defines whether the performance metric measures calibration or discrimination ability (moving down Figure 2 and Table 2); the second axis defines whether it requires binary predictions $\hat{\mathbf{p}}$ (moving across Figure 2 and Table 2). Binary predictions $\hat{\mathbf{p}}$ are obtained by applying a threshold τ to the predicted probabilities of presence **p**:

$$\hat{p}_i = \begin{cases} 1, & \textit{if } p_{\mathrm{i}} \ge \tau \\ 0, & \text{otherwise} \end{cases}$$

With binary predictions, a confusion matrix (contingency table) can be assembled, representing the joint distribution of prediction-observation pairs $(\hat{\mathbf{p}}, \mathbf{y})$ (Fig. 2c), from which all 'threshold-dependent' metrics are calculated (Liu *et al.* 2005; Peterson *et al.* 2011).

It is usually assumed that binary metrics measure discrimination but not calibration, because they will always be perfectly calibrated ($\hat{p}_i = y_i$) or entirely miscalibrated ($|\hat{p}_i - y_i| = 1$) for any given observation (Pearce & Ferrier 2000; Vaughan & Ormerod 2005; Liu, White & Newell 2011). However, this view neglects the fact that, as with continuous predictions, the calibration of binary metrics should be defined over many cells and varies continuously

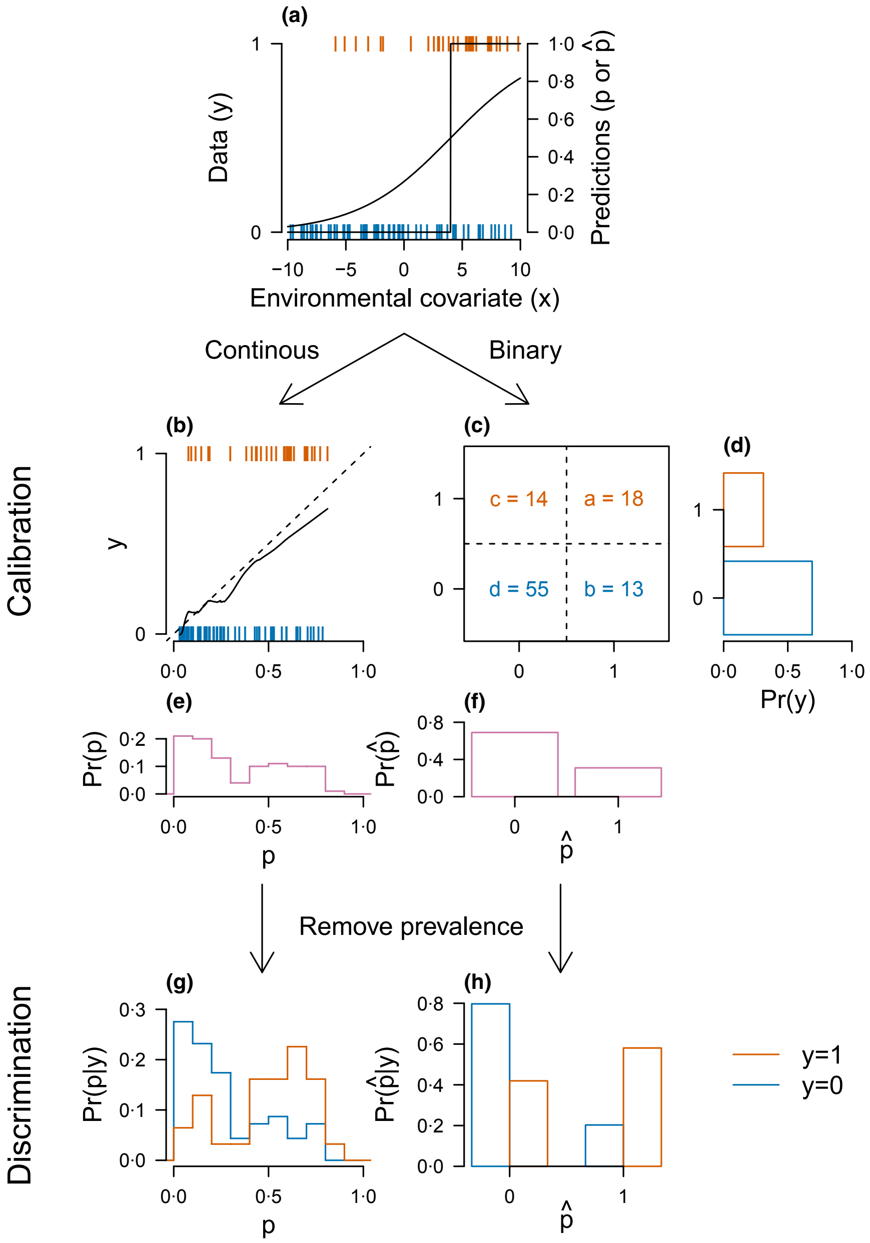

Fig. 2. A conceptual framework for performance metrics. A data set ***y*** of presences (red) and absences (blue), collected over different environments **x**, is compared to predicted probabilities of presence **p** or binary presence–absence predictions $\hat{\mathbf{p}}$ (a). The calibration (numerical accuracy) of the predictions can be assessed using the joint distribution of predictions and observations (b, c; solid line indicates a smoothed relationship between predictions and observations, dotted line indicates a 1:1 relationship representing 'perfect' calibration). The marginal distributions show how frequently each prediction was made (e, f) and how frequently the species is present overall (prevalence; d). Factoring out prevalence from the joint distribution gives the conditional distributions of predictions for presence and absence observations (g, h), which can be compared to assess the discrimination ability of the predictions, but not their calibration.

with the true occurrence probability, denoted as ψ; for example, if $\hat{p} = 1$, higher calibration will be achieved if ψ = 0·9 than if ψ = 0·5. As we will demonstrate later, this distinction has important consequences, because it means that two binary models may have equal discrimination ability, but differ in calibration.

The above classification scheme identifies four types of SDM performance metric (continuous calibration, continuous discrimination, binary calibration and binary discrimination metrics; Fig. 2, Table 2). The following sections explore the connections between these categories.

THRESHOLD-DEPENDENT METRICS DO NOT REQUIRE THRESHOLD TRANSFORMATIONS

Binary and continuous metrics have been treated as separate and incomparable ways of measuring SDM performance (Liu, White & Newell 2011; Peterson *et al.* 2011). However, we show

Table 2. Classification scheme for performance metrics with examples. See Appendix S2 for definitions of specific performance metrics.

	Continuous	Binary
Calibration	Mean squared error (MSE) Root mean squared error (RMSE) Likelihood or deviance Akaike's Information Criterion (AIC) Bayesian information criterion (BIC) *R*-squared (standard or likelihood-based)	Mean accuracy Cohen's kappa
Discrimination	Area under the curve (AUC) Pearson's correlation Spearman's correlation	Sensitivity Specificity True Skill Statistic (TSS)

that by rewriting the confusion matrix (Table 3a) as a probabilistic function of the continuous predictions **p** and observations **y** (Table 3b), binary metrics can be calculated using probabilistic predictions $0 < \mathbf{p} < 1$. For the special case in which predictions are binary, the usual (whole-number) confusion matrix results, but when predictions are probabilistic, each prediction-observation pair is 'split' between two cells of the confusion matrix, with a presence predicted with probability p_{i} and absence predicted with probability $1-p_{\mathrm{i}}$.

The probabilistic confusion matrix brings several advantages: it makes the threshold selection process redundant; it enables a direct comparison of binary and continuous metrics, using them to estimate the same parameters; and it conceptually unifies binary and continuous metrics, showing, for example, that overall accuracy (a binary metric) and mean absolute prediction error (formerly a continuous metric; Liu, White & Newell 2011) are actually equivalent measures of performance (Appendix S1 in Supporting Information). Within this new framework, it can be shown that any metric that can be formulated as a confusion matrix is always maximized by binary (rather than probabilistic) predictions (Appendix S1), such that even when the true probability of species presence ψ lies between 0 and 1, binary predictions ($p \in 0,1$) are necessary to achieve high performance (i.e. binary metrics are 'improper' performance measures; Wilks 2011). In short, binary metrics can be used to evaluate the performance of probabilistic predictions, but they always favour binary predictions, with important impacts on which SDMs are selected for prediction (discussed in 'Consequences of performance metric choice').

PREVALENCE-INDEPENDENT METRICS CANNOT MEASURE CALIBRATION

We now show that prevalence-dependence defines the difference between calibration and discrimination metrics (Fig. 2). The consequences of mathematically factoring out prevalence $\Pr(\mathbf{y} = 1)$ are implicit in Murphy and Winkler's *likelihood-base rate factorisation* (1987):

$$\Pr(\mathbf{p}, \mathbf{y}) = \Pr(\mathbf{p}|\mathbf{y})\Pr(\mathbf{y})$$

This equation shows that factoring out prevalence $\Pr(\mathbf{y}{=}1)$ from the joint distribution $\Pr(\mathbf{p},\mathbf{y})$ leaves the conditional distributions $\Pr(\mathbf{p}|\mathbf{y})$, which are used to calculate discrimination ability (Fig. 2g; Murphy & Winkler 1987; Pearce & Ferrier 2000; Wilks 2011). It also shows that once prevalence is factored out, the joint distribution $\Pr(\mathbf{p},\mathbf{y})$ cannot be recovered, so the numerical match between **p** and **y** (calibration; Fig. 2b) cannot be assessed (note that some prevalence-independent metrics, such as Pearson's correlation, permit the 'calibration' of relative occurrence probabilities to be assessed, but that knowledge of prevalence is required to assess true calibration; Phillips & Elith 2010, 2013). Thus, calibration metrics

Table 3. Traditional (a) and probabilistic (b) confusion matrices. The probabilistic confusion matrix allows 'threshold-dependent' metrics to be used with probabilistic predictions (**p**) and reverts to the traditional confusion matrix when binary predictions ($\hat{\mathbf{p}}$) are used. This removes the need to select thresholds and unifies continuous and binary performance metrics

		Observed (**y**)		
		Present (**y** = 1)	Absent (**y** = 0)	
(a) Traditional confusion matrix				
Predicted ($\hat{\mathbf{p}}$)	Present ($\hat{\mathbf{p}}=1$)	True presence (*a*)	False presence (*b*)	Mean prediction $= \frac{a+b}{n}$
	Absent ($\hat{\mathbf{p}}=0$)	False absence (*c*)	True absence (*d*)	$n = a + b + c + d$
		Prevalence $= \frac{a+c}{n}$		
(b) Probabilistic confusion matrix				
Predicted (**p**)	Present (**p** = 1)	$a = \sum_{i=1}^{n} \Pr(p_i = 1 \cap y_i = 1)$	$b = \sum_{i=1}^{n} \Pr(p_i = 1 \cap y_i = 0)$	
		$\sum_{i=1}^{n} p_i y_i$	$= \sum_{i=1}^{n} p_i(1 - y_i)$	
	Absent (**p** = 0)	$c = \sum_{i=1}^{n} \Pr(p_i = 0 \cap y_i = 1)$	$d = \sum_{i=1}^{n} \Pr(p_i = 0 \cap y_i = 0)$	
		$= \sum_{i=1}^{n} (1 - p_i)y_i$	$= \sum_{i=i}^{n} (1 - p_i)(1 - y_i)$	

are dependent on prevalence, and prevalence-independent metrics cannot measure calibration.

As with continuous metrics, binary calibration metrics are prevalence-dependent. The confusion matrix $\Pr(\hat{\mathbf{p}}, \mathbf{y})$ is also subject to Murphy and Winkler's likelihood-base rate factorisation:

$$\Pr(\hat{\mathbf{p}}, \mathbf{y}) = \Pr(\hat{\mathbf{p}}|\mathbf{y})\Pr(\mathbf{y})$$

Again, the values of binary metrics based on the conditional distributions $\Pr(\hat{\mathbf{p}}|\mathbf{y})$ will be independent of prevalence $\Pr(\mathbf{y})$. This includes *sensitivity* $\Pr(\hat{\mathbf{p}} = 1|\mathbf{y} = 1)$, which measures the chance of making a presence prediction given that the species was present (left-most bar, Fig. 2h), *specificity* $\Pr(\hat{\mathbf{p}} = 0|\mathbf{y} = 0)$, which measures the chance of making an absence prediction given that the species was absent (right-most bar, Fig. 2h), and the true skill statistic (TSS), which is calculated by adding sensitivity and specificity together and subtracting 1 (Appendix S2; Allouche, Tsoar & Kadmon 2006). There exists a close relationship between these prevalence-independent binary metrics and continuous prevalence-independent metrics such as the area under the curve (AUC; Fielding & Bell 1997), which we confirm by showing that $TSS = 2(AUC{-}0{\cdot}5)$ when predictions are binary (Appendix S3). In summary, factoring out prevalence precludes calibration assessment, whether continuous or binary metrics are used.

Literature review

To assess current trends in SDM performance assessment, we systematically reviewed 100 SDM studies that employed single-number measures of performance (Appendix S4). We reviewed both presence–absence studies and presence-only studies that used presence–absence performance metrics (based on simulated 'pseudo-absences' or 'background points'), but the patterns that we report are similar when only presence–absence studies (n = 43) are considered (Appendix S4). We classified each performance metric into one of the four categories identified by our framework; the most common metrics are categorized in Table 2, and definitions and formulae for specific metrics are given in Appendix S2.

Across all studies, continuous discrimination metrics were more widely used than calibration metrics (Fig. 3a), largely due to the widespread application of the area under the curve (AUC; Appendix S2; see also Yackulic *et al.* 2012). Two compatible explanations are as follows: (i) the wide encouragement of prevalence-independent (discrimination) performance metrics in the SDM literature and (ii) that discrimination metrics are provided by SDM software packages; for example, MaxEnt uses AUC (Yackulic *et al.* 2012), and studies using MaxEnt were more likely to use continuous discrimination metrics (94%) than those that did not (74%). Calibration metrics were preferred for assessing which environmental variables contributed most to variation in occupancy, but discrimination metrics were preferred for selecting among competing models and quantifying the predictive performance of a given model (Fig. 3b). Moreover, of the 74 studies that relied on the absolute values of SDM predictions (such as those that estimated occupancy), 64% did not assess calibration, relying solely on discrimination metrics. These observations suggest a widespread belief that discrimination metrics offer an advantage when quantitative predictions are required, even though the numerical accuracy of predictions can only be assessed by a calibration metric.

Binary metrics were frequently used to assess SDM performance (50% of studies; Fig. 3a). However, only one study was restricted to binary predictions by software outputs, whilst only two studies required binary predictions for use with conservation algorithms or applications. Thus, the widespread use of binary metrics may be due to perceived correct practice for performance assessment, rather than practical necessity.

Consequences of performance metric choice

In this section, we illustrate how the choice of performance metric type influences the predictions of SDMs. Motivated by our literature review findings, we aim to highlight the circumstances under which (i) binary metrics will carry drawbacks and (ii) prevalence-dependent (calibration) metrics should be employed.

We base our discussion around a scenario in which there are two environments with true occurrence probabilities ψ_1 and ψ_2, of which estimates p_1 and p_2 are to be made. For illustration, we chose values $\psi_1 = 0{\cdot}6$ and $\psi_2 = 0{\cdot}8$ and assumed that the two environments were equally common. Figure 4 shows the scores allocated to all combinations of p_1 and p_2 by a performance metric from each category: log-likelihood for continuous calibration; AUC for continuous discrimination; mean overall accuracy (hereafter 'accuracy') for binary calibration; and TSS for binary discrimination (scores are calculated analytically, assuming a large sample size).

THE DISADVANTAGES OF BINARY METRICS

We start by comparing continuous and binary metrics. Figure 4 shows that the binary metrics favour binary predictions: both accuracy and TSS are maximized by binary predictions $\hat{p} \in (0, 1)$. In contrast, the log-likelihood is maximized by the underlying values of p_1 = 0·8 and p_2 = 0·6, and AUC does not favour either continuous or binary predictions (explanation given below).

Favouring binary predictions in this way can present considerable disadvantages due to the information lost when probabilistic predictions are dichotomized (Hand 1997; Fielding 2007). First, binary predictions have reduced discrimination ability; for instance, in Figure 4, accuracy results in predictions of $p_1 = p_2 = 1$ in both environments and as such does not indicate that environment 1 is more likely to contain the species. In our simplified example, the TSS-selected model maintains discrimination ability because it makes different predictions in the two environments (p_1 = 1 and p_2 = 0). However, binary predictions will always have lower maximum discrimination than continuous predictions when there are more than two different true probabilities of occurrence ψ (proved graphically in Appendix S5), because binary predictions are restricted to two

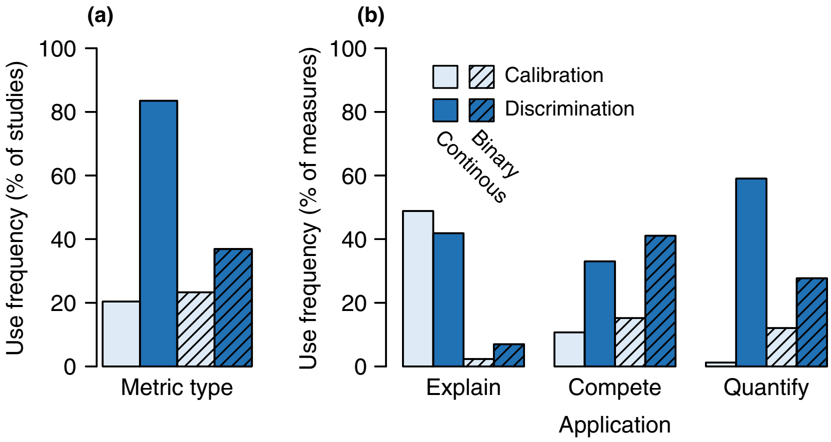

Fig. 3. Frequency of use of different performance metric types, based on a literature survey of 100 SDM studies. Panels show overall frequency of use per study (a) and which metrics were used most frequently for particular applications (b; Explain = to test hypotheses on environmental determinants of species presence; Compete = to compare the predictive ability of different models; Quantify = to summarize the performance of a single model; details in Appendix S2).

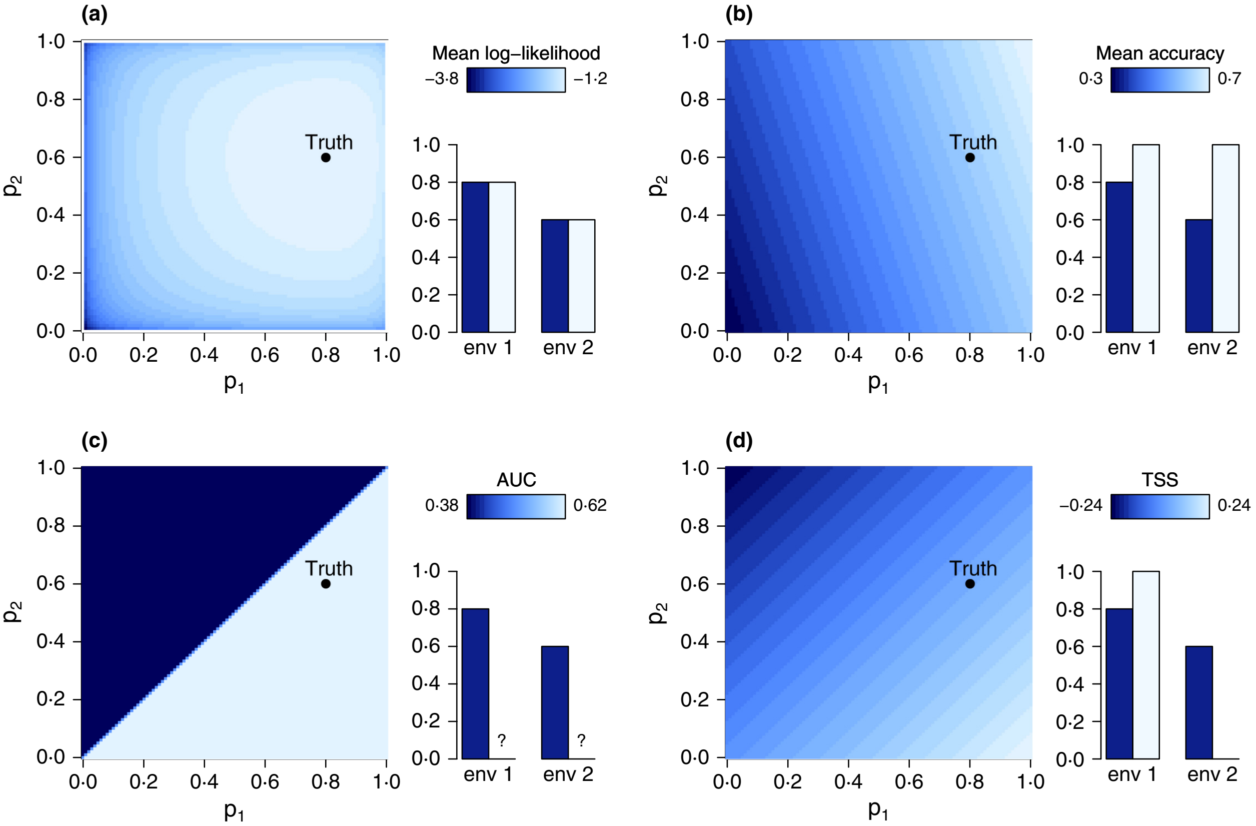

Fig. 4. Predictions favoured by different types of performance metrics in a simplified modelling scenario. (a) Mean log-likelihood (continuous calibration); (b) mean accuracy (binary calibration); (c) AUC (continuous discrimination); (d) TSS (binary discrimination). Large squares show the performance of all predicted probability of presence combinations for environment 1 (p_1) and environment 2 (p_2), with the true combination $\psi_1 = 0{\cdot}8$, $\psi_2 = 0{\cdot}6$, indicated by dots. Bars display the true occurrence probabilities (dark blue) and highest-performing predictions (light blue).

ranks ($\hat{\mathbf{p}} \in 0, 1$). Thus, in all cases in which there are more than two occurrence probabilities (e.g. continuous gradients of habitat suitability such as that presented in Fig. 1), choosing a binary performance metric will reduce the discrimination ability of SDM predictions.

The other disadvantages of binary predictions relate to their reduced calibration when true occurrence probabilities lie between 0 and 1 ($0 < \psi < 1$). First, the proportion of occupied cells in a given environment will be incorrect; for example, in Figure 4b and 4d, all cells with environment 1 are predicted to be occupied, but in reality, 20% will be unoccupied. Secondly, predictions for any single cell will be made with false certainty, because the variance of the Bernoulli distribution, which forms the stochastic part of a presence–absence models, reaches zero for $p = 1$ or $p = 0$ (Appendix S6; Bolker 2008); that is, the possibility of a 'surprise' (e.g. an absence in a cell with environment 1 in Figure 4) is disregarded. In summary, models developed using binary performance metrics will often have both reduced discrimination and calibration.

PREVALENCE AND CALIBRATION

We next examine the consequences of using prevalence-dependent or prevalence-independent performance metrics. In Figure 4, the continuous discrimination metric, AUC, correctly identifies that environment 1 has a higher probability of containing the species than environment 2 (models in which $p_1 > p_2$ receive higher AUC scores). However, because AUC is prevalence-independent, it is indifferent to the absolute values of p_1 and p_2; in contrast, the log-likelihood is prevalence-dependent and maximized by the correct combination of $p_1 = 0{\cdot}8$ and $p_2 = 0{\cdot}6$. The binary discrimination metric, TSS, favours the combination of binary predictions which best separate presences from absences and is maximized by allocating $p_1 = 1$ to the environment in which presences are most common (environment 1) and $p_2 = 0$ to the environment in which presences are least common (environment 2); in contrast, the binary calibration metric, accuracy, recognizes that the species is more likely to be present than absent in both environments ($p_1 = p_2 = 1$). This simple SDM parameterization example demonstrates that only prevalence-dependent metrics will assess the calibration of predictions.

If two competing SDMs have the same discrimination ability but make different predictions, prevalence dictates which of the two models is better-calibrated. Figure 5 illustrates a case in which one model predicts a low prevalence and thus is more likely to make a correct prediction given that the species is absent (has higher specificity), and another model predicts a high prevalence and thus is more likely to make a correct prediction when the species is present (has higher sensitivity). Both models have equal discrimination ability (equal TSS and AUC values), but the higher the probability that any given cell contains a presence, the better-calibrated the high-prevalence (high-sensitivity) model will be compared to the low-prevalence (high-specificity) model (Figure 5). This shows how ignoring prevalence during performance assessment masks differences in calibration (an example of the base rate fallacy; Gigerenzer 2003; Bolker 2008).

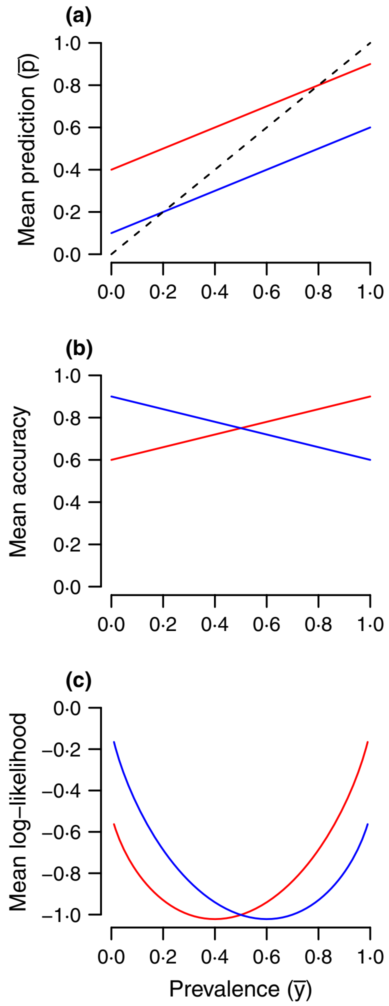

Fig. 5. The role of prevalence in model calibration. The calibration of two models with equal discrimination ability (TSS = 0·5; AUC = 0·75) but different prevalence predictions (red: higher prevalence and higher sensitivity; blue: lower prevalence and higher specificity) is compared. Panels show how prevalence $\bar{y}$ alters: (a) the predicted prevalence $\bar{p}$ in each model, with the true prevalence shown by dashed line, (b) the mean overall accuracy; and (c) the mean log-likelihood. As prevalence increases, the high-prevalence model becomes better-calibrated than the low-prevalence model.

The prevalence predicted by an SDM ($\bar{p}$), and thus its calibration, depends not only on the observed prevalence $\bar{y}$, but also on the relative value or costs of presence and absence predictions (Hand 1997; Fielding 2007). It is widely believed that biased predictions can be avoided by adopting different costs depending on the observed prevalence (Jiménez-Valverde & Lobo 2006; Lobo, Jiménez-Valverde & Real 2008), but an explicit consideration of costs shows why this is incorrect. If correctly predicting presences is more valuable than correctly predicting absences (or the costs of false absences are higher than the costs of false presences), it is better to classify a greater proportion of cells as present instead of absent (Hand 1997; Fielding 2007). This relationship between error costs and predicted prevalence is reflected in the common practice of selecting classification thresholds τ using receiver operating characteristic (ROC) plots: higher costs for false absences encourage the selection of lower thresholds and thus predictions with higher prevalence $\bar{p}$ (Hand 1997; Fielding 2007). Generalising these ideas, a calibration metric in which the value of presences is V_1 and the value of absences is V_0 will be maximised by a predicted prevalence of $\bar{p} = \frac{V_1\bar{y}}{V_1\bar{y}+V_0(1-\bar{y})}$. This means that unbiased (perfectly calibrated) predictions ($\bar{p} = \bar{y}$) will only outperform biased predictions ($\bar{p} \neq \bar{y}$) if presences and absences have equal value ($V_1 = V_0$); adopting unequal costs ($V_1 \neq V_0$), will favour predictions that over-estimate prevalence if ($V_1 > V_0$), or that under-estimate prevalence if ($V_1 < V_0$). The key distinction is that whilst the costs of errors in practical applications of SDM predictions may be unequal (for example, the cost of protecting an unoccupied cell may be less than leaving an occupied cell unprotected), during SDM performance evaluation it is necessary to assume balanced costs to obtain unbiased, well-calibrated presence–absence predictions.

Prevalence-dependent metrics have been criticized for making SDM performance difficult to compare across data sets (McPherson, Jetz & Rogers 2004; Vaughan & Ormerod 2005), but this assumes that the absolute values of performance metrics are comparable. In fact, any comparisons of performance across data sets will necessarily be subjective, because performance scores are conditional on the data; for example, altering the frequency of different environments alters predictive performance even if the SDM itself remains identical (Appendix S7). This limitation applies to all performance metrics and makes searching for an absolute grading for SDM performance a questionable goal (see also Wolpert & Macready 1997; Yackulic *et al.* 2012).

Performance metrics in predictive applications

A key motivation for SDM development is to facilitate practical decision-making by predicting species distributions in places or times for which occupancy data are unavailable (Fig. 1). Practical applications often require binary classification decisions, such as whether to allocate statutory protection, search for an endangered species, or conduct an invasive species eradication programme in a given area (Moilanen, Wilson & Possingham 2009; Peterson *et al.* 2011). The complexities of conservation decision-making frequently demand the use of sophisticated classification algorithms (e.g. machine learning approaches: Fielding 2007; specialized software such as Zonation: Moilanen, Wilson & Possingham 2009), but, as we discuss in this section, their success is dependent on the predictive performance of the SDM used.

By analogy with SDM performance metrics, it is possible to distinguish ‘calibration applications’, which require estimated occurrence probabilities p, and ‘discrimination applications’, which require only relative or ranked occurrence probabilities. Consider a prioritization task in which the objective is to protect as many occupied cells as possible, based on estimated occurrence probabilities p and resource limitations on the proportion of cells that can be protected r. If all cells have equal protection costs, efficient prioritization depends only on which environments (cells) the species is most likely to occur in; decision-making is not influenced by the absolute values of occurrence probabilities (two-environment example presented in Appendix S8). In this scenario, the optimal strategy will be governed by the discrimination ability, not the calibration ability, of the predictions, and prevalence-independent discrimination metrics such as AUC (Fielding & Bell 1997) will provide adequate performance assessment. Contrast to this prioritization task with the situation in which the predictions are used to determine the minimum number of cells $m = rn$ that must be protected to attain a target number T of protected populations. If there are n cells with probability of presence p, the expected number of occupied protected cells is mp, requiring the protection of $m_T = \frac{T}{p}$ cells to achieve the target. The number of cells that must be protected m_T is critically dependent on the probability of presence p: if a smaller proportion of cells are occupied (p is lower), more cells must be protected (m_T must be higher; Appendix S8). Therefore, in this scenario, numerically accurate estimates of p, and thus calibration assessment, are essential to develop an efficient solution.

The application of SDM predictions to classification problems has been frequently conflated with the assessment of the predictive performance of SDMs, generating a widespread belief that applications that require binary classifications also require binary predictions (e.g. Liu *et al.* 2005; Allouche, Tsoar & Kadmon 2006; Lobo, Jiménez-Valverde & Real 2008; Bean, Stafford & Brashares 2011). However, clearly distinguishing between the development of SDM predictions and their subsequent application to classification problems (Fig. 1) reveals a fundamental problem with using binary predictions in classification applications: the information available on which to base decisions is often reduced (see ‘The disadvantages of binary metrics’). When dichotomization results in equal absolute values of the predictions (e.g. $\hat{p}_1 = \hat{p}_2 = 1$ in Fig. 4b), discrimination among presences and absences, and thus prioritization of different cells, is precluded. When dichotomization results in reduced calibration ($\hat{p} \neq \psi$), the success of a given conservation action (e.g. protection) will be systematically under- or overestimated, and the uncertainty of success will be underestimated (Appendix S8). The implication is that it is better to use continuous than binary performance metrics to develop SDM predictions for use in binary classification applications.

Discussion and recommendations

We have developed a conceptual framework for classifying performance metrics and identifying the circumstances under which they should be employed (identified as a key target for SDM research by Elith & Leathwick 2009). We applied this framework to explore the theoretical and practical implications of using prevalence-dependent and binary (‘threshold-dependent’) metrics to evaluate SDM performance.

Although prevalence-independence has been regarded as a desirable property of performance metrics (Vaughan & Ormerod 2005; Allouche, Tsoar & Kadmon 2006; Lobo, Jiménez-Valverde & Real 2008), we showed that consideration of prevalence is essential to measure model calibration (the ability to correctly predict the number of occupied cells) and thus to develop unbiased probability of presence predictions. Conversely, all prevalence-independent metrics, including both continuous metrics such as AUC and binary metrics such as TSS, are limited to measuring discrimination ability (the ability to distinguish presences above absences) and cannot measure calibration; this finding complements recent work showing that prevalence-independent SDM parameterization methods cannot estimate absolute occurrence probabilities except under highly restrictive assumptions (Phillips & Elith 2013). Obtaining accurate prevalence estimates can be challenging, particularly for species with highly dynamic distributions (Fielding 2007) and/or low detection probabilities (Yackulic *et al.* 2012), and is unnecessary for some applications of SDMs; for instance, prevalence-independent evaluation of discrimination ability is sufficient when SDM predictions are applied to qualitative prioritization problems in which occupancy estimates are superfluous. Nonetheless, accurate estimates of probability of presence are required for many quantitative applications of SDMs, such as estimating areas of occupancy, or, as in our example, designing conservation strategies to protect a minimum number of populations; in all such cases, prevalence-dependent assessment of SDM calibration is crucial.

Binary ('threshold-dependent') metrics are as widely used to evaluate SDM performance, often on the grounds that binary predictions are necessary for conservation planning (Liu *et al.* 2005; Allouche, Tsoar & Kadmon 2006; Lobo, Jiménez-Valverde & Real 2008; Bean, Stafford & Brashares 2011), but this conflates the development of SDM predictions with their application to practical problems. We showed that binary metrics can be calculated directly from probabilistic predictions without applying threshold transformations, but always favour binary predictions over probabilistic predictions. The consequent loss of information (Hand 1997; Fielding 2007) means that using binary metrics instead of continuous metrics to parameterize or select among SDMs results in predictions with reduced calibration and misleading uncertainty when occurrence probabilities lie between zero and one, and reduced discrimination ability when there are more than two different occurrence probabilities (e.g. continuous gradients of habitat suitability). Under this broad set of circumstances, dichotomizing predictions will reduce the efficiency of any applications of SDMs, and given that binary predictions are rarely required to develop binary classification decisions (Fielding 2007; Moilanen, Wilson & Possingham 2009), we urge that continuous metrics be used in place of binary metrics wherever possible.

Combined, our findings encourage the wider uptake of continuous calibration metrics in SDM assessment, for which we recommend likelihood functions (Bolker 2008). In Appendix S9, we show that given predictions and a test data set, it is straightforward to calculate log-likelihoods, information criteria with model complexity penalties such as Akaike's Information Criterion (AIC), and likelihood-based *R*-squared values (see also Nakagawa & Schielzeth 2012). A recent simulation study (Warren & Seifert 2011) found that AIC was superior to AUC at uncovering the relationships underlying species distributions – a result that supports our emphasis on the importance of calibration assessment (AIC measures calibration, but AUC does not). The strong body of statistical theory on likelihood functions also serves as an excellent basis for new developments in SDM performance assessment, such as relaxing the assumption of spatial independence between grid cells (see Bennie *et al.* 2013 for an example).

It is, however, essential to be aware of the limitations of SDM performance metrics. First, they are designed to compare predictions conditional on given test data set (Bolker 2008; Nakagawa & Schielzeth 2012), and interpretation of their absolute values is necessarily subjective (Wolpert & Macready 1997; Yackulic *et al.* 2012). The wish to compare model performance among species or regions (as occurred in 31% of studies we reviewed) has fuelled a search for a 'holy grail' performance metric that is independent of data properties, but trade-offs always exist between eliminating those properties and preserving the information required to test predictions, as we have shown for prevalence. In short, no metric provides a universal grading for SDM performance. Secondly, graphical quantities can convey more information about performance than scalar metrics, and we encourage the wider uptake of calibration plots (used in just 2% studies we reviewed; Fig. 2b, Pearce & Ferrier 2000; Wilks 2011), and plots of predictions against data in environmental space (6% of studies; Figs 1a and 2b; Wilks 2011; Yackulic *et al.* 2012), preferably together with confidence intervals on predictions (Bolker 2008). These techniques, coupled with the wider application of continuous calibration metrics, will lead to improved understanding and prediction from species distribution models.

Acknowledgements

We thank Alan Fielding, Dave Hodgson and Justin Travis for advice that greatly improved the content and clarity of the manuscript, and Nick Golding for useful discussion of the role of prevalence in species distribution modelling. CRL was funded by a CCSF grant from the University of Exeter.

References

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

Bean, W.T., Stafford, R. & Brashares, J.S. (2011) The effects of small sample size and sample bias on threshold selection and accuracy assessment of species distribution models. *Ecography*, **35**, 250–258.

Bennie, J., Hodgson, J.A., Lawson, C.R., Holloway, C.T.R., Roy, D.B., Brereton, T., Thomas, C.D. & Wilson, R.J. (2013) Range expansion through fragmented landscapes under a variable climate. *Ecology Letters*, **16**, 921–929 in press.

Bolker, B.M. (2008) *Ecological Models and Data in R*. Princeton University Press, Princeton, NJ.

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

Fielding, A. (2007) *Cluster and Classification Techniques for the Biosciences*. Cambridge University Press, Cambridge.

Fielding, A.H. & Bell, J.F. (1997) A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environmental conservation*, **24**, 38–49.

Foody, G.M. (2011) Impacts of imperfect reference data on the apparent accuracy of species presence–absence models and their predictions. *Global Ecology and Biogeography*, **20**, 498–508.

Gigerenzer, G. (2003) *Reckoning with Risk: Learning to Live With Uncertainty*. Penguin, London.

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

Hand, D.J. (1997) *Construction and Assessment of Classification Rules*. Wiley, Chichester.

Hirzel, A.H., Le Lay, G., Helfer, V., Randin, C. & Guisan, A. (2006) Evaluating the ability of habitat suitability models to predict species presences. *Ecological Modelling*, **199**, 142–152.

Jiménez-Valverde, A. & Lobo, J. (2006) The ghost of unbalanced species distribution data in geographical model predictions. *Diversity and Distributions*, **12**, 521–524.

Kunin, W.E. (1998) Extrapolating species abundance across spatial scales. *Science*, **281**, 1513–1515.

Liu, C., White, M. & Newell, G. (2011) Measuring and comparing the accuracy of species distribution models with presence–absence data. *Ecography*, **34**, 232–243.

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

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

McPherson, J., Jetz, W. & Rogers, D.J. (2004) The effects of species' range sizes on the accuracy of distribution models: ecological phenomenon or statistical artefact? *Journal of Applied Ecology*, **41**, 811–823.

Moilanen, A., Wilson, K.A. & Possingham, H.P. (2009) *Spatial Conservation Prioritization: Quantitative Methods and Computational Tools*. Oxford University Press, Oxford.

Murphy, A. & Winkler, R. (1987) A general framework for forecast verification. *Monthly Weather Review*, **115**, 1330–1338.

Nakagawa, S. & Schielzeth, H. (2012) A general and simple method for obtaining R2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution*, **4**, 133–142.

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

Peterson, A.T., Soberón, J., Pearson, R.G., Anderson, R.P., Martínez-Meyer, E., Nakamura, M. & Araújo, M.B. (2011) *Ecological Niches and Geographic Distributions*. Princeton University Press, Oxford.

Phillips, S.J. & Elith, J. (2010) POC plots: calibrating species distribution models with presence-only data. *Ecology*, **91**, 2476–2484.

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

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.

Vaughan, I. & Ormerod, S. (2005) The continuing challenges of testing species distribution models. *Journal of Applied Ecology*, **42**, 720–730.

Warren, D.L. & Seifert, S.N. (2011) Ecological niche modeling in Maxent: the importance of model complexity and the performance of model selection criteria. *Ecological Applications*, **21**, 335–342.

Wilks, D.S. (2011) *Statistical Methods in the Atmospheric Sciences*, 3rd edn. Elsevier/Academic Press, London.

Wolpert, D.H. & Macready, W.G. (1997) No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, **1**, 67–82.

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

Received 27 July 2013; accepted 30 September 2013
Handling Editor: Robert Freckleton

Supporting Information

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

Data S1. Results from literature review of trends in SDM performance assessment.

Appendix S1. Proof that binary performance measures favour binary predictions.

Appendix S2. Description of performance metrics.

Appendix S3. Proof that AUC is equivalent to TSS for binary models.

Appendix S4. Literature review methods.

Appendix S5. Proof that binary predictions frequently have reduced discrimination ability.

Appendix S6. Stochasticity in presence-absence models.

Appendix S7. Effects on the frequency of different environments on performance.

Appendix S8. Further details on conservation examples.

Appendix S9. Demonstration of likelihood performance metrics.