Journal of Biogeography (J. Biogeogr.) (2013)

ORIGINAL ARTICLE

# Selecting thresholds for the prediction of species occurrence with presence-only data

Canran Liu*, Matt White and Graeme Newell

*Arthur Rylah Institute for Environmental Research, Department of Sustainability and Environment, Heidelberg, Victoria 3084, Australia*

*Correspondence: Canran Liu, Arthur Rylah Institute for Environmental Research, Department of Sustainability and Environment, 123 Brown Street, Heidelberg, Victoria 3084, Australia.
E-mail: canran.liu@dse.vic.gov.au

## ABSTRACT

**Aim** Species distribution models have been widely used to tackle ecological, evolutionary and conservation problems. Most species distribution modelling techniques produce continuous suitability predictions, but many real applications (e.g. reserve design, species invasion and climate change impact assessment) and model evaluations require binary outputs, and thresholds are needed for these transformations. Although there are many threshold selection methods for presence/absence data, it is unclear whether these are suitable for presence-only data. In this paper, we investigate mathematically and empirically which of the existing threshold selection methods can be used confidently with presence-only data.

**Location** We used real spatially explicit environmental data derived from the western part of the state of Victoria, south-eastern Australia, and simulated species distributions within this area.

**Methods** Thirteen existing threshold selection methods were investigated mathematically to see whether the same threshold can be produced using either presence/absence data or presence-only data. We further adopted a simulation approach, created many virtual species with differing prevalences in a real landscape in south-eastern Australia, generated data sets with different proportions of pseudo-absences, built eight types of models with four modelling techniques, and investigated the behaviours of four threshold selection methods in these situations.

**Results** Three threshold selection methods were not affected by pseudo-absences, including max *SSS* (which is based on maximizing the sum of sensitivity and specificity), the prevalence of model training data and the mean predicted value of a set of random points. Max *SSS* produced higher sensitivity in most cases and higher true skill statistic and *kappa* in many cases than the other methods. The other methods produced different thresholds from presence-only data to those determined from presence/absence data.

**Main conclusions** Max *SSS* is a promising method for threshold selection when only presence data are available.

**Keywords**
**Evaluation, lift curve, presence-only, ROC curve, sensitivity, species distribution model, specificity, threshold.**

## INTRODUCTION

Information concerning the distributions of species is fundamental to many ecological, evolutionary and conservation problems (Graham *et al.*, 2004; Guisan & Thuiller, 2005; Franklin, 2009). Numerous modelling techniques have been used to predict species distributions (e.g. Segurado & Araújo, 2004; Elith *et al.*, 2006; Thuiller *et al.*, 2009). Most modelling techniques, particularly the newer, more powerful techniques (e.g. Maxent, boosted regression trees, random forest, artifi-




Journal of Biogeography

cial neural networks), produce continuous (or at least non-binary) predictions. Although continuous results convey more information than binary outputs (Vaughan & Ormerod, 2005) and are potentially useful for many conservation applications (e.g. Araújo *et al.*, 2002), binary outputs classified by the use of a model-specific threshold have become integrated into applications such as biodiversity assessments, reserve selection, and climate change impact assessments and investment programmes by government agencies (Lobo *et al.*, 2008; Rebelo & Jones, 2010). In addition, binary predictions are required to evaluate models when using accuracy measures derived from confusion matrices (e.g. Phillips *et al.*, 2006; Pearson *et al.*, 2007).

If both presence and absence data are available, there are many approaches for selecting a threshold or cut-off value to transform non-binary into binary predictions (Fielding & Bell, 1997; Liu *et al.*, 2005; Jiménez-Valverde & Lobo, 2007; Pearson, 2007; Freeman & Moisen, 2008; Nenzén & Araújo, 2011). For many species, however, reliable absence data are not available (Pearson *et al.*, 2007; Peterson *et al.*, 2008) and therefore conventional confusion matrices cannot be formed. In this situation, it is commonly accepted that the methods developed for presence/absence data are no longer applicable (Jiménez-Valverde & Lobo, 2007; Pearson *et al.*, 2007), and that selecting an appropriate threshold is problematic (Papeş & Gaubert, 2007; Rebelo & Jones, 2010).

It is often believed that when only presence data are available, threshold selection rules can only be based on those data. Phillips *et al.* (2006) used the minimum predicted value for the training sites as the threshold, which was termed 'lowest presence threshold' (LPT) by Pearson *et al.* (2007), but this is extremely sensitive to low sample sizes (Bean *et al.*, 2012). It can also be considered a special case of the fixed (or required) sensitivity method (Pearson *et al.*, 2004), where the required sensitivity is set at 100%. However, the fixed sensitivity method requires an arbitrary assignment of a level of sensitivity (e.g. 90%).

The sensitivity–specificity difference minimization method has been used with presence/pseudo-absence data (e.g. Chefaoui & Lobo, 2008), as would generally be used with presence/absence data. However, because the determination of 'specificity' is based on pseudo-absence rather than true absence data, this approach is unlikely to generate a meaningful threshold (as discussed in the next section). Pseudo-absences are the points that are taken as absences but may not all be true absences.

Braunisch & Suchant (2010) used two threshold selection methods in their study using presence-only data. One is provided within the software BIOMAPPER (Hirzel *et al.*, 2002), which is based on the continuous *P*/*E* curve, where *P* is the predicted frequency of evaluation points and *E* is the expected frequency of evaluation points (which is actually the predicted frequency of a sample of random points). Specifically, they assigned 'presence' to all the points with predicted suitability values larger than the value where *P*/*E*, including its 90% confidence interval, exceeds 1. This approach only guarantees that the transformed model is better than the random model. The other approach is provided within the MAXENT software package, and is based on maximizing the sum of sensitivity and specificity (max *SSS*). However, Braunisch & Suchant (2010) stated that 'without true absence data, specificity and commission error cannot be calculated' (p. 836), and consequently 'threshold selection methods for presence-only data are targeted at optimizing the discrimination between predicted presence and random' (pp. 836–837), and the 'two approaches employed are both based on this principle' (p. 837).

In this paper, we prove mathematically that the threshold selection method max *SSS* produces the same threshold using either presence/absence data or presence-only data. This is confirmed by simulation results using different modelling techniques and different simulated species. Furthermore, max *SSS* is an objective method, and it optimizes the discrimination between presence and absence rather than between presence and random point. In contrast, none of the other threshold selection methods has all these properties.

## MATERIALS AND METHODS

### Theoretical consideration of the threshold selection methods

More than a dozen threshold selection methods have been used with presence/absence data (see Liu *et al.*, 2005; Jiménez-Valverde & Lobo, 2007; Pearson, 2007; Freeman & Moisen, 2008; Nenzén & Araújo, 2011). The fixed threshold method is essentially arbitrary and is not considered in this study (see Liu *et al.*, 2005). The methods we evaluate include: (1) training data prevalence (trainPrev), (2) mean predicted value for a set of random points over the whole study area (meanPred), (3) mid-point between the average predicted values for the presences and the absences (midPoint), (4) maximizing *kappa* (max *kappa*), (5) maximizing overall accuracy (max *OA*), (6) maximizing the *F* measure (max *F*), (7) maximizing the sum of sensitivity and specificity (max *SSS*), (8) minimizing the difference between sensitivity and specificity (min *DSS*), (9) minimizing the difference between precision and recall (min *DPR*), (10) minimizing the distance between the receiver operating characteristic (ROC) curve and the point (0,1) (min $D_{01}$), (11) minimizing the distance between the precision–recall curve and the point (1,1) (min $D_{11}$), and (12) the predicted and the observed prevalence equalization (equalPrev) (see Liu *et al.*, 2005, and Nenzén & Araújo, 2011, for detailed explanation). In this section, we investigate mathematically whether the same threshold is selected by each of these methods using either presence/absence data or presence-only data (see also Appendix S1 in Supporting Information).

Because trainPrev uses only training points, a unique threshold should be obtained by this method for a specific model. Similarly, meanPred uses only random points and, provided that a large number of random points are used, a unique threshold should be obtained by this method for a specific model.

The test data can be a large random sample from the study area, but a more realistic situation is where the test data set contains two separate samples, with one randomly sampled from all the presences and the other from the whole study area. It can be proved that the same result will be derived from these two sampling schemes (see Appendix S1). For simplicity and convenience of explanation, we assume in the following that the test data set is a large random sample.

For the purposes of this investigation, we also assume that a species' realized distribution and potential distribution completely overlap, and we consider the points within the realized distribution as true presences and those outside as true absences. In the Discussion, we extend this to the general situation where the potential distribution is larger than the realized distribution. We will prove that the conclusions obtained here will remain valid in the general situation if we assume that the occupied and the unoccupied suitable areas are statistically similar in terms of the environmental variables selected for modelling (i.e. they are environmentally similar).

Suppose $p$ is the species' prevalence within the test data (i.e. the proportion of presences in the entire test data), but only a part of the total presences accounting for a proportion $p - r$ ($0 < r < p < 1$) of the entire test data points are taken as presences. The remaining presences, accounting for a proportion $r$ of the entire test data points, and the true absences, accounting for a proportion $1 - p$ of the entire test data points, are together considered pseudo-absences. Let $M$ and $M'$ represent the estimates of a metric calculated with presence/absence data and with presence-only data, respectively. From the true presence component of the test data, we can estimate the sensitivity ($Se$) of the model, which is appropriate for both presence/absence data and presence-only data because no absence data are required for its estimation, i.e. $Se = Se'$. However, this is not the case for specificity ($Sp$), which can only properly be estimated with presence/absence data. The pseudo-specificity can be formulated as:

$$Sp' = [r(1 - Se) + (1 - p)Sp]/(1 - p + r).$$

Thus,

$$\begin{aligned} Se' + Sp' &= Se + Sp' \\ &= r/(1 - p + r) + [(1 - p)/(1 - p + r)](Se + Sp). \end{aligned}$$

Because $r$ is constant for a data set and $(1 - p)/(1 - p + r) > 0$, $SSS' \equiv Se' + Sp'$ is a monotonically increasing function of $SSS \equiv Se + Sp$. Therefore, if a threshold maximizes $SSS'$, it will also maximize $SSS$, and vice versa. That is, the same threshold value is selected by max $SSS$ using either presence/absence data or presence-only data.

We know that the vertical distance from a point on the ROC curve to the diagonal line is

$$VDr = Se - (1 - Sp) = Se + Sp - 1 = SSS - 1 = TSS$$

(i.e. the true skill statistic) for presence/absence data, and

$$VDr' = Se' - (1 - Sp') = Se' + Sp' - 1 = SSS' - 1 = TSS'$$

for presence-only data. Therefore, if a threshold maximizes $SSS'$ and therefore $VDr'$ and $TSS'$, it will also maximize $SSS$ and therefore $VDr$ and $TSS$, and vice versa. This means that maximizing any of the measures $SSS'$, $VDr'$, $TSS'$, $SSS$, $VDr$ or $TSS$ is equivalent for threshold selection.

A special type of presence-only data may be considered where the absence component contains only random points selected from the study area. For this type of data, the above statements remain valid, and we use the terms lift curve and $VDl$ (see Liu *et al.*, 2012, for an explanation of these terms) instead of ROC curve and $VDr$.

From the following equations, we can conclude that the four methods (max $OA$, max $kappa$, max $F$ and min $DSS$) are not suitable for presence-only data because maximizing $OA$, $kappa$ and $F$ and minimizing $DSS$ ($= Se - Sp$) may not be equivalent to maximizing $OA'$, $kappa'$ and $F'$ and minimizing $DSS'$ ($= Se' - Sp'$) respectively:

$$OA' = OA + r(1 - 2Se),$$

$$DSS' = [(1 - p)/(1 - p + r)]DSS + [r/(1 - p + r)](2Se - 1),$$

$$kappa' = kappa - r^2 p_{+1}/[(1 - EA)^2 + r(2p_{+1} - 1)(1 - EA)],$$

$$F' = F - 2rp_{+1}Se/[(p + p_{+1})(p + p_{+1} - r)],$$

where $EA = 1 - p + (2p - 1)p_{+1}$ and $p_{+1} = pSe + (1 - p)(1 - Sp)$.

For min $D_{01}$, because

$$D_{01} = [(1 - Se)^2 + (1 - Sp)^2]^{1/2}$$

and

$$D'_{01} = \{(1 - Se)^2 + [rSe + (1 - p)(1 - Sp)]^2/(1 - p + r)^2\}^{1/2},$$

both higher $Se$ and higher $Sp$ tend to be obtained by minimizing $D_{01}$, but $Se$ is likely to be compromised by minimizing $D'_{01}$. Therefore, different thresholds may be produced by this method with presence/absence data and presence-only data.

For min $D_{11}$, because

$$D_{11} = [(1 - PPV)^2 + (1 - Se)^2]^{1/2}$$

and

$$D'_{11} = \left\{[1 - (PPV - rSe/p_{+1})]^2 + (1 - Se)^2\right\}^{1/2},$$

both higher $PPV$ (positive predictive value or precision) and higher $Se$ (sensitivity or recall) tend to be obtained by minimizing $D_{11}$, but $Se$ may be compromised by minimizing $D'_{11}$.

For min $DPR$, because $DPR = PPV - Se$ and $DPR' = [(p - r)/p]PPV - Se$, minimizing $DPR$ may be inconsistent with minimizing $DPR'$. They may therefore produce different thresholds.

For equalPrev, where the predicted prevalence is $pSe + (1 - p)(1 - Sp) \equiv PP$, the observed prevalence is $p$ when the data set is used as presence/absence data and $p - r$ when the data set is used as presence-only data (i.e. presence/

pseudo-absence data here). If $r \neq 0$, no threshold can be found to satisfy both $PP = p$ and $PP = p - r$. Therefore, different thresholds will be produced by this method using presence/absence data and presence-only data.

Now we consider midPoint. Suppose the data set contains $n_1$ presences with model-predicted values $x_i$ ($i$ = 1, 2, ⋯, $n_1$) and $n_0$ absences with model-predicted values $y_i$ ($i$ = 1, 2, ⋯, $n_0$). The mid-point value between $\overline{x} = \sum_{i=1}^{n_1} x_i/n_1$ and $\overline{y} = \sum_{i=1}^{n_0} y_i/n_0$ is $a = (\overline{x} + \overline{y})/2$. If we only use $n_1 - m$ presences (e.g. $x_i$, $i$ = 1, 2, ⋯, $n_1 - m$) as true presences, and take the remaining presences and the absences together as pseudo-absences, the new mid-point value becomes $a' = (\overline{x}' + \overline{y}')/2$, where $\overline{x}' = \sum_{i=1}^{n_1-m} x_i/(n_1 - m)$ and $\overline{y}' = \left(\sum_{i=n_1-m+1}^{n_1} x_i + \sum_{i=1}^{n_0} y_i\right)/(m + n_0)$. For a large random sample, we can assume $\overline{x}' = \overline{x} = \sum_{i=n_1-m+1}^{n_1} x_i/m$, then $a' = a + m(\overline{x} - \overline{y})/[2(m + n_0)]$. For a reasonable model, we can further assume $\overline{x} > \overline{y}$; thus, $a' > a$. Therefore, different thresholds may be selected by this method using presence/absence data and presence-only data.

## Generation of virtual species

In this study, we used virtual species distributed in a 250 km × 250 km area, in central–western Victoria, Australia. This study area spans a range of distinct geomorphological and climatic contexts from near sea level to 1100 m in elevation. It was primarily selected to incorporate environmental heterogeneity. Its dimensions were chosen to expedite data manipulation and iterative modelling. Eighteen environmental variables, including bioclimatic, topographical and radiometric variables (see Liu *et al.*, 2012, for details), were used in principal components analysis (PCA) to extract six principal components ($x_j$, $j$ = 1, 2, ⋯, 6), which accounted for more than 85% of the total variation. These principal components were used for modelling. All of the environmental data were resolved to 1 km × 1 km, creating a total of $n_T$ = 62,500 cells.

Virtual species were simulated as follows. For each species, we selected random values $a_j$ ($j$ = 0, 1, ⋯, 6) and $b_{jk}$ ($j$, $k$ = 1, 2, ⋯, 6), which were uniformly distributed on the interval (0,1). For site $i$ ($i$ = 1, 2, ⋯, $n_T$) with environmental data $X_i = (x_{i1}, x_{i2}, \cdots, x_{i6})$, the suitability for the species was calculated as

$$P(X_i) = \frac{1}{1 + e^{-f(X_i)}},$$

where

$$f(X_i) = a_0 + \sum_{j=1}^{6} a_j x_{ij} + \sum_{j,k=1}^{6} b_{jk} x_{ij} x_{ik}. \quad (1)$$

To make the prevalence of the species a specific value $p$, the $pn_T$ cells with the highest suitabilities were labelled as presences, and all other cells as absences. Some examples of the simulated species are shown in Appendix S2.

## Eight types of models

Four modelling techniques were used: (1) a method based on Mahalanobis distance (MD); (2) ecological niche factor analysis (ENFA); (3) generalized additive model (GAM); and (4) random forest (RF). For MD, only presence data were used to build the models. For ENFA, both presence data and random point data were required. Three types of models for GAM and RF were built with presences/absences (GAM_PA and RF_PA), presences/pseudo-absences filtered with MD (GAM_POf and RF_POf, see the detailed description in the next sub-section), and presences/pseudo-absences randomly sampled from the study area (GAM_POr and RF_POr). The same random-point training data were used for ENFA, GAM_POr and RF_POr models. All calculations were carried out in R 2.10.1 (R Development Core Team, 2010). The R packages mgcv 1.6-1 and randomForest 4.5-34 were used to implement GAM and RF, respectively. MD and ENFA were implemented with our custom programming of the algorithms provided by Farber & Kadmon (2003) and Hirzel *et al.* (2002), respectively.

Preliminary results determined that the accuracy of MD and ENFA models for the virtual species generated with equation (1) was very low. Consequently, virtual species for these two types of models were simulated with a simpler linear function, i.e. setting all the coefficients $b_{jk} = 0$ ($j$, $k$ = 1, 2, ⋯, 6) in equation (1) (see Appendix S2 for the examples of simulated species distributions).

## Creation of different data sets

To verify the theoretical results, we conducted two groups of simulations. For Group 1 simulations, we simulated the distribution of 1000 species at three levels of species prevalence ($p$): 0.05, 0.25 and 0.75. Six data sets were generated for each species: (1) training data [with 0.03 $pn_T$ presences and 0.03 $(1-p)$ $n_T$ absences]; (2) auxiliary data [with 0.03 $pn_T$ presences and 0.03 $(1 - p)$ $n_T$ absences]; (3) test data [with 0.05 $pn_T$ presences and 0.05 $(1 - p)$ $n_T$ absences]; (4) pre-processed random-point training data (with 20,000 $p$ random points); (5) random-point training data (with 0.06 $pn_T$ random points); and (6) auxiliary random-point data (with 0.03 $n_T$ random points). Stratified random sampling was employed to generate the first three data sets, and completely random sampling was used to generate the last three data sets. The first three data sets did not intersect with each other, i.e. they were made sequentially without replacement, while the last three data sets may intersect with each other and with the first three data sets.

For Group 2 simulations, we only simulated one species for each of the three levels of species prevalence, i.e. we only simulated three species. We created four of the above six data sets for each simulated species, including data sets 1, 3,

4 and 5 as described above, which were model-training data and test data. That is, we fixed the model-training data (and therefore, the model itself) and test data. We then created auxiliary data (data sets 2 and 6 in the above) for threshold selection, and this was replicated 1000 times.

Model-filtered pseudo-absences were generated for both groups of simulations. Training presences were used to build an MD model and a threshold was set to ensure that the sensitivity of this MD model was at least 0.95. The MD model was applied to the pre-processed random-point training data. Sites with predicted suitability values lower than the selected threshold were taken as candidate pseudo-absences, from which the required number ($0.06\ pn_T$) of points were randomly selected.

In this study, the number of model-filtered pseudo-absences and the number of random points used for model training were always twice the number of the presences in the training data. Whereas others have employed more pseudo-absences or random points (e.g. Ferrier *et al.*, 2002; Phillips *et al.*, 2006; Raes & ter Steege, 2007), we found this level of the pseudo-absence/true presence ratio to provide consistently reliable results. Furthermore, because our purpose was to investigate the performance of various threshold selection methods and not the models themselves, it was not deemed necessary to fine-tune all the models to their optimal capacity.

The original auxiliary data were manipulated to make two additional auxiliary data sets with different levels of 'absence pseudoness' (the degree that the pseudo-absences are not true absences; the more true presences are included in the pseudo-absences, the higher the pseudoness). We randomly selected specific proportions (25% and 75%) of the original presences in the auxiliary data set and combined them with the original true absences in that data set to create new pseudo-absences. These pseudo-absences and the remaining presences made up the two new auxiliary data sets.

### Threshold selection

We have shown that max *kappa* and min $D_{01}$ are theoretically unsuitable for presence-only data, whereas max *SSS* is suitable, and we wanted to determine if the same holds for empirical tests. We have also shown that unique threshold can be selected with meanPred, but its ability to differentiate between presence and absence remains unknown, and this therefore needs to be established empirically. Therefore, these four methods were chosen for further investigation. In the presence/absence situation, max *kappa* is popular and was recommended by Freeman & Moisen (2008) but not by Liu *et al.* (2005) and Jiménez-Valverde & Lobo (2007). The other three methods were all recommended by Liu *et al.* (2005).

The auxiliary data and the auxiliary random-point data were used for threshold selection. The auxiliary data were used for max *kappa*, and the auxiliary random point data were used for meanPred. Two versions were calculated for max *SSS* and min $D_{01}$. The first version was obtained with the auxiliary data and the data sets manipulated from them (max $SSS^r$ and min $D^r_{01}$). The second version was obtained with the auxiliary random-point data and the presence component of the auxiliary data and the data sets created from them (max $SSS^l$ and min $D^l_{01}$).

For the two versions of max *SSS* and min $D_{01}$ to be comparable, the same presence data were always used for all threshold selection methods at each level of absence pseudoness.

### Model assessment

The selected thresholds were applied to independent test data, and the accuracy of the binary predictions was evaluated with four accuracy measures (see Liu *et al.*, 2011, for details): sensitivity (*Se*), specificity (*Sp*), true skill statistic (*TSS*) and *kappa*.

## RESULTS

The results identified similar trends for each of the eight model types with respect to the selected thresholds. There was a clear effect of absence pseudoness for max *kappa* and min $D^r_{01}$, which frequently produced higher thresholds using manipulated data than using unmanipulated data, and almost no such effect for max $SSS^r$, and there was no obvious effect of the number of presences for max $SSS^l$ and min $D^l_{01}$ (Fig. 1, Appendix S3). At each level of species prevalence, the thresholds selected by max *kappa* and min $D^r_{01}$ always increased with increased absence pseudoness, and the thresholds selected by max *SSS* (both max $SSS^r$ and max $SSS^l$) and min $D^l_{01}$ remained almost unchanged when either absence pseudoness increased (for max $SSS^r$), or the number of known presences decreased (for max $SSS^l$ and min $D^l_{01}$), and those selected by max $SSS^r$ and max $SSS^l$ were almost the same (Fig. 2). Furthermore, when species prevalence was very low (0.05), the thresholds selected by max *SSS* and min $D_{01}$ were always substantially different from those selected by max *kappa*. When species prevalence was not high, meanPred produced lower thresholds than the other methods in most situations.

The accuracy of the binary results transformed with the selected thresholds is shown in Figs 3 & 4. For each level of species prevalence, the two versions of max *SSS* produced an almost identical level of accuracy. When species prevalence was low, max *kappa* almost always produced lower *Se* and *TSS* and higher *Sp* and *kappa*, and meanPred usually produced higher *Se* and lower *Sp*. For any level of species prevalence, meanPred sometimes produced both lower *TSS* and lower *kappa*, e.g. for ENFA at very high prevalence (Fig. 3). When species prevalence was very low, max *SSS* and min $D_{01}$ produced very similar levels of accuracy. When species prevalence was not low, max *SSS* almost always produced higher *Se* and *TSS* and relatively lower *Sp* than other methods (the only exception is for MD at prevalence 0.75);

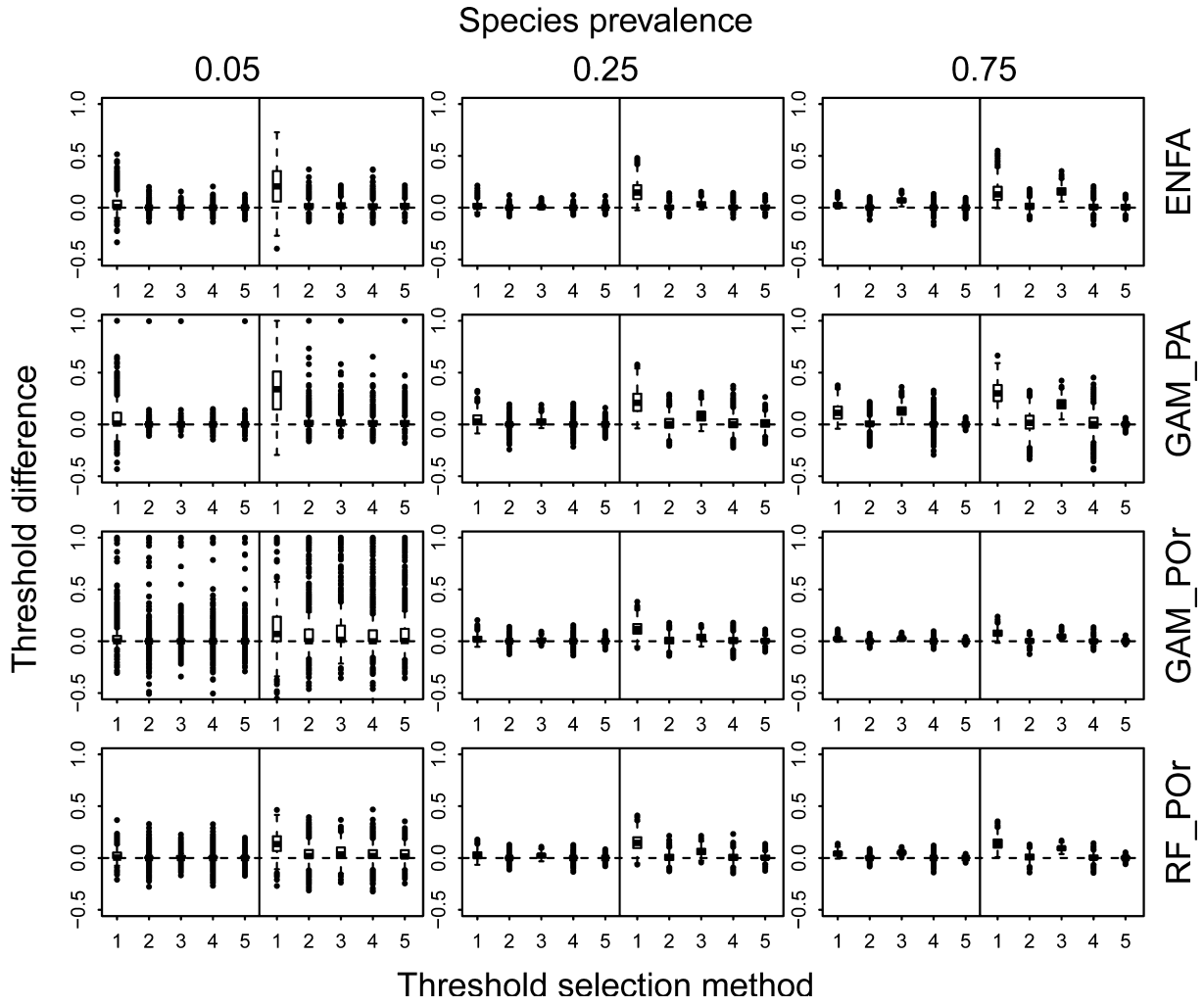


**Figure 1** Difference of threshold between that calculated with manipulated data and unmanipulated data for three levels of species prevalence (0.05, 0.25 and 0.75) from Group 1 simulations (i.e. 1000 different species at each level of prevalence) for four types of models: ecological niche factor analysis (ENFA), generalized additive models built with presences/absences (GAM_PA), and generalized additive and random forest models built with presences/pseudo-absences randomly sampled from the study area (GAM_POr and RF_POr, respectively). For each individual plot, there are two sections, corresponding to two levels of absence pseudoness (left, 25%; right, 75%). The codes for the five threshold selection methods (or variates) correspond to max *kappa*, max $SSS^r$, min $D^r{}_{01}$, max $SSS^l$ and min $D^l{}_{01}$, respectively. In each individual plot, the height of the horizontal dashed line is 0.

when species prevalence was very high, it also produced higher *kappa*.

The threshold selected using the training data was very similar to that selected using independent data for each threshold selection method and for each of the five types of models (MD, ENFA, GAM_PA, GAM_POf and GAM_POr) at each level of prevalence except GAM_POf and GAM_POr at very low prevalence (Fig. 5). In the exceptional situations (including GAM_POf and GAM_POr at very low prevalence and RF_PA, RF_POf and RF_POr at almost all levels of prevalence), the threshold selected using the training data was very different from that selected using independent data for each threshold selection method. In these situations, meanPred performed much better than the other threshold selection methods using training data.

## DISCUSSION

### Properties of the threshold selection methods

There are two main purposes for the conversion of continuous model outputs to binary results. The first is for 'real-world' applications (i.e. likely distributions or not), and the second is for evaluating model accuracy using a confusion table and measures derived from it. Thresholds should not be chosen arbitrarily (Hernandez *et al.*, 2006), and their determination should be attentive to the relative importance of the two primary forms of errors: omission and commission.

Many species distribution models (SDMs) are developed where reliable absence data are unavailable. In these cases, only omission error can be estimated. Therefore, even if we

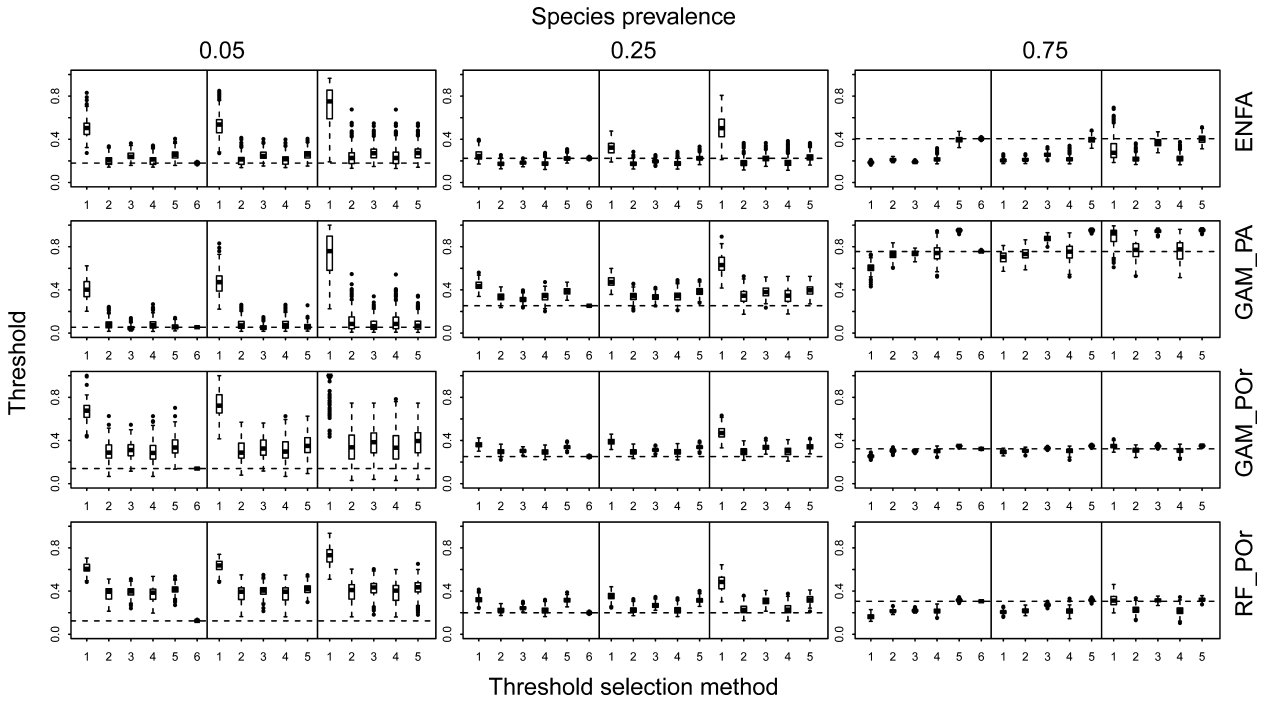


**Figure 2** Threshold selected using independent data for three levels of species prevalence (0.05, 0.25 and 0.75) from Group 2 simulations (i.e. one species at each level of prevalence) for four types of models: ecological niche factor analysis (ENFA), generalized additive models built with presences/absences (GAM_PA), and generalized additive and random forest models built with presences/pseudo-absences randomly sampled from the study area (GAM_POr and RF_POr, respectively). For each individual plot, there are three sections corresponding to three levels of absence pseudoness (0%, 25% and 75% from left to right). The codes for the six threshold selection methods (or variates) correspond to max *kappa*, max $SSS^r$, min $D^r{}_{01}$, max $SSS^l$, min $D^l{}_{01}$ and meanPred, respectively. In each individual plot, the dashed horizontal line corresponds to the median of the meanPred.

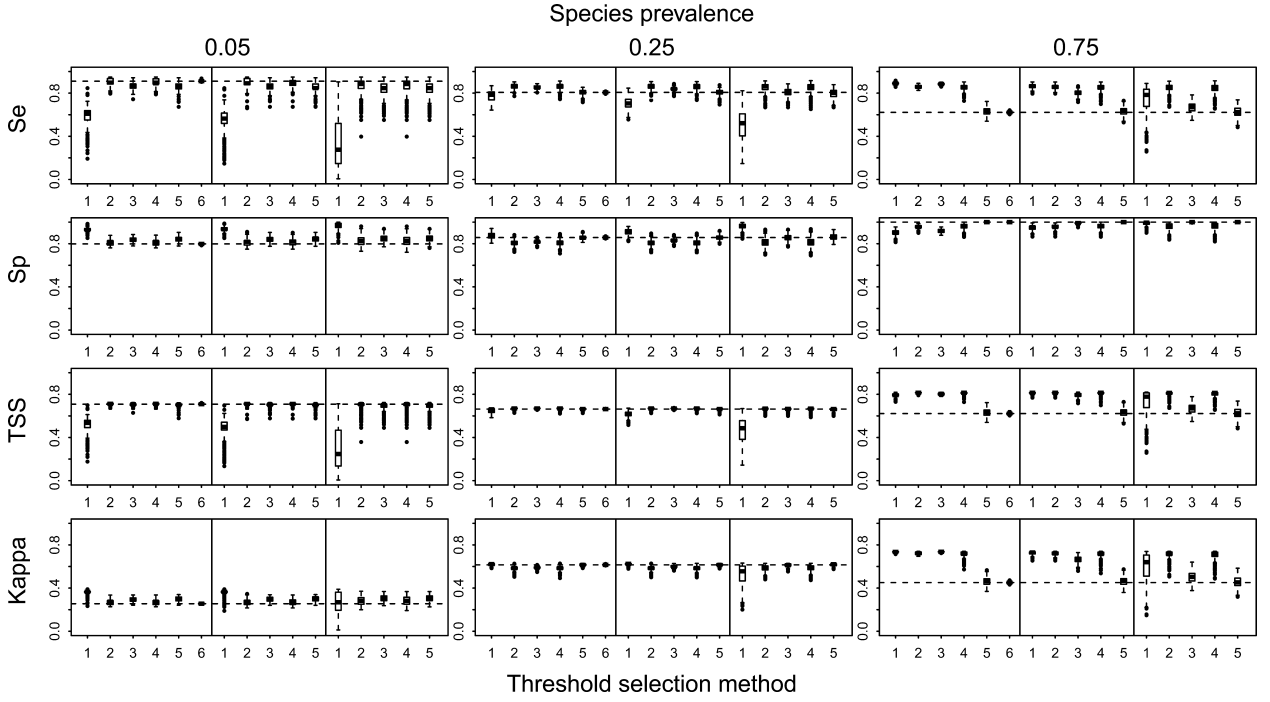


**Figure 3** Accuracy for ecological niche factor analysis (ENFA) models after the modelling results were transformed using the six threshold selection methods (or variates) from Group 2 simulations (i.e. one species at each level of prevalence). See the explanation in Fig. 2 for more information. *TSS*, true skill statistic; *Sp*, specificity; *Se*, sensitivity.

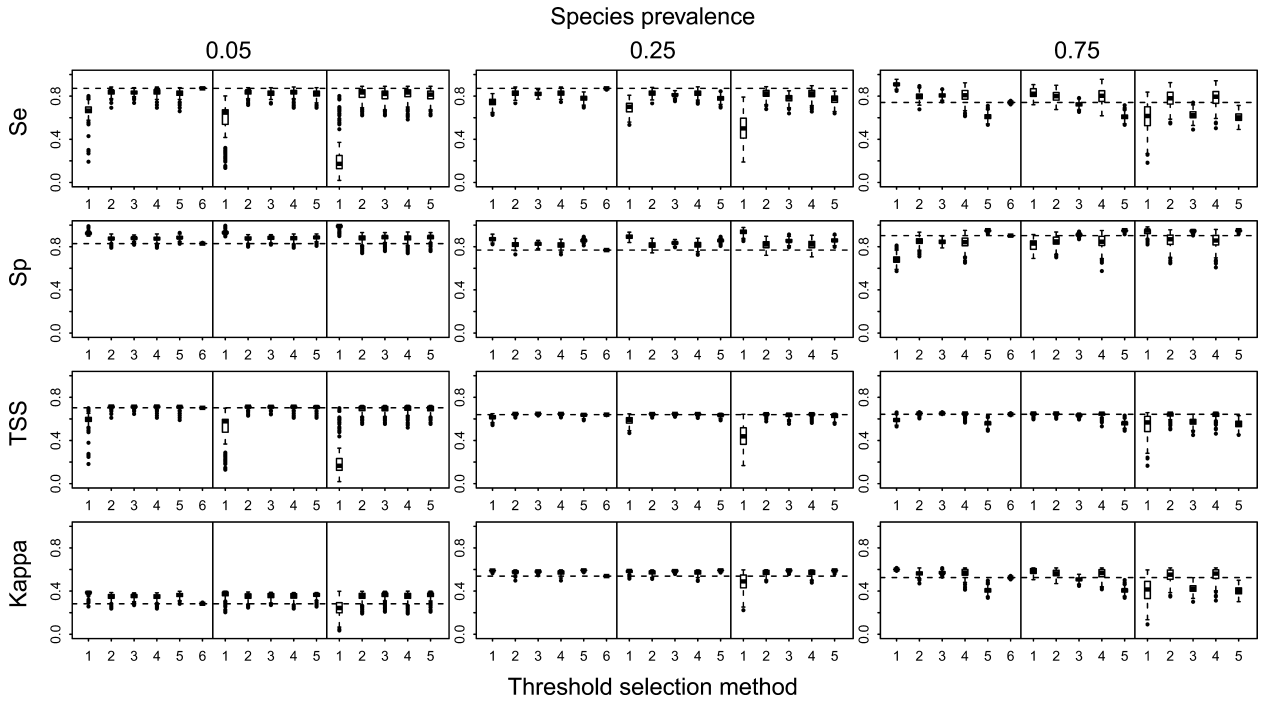


**Figure 4** Accuracy for GAM_POr models after the modelling results were transformed using the six threshold selection methods (or variates) from Group 2 simulations (i.e. one species at each level of prevalence). See the explanation in Fig. 2 for more information.

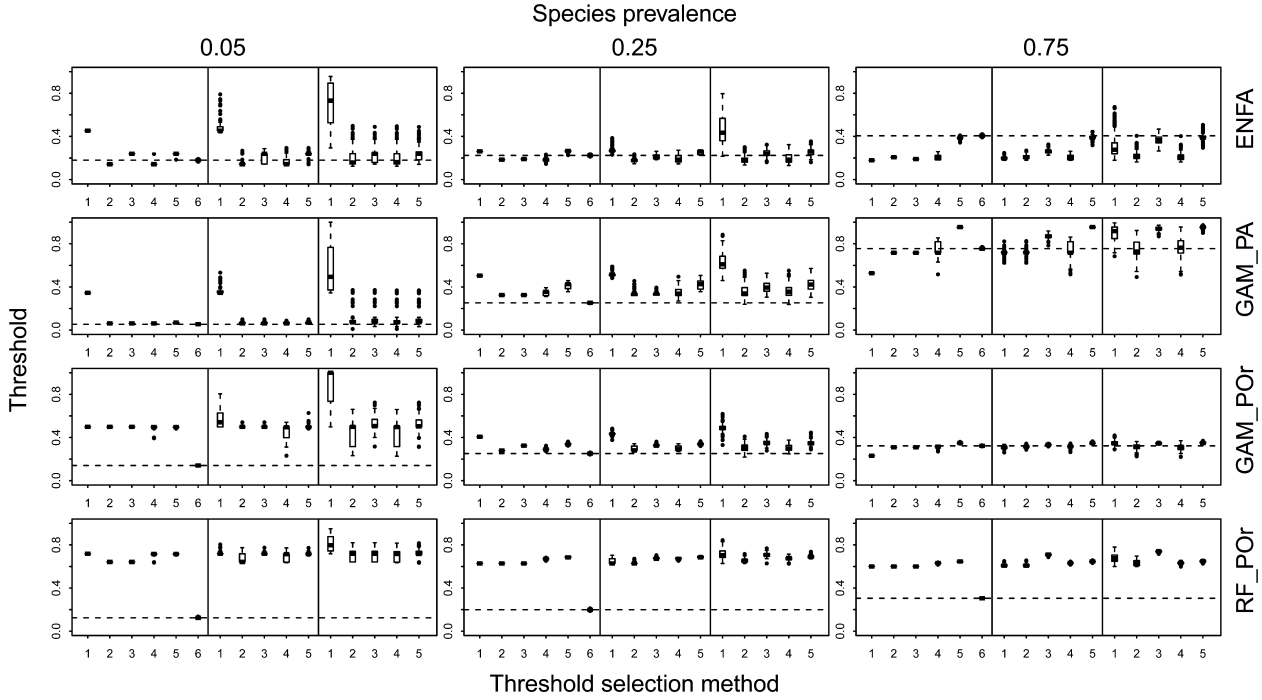


**Figure 5** Threshold selected using model training data for three levels of species prevalence (0.05, 0.25 and 0.75) from Group 2 simulations (i.e. one species at each level of prevalence). See the explanation in Fig. 2 for more information.

understand the importance and implications of each of the two types of errors, it is not possible to evaluate their relative influence. To select an optimal threshold for a particular application, a sound method should be adopted, and the resulting threshold can be used at least as a benchmark and adjusted to satisfy the specific purpose.

In this paper we attempt to characterize the approaches to selecting thresholds for SDMs when only presence data are

available. In this context, what constitutes a sound threshold selection method? First, the threshold should be objectively selected. Second, at least for large samples, the threshold selected should be identical irrespective of whether we are using presence/absence data or presence-only data. Third, discrimination between presence and absence rather than between presence and random point should be optimized. These are called objectivity, equality and discriminability criteria, respectively.

It is obvious that the max *SSS* threshold selection method satisfies the first criterion. We have proved mathematically and demonstrated empirically that it also satisfies the second criterion. Because max *SSS* is equivalent to maximizing the true skill statistic (*TSS*), and *TSS* is a well-accepted accuracy measure, discrimination between presence and absence should be optimized with this method. It is also supported by our simulation results, where both higher *Se* and higher *Sp* were produced. This means that max *SSS* also satisfies the third criterion. Therefore, max *SSS* is a promising threshold selection method when only presence data are available.

We have also shown that trainPrev and meanPred are not affected by the pseudo-absence, and a unique threshold should be selected for each model by each of these two methods. The objectivity and equality criteria are certainly met by them, but the discriminability criterion is not guaranteed to be met. For example, meanPred usually produced lower *Se* when species prevalence was very high. Although we did not include trainPrev in our simulation study, its performance can be easily inferred from the simulation results, because the model training data prevalence is known for every situation (except MD models which only used presence data). That is, for GAM_PA and RF_PA models, the model training data prevalence was the same as the species prevalence in the whole study area (i.e. 0.05, 0.25 and 0.75), and for all the other models, the model training data prevalence (here the apparent prevalence) was 0.33. Therefore, the thresholds produced by meanPred and trainPrev were very similar for GAM_PA and RF_PA models at all three levels of species prevalence and similar for the ENFA, GAM_POf, GAM_POr, RF_POf and RF_POr models at high level of species prevalence, and the thresholds selected by meanPred were lower than those selected by trainPrev when species prevalence was low and moderate. In the latter situation, the thresholds selected by max *SSS* usually lay between those selected by meanPred and trainPrev. However, meanPred (and also trainPrev) usually produced lower *Se* when species prevalence was high. Therefore, these two methods are not as useful as max *SSS*.

We have stated that the fixed threshold method is not objective and we have also theoretically proved that all the other methods do not meet at least the equality criterion. Therefore, none of these methods are suitable for presence-only data.

In the introduction, we mentioned other threshold selection methods. The least presence threshold (LPT) method is an objective method. It produces the highest sensitivity ($Se = 1$) but very low specificity. Therefore, this method cannot be recommended. The other threshold selection methods mentioned in the introduction are also not recommendable, including the required-sensitivity method and the method based on Hirzel *et al.*'s (2006) continuous *P*/*E* curve.

In general, model accuracy increases as training data set size increases (e.g. Stockwell & Peterson, 2002; Hernandez *et al.*, 2006; Wisz *et al.*, 2008). For some taxa, especially rare and restricted species, the limits to data availability may require the entire data set to be used for model training (e.g. Papeş & Gaubert, 2007), leaving no data for threshold selection. In this situation, if the results from training data are comparable (to some extent) with those from independent data, selecting a threshold with training data will provide a method for these data-limited taxa. Our simulation results show that the success of this approach can be dependent on model type. The difference between the threshold selected with training data and that selected with independent data was substantial for the three types of RF models and was slight for the other five types of models. This means that training data cannot be used to select the threshold for RF models, but can be used for the other five types of models investigated. For RF models, the results from meanPred and trainPrev were much better than those from the other threshold selection methods using training data. We recommend that trainPrev be used for RF_PA models, while for RF_POf and RF_POr models, meanPred should be used when species prevalence is moderate and high and trainPrev should be used when species prevalence is very low.

We cannot, however, give a general recommendation on how to choose a threshold with model training data for other modelling techniques that were not investigated in this study, and further investigation is needed.

### Output represents potential distribution

We have previously suggested two distinct uses for transforming the continuous modelling results into binary ones: to assess model accuracy, and to provide binary models for use in 'real-world' applications (e.g. conservation planning and management). However, the question remains whether the transformed binary output represents a species' potential or realized distribution. We believe that with the approaches outlined in this paper, the transformed models represent species' potential distributions for almost all (large-scale) correlative species distribution models. This is primarily a reflection of our assumption that the sites in the potential distributions are indistinguishable from those in the realized distributions when examined against environmental variables (usually abiotic) used in the model training and evaluation. Using either the sites in the realized or potential distributions as true presences, the same threshold will be selected, which can be explained as follows.

Suppose the study area consists of three types of areas: occupied suitable areas (A) with proportion $p-r$, unoccupied suitable areas (B) with proportion $r$, and unsuitable areas

(C) with proportion $1-p$. If a random sample is taken from the study area, it is reasonable to assume that the proportions of sampling points falling into the three types of areas (A, B and C) are $p - r$, $r$ and $1 - p$, respectively. If we focus on the potential distribution (Potential Scenario), the presences are from suitable areas (both occupied and unoccupied, i.e. $A \cup B$) with proportion $p$ and absences are from the unsuitable areas (i.e. C) with proportion $1 - p$. If we focus on the realized distribution (Realized Scenario), the presences are from the occupied suitable areas (i.e. A) with proportion $p - r$ and the absences are from both the unoccupied suitable and the unsuitable areas (i.e. $B \cup C$) with proportion $1 - p + r$. According to our assumption that the sites from the occupied and unoccupied suitable areas (i.e. A and B) are environmentally similar, all these sites can simply be considered as presences, and the sensitivity ($Se$) calculated for the two scenarios (Realized and Potential scenarios) should be the same. The absences for the Realized Scenario (i.e. the unoccupied suitable and unsuitable sites $B \cup C$) include the absences for the Potential Scenario (i.e. the unsuitable sites C) and the unoccupied suitable sites (i.e. B), and the latter are just a part of presences. If sites from C (i.e. the absences for the Potential Scenario) are considered true absences, the sites from $B \cup C$ (i.e. the absences for the Realized Scenario) can be considered as pseudo-absences (because they include some presences). If we denote the Realized and Potential scenarios with R and P, respectively, and use the reasoning similar to that used in the Methods section, we have $Se^{(R)} = Se^{(P)}$, and $Sp^{(R)} = [r(1 - Se^{(P)}) + (1 - p)Sp^{(P)}]/(1 - p + r)$. Thus, $Se^{(R)} + Sp^{(R)} = r/(1 - p + r) + [(1 - p)/(1 - p + r)](Se^{(P)} + Sp^{(P)})$. Because $(1 - p)/(1 - p + r) > 0$, $SSS^{(R)}$ $[\equiv Se^{(R)} + Sp^{(R)}]$ is a monotonically increasing function of $SSS^{(P)}$ $[\equiv Se^{(P)} + Sp^{(P)}]$. Therefore, maximizing $SSS^{(R)}$ (calculated with the presences from the realized distribution) is equivalent to maximizing $SSS^{(P)}$ (calculated with the presences from the potential distribution). In other words, the same threshold will be selected using presence data from either the potential distribution (i.e. all suitable areas, although it is impossible in practice because we do not know the unoccupied suitable areas) or the realized distribution (i.e. the occupied suitable areas).

From this logic, it is clear that the predicted distribution using the selected threshold represents the species' potential distribution. Without incorporating detailed information about the biotic and anthropogenic factors that affect the species distribution in model development, it is impossible to predict the species' realized distribution, because the model always treats the occupied suitable areas and the unoccupied suitable areas equally. Therefore, even if we use presence data from the realized distribution to select the threshold, the threshold can only possibly delimit the potential distribution. If an estimate of the realized distribution is the primary objective for a particular application, then the reclassified distribution using the selected threshold can be post-processed with other ancillary information (regarding species dispersal, biotic interaction and human modification of the environment) through a series of steps (e.g. Phillips *et al.*, 2006; Guisan & Rahbek, 2011; Boulangeat *et al.*, 2012).

### The sampling assumption and simulating species distributions

A critical point of this work is the assumption of random sampling. That is, either the entire data set is a single random sample from the whole study area, or it contains two separate random samples with one randomly sampled from all the presence points (i.e. random presences) and the other randomly sampled from the whole study area (i.e. random points). Certainly the latter sampling scheme is more realistic. Under these two sampling schemes, max *SSS* produces the same threshold as with known true presences and true absences, provided the sample is large. However, the same threshold cannot be guaranteed for small-sized samples. Therefore, the small-sample-size effect needs to be further investigated, as in Bean *et al.* (2012).

For the presence-only situation, this random sampling assumption is almost always violated, but as we know, random sampling in geographical space is not essential for building species distribution models. Spatial bias in the records may not be a problem if the data are not environmentally biased (Newbold, 2010). Although some previous studies have shown that museum records did not completely capture the environmental conditions inhabited by the target species (e.g. Hortal *et al.*, 2008), others have shown that spatially biased museum record data are unrelated to major biases in environmental space (e.g. Kadmon *et al.*, 2004). Intuitively, good models will be difficult to construct from environmentally biased data. Therefore, data that are unbiased in the environmental space are ideal for building good species distribution models.

Following this logic, our random sampling assumption can be relaxed accordingly. Because random points are always randomly sampled, we may be able to partially relax the random sampling assumption for presences. Whichever sampling strategy is adopted, we are only concerned whether sensitivity can be accurately estimated (to some extent) from the sample of presences, and random sampling provides a simple way to obtain a good estimate of sensitivity. Actually, any sampling strategy could be used, if sensitivity can be accurately estimated. More work is still required to further verify this approach, particularly with data from spatially biased sampling.

Although virtual species have been used in previous studies, there is still no widely accepted method to develop such entities. In this paper, we simulated species distributions by only taking into account environmental conditions and have not considered broader ecological processes. Therefore, these virtual species are unlikely to be as nuanced in their distributions or associations as real-world taxa. However, the species–environment relationships examined in this paper were quite complex because both linear and non-linear (i.e. quadratic and interaction) terms were included and the coefficients were randomly generated for

each species. We therefore believe that these virtual species should mimic the characteristics of many real world species.

## CONCLUSIONS

In conclusion, the threshold selection method based on maximizing the sum of sensitivity and specificity is a promising method for use when reliable absence data are unavailable. This is equivalent to maximizing the vertical distance from a point on the ROC curve or lift curve to the diagonal line (*VDr* and *VDl*, respectively) or maximizing the true skill statistic (*TSS*). We have theoretically and empirically demonstrated that the same threshold will be selected with this method using either presence/absence data or presence-only data. When there are no independent presence data for threshold selection, the training data can be used for this purpose. In this situation, max *SSS* can be used for MD, ENFA and GAM models, and the training data prevalence (trainPrev) and the mean predicted suitability (meanPred) can be used as the threshold for RF models. Interpretations of the converted distribution models using thresholds are best considered in the majority of circumstances as potential distributions, rather than as realized distributions.

## ACKNOWLEDGEMENTS

We thank Richard Pearson, Bradford A. Hawkins, Robert J. Whittaker, Charles Todd and three anonymous referees for their valuable comments that improved this manuscript. We also thank Peter Griffioen for some data preparation. The authors are primarily supported by the Biodiversity and Ecosystem Services Division, Department of Sustainability and Environment, Victoria, Australia.

## REFERENCES

Araújo, M.B., Williams, P.H. & Fuller, R.J. (2002) Dynamics of extinction and the selection of nature reserves. *Proceedings of the Royal Society B: Biological Sciences*, **269**, 1971–1980.

Bean, W.T., Stafford, R. & Brashares, J.S. (2012) The effects of small sample size and sample bias on threshold selection and accuracy assessment of species distribution models. *Ecography*, **35**, 250–258.

Boulangeat, I., Gravel, D. & Thuiller, W. (2012) Accounting for dispersal and biotic interactions to disentangle the drivers of species distributions and their abundances. *Ecology Letters*, **15**, 584–593.

Braunisch, V. & Suchant, R. (2010) Predicting species distributions based on incomplete survey data: the trade-off between precision and scale. *Ecography*, **33**, 826–840.

Chefaoui, R.M. & Lobo, J.M. (2008) Assessing the effects of pseudo-absences on predictive distribution model performance. *Ecological Modelling*, **210**, 478–486.

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.

Farber, O. & Kadmon, R. (2003) Assessment of alternative approaches for bioclimatic modeling with special emphasis on the Mahalanobis distance. *Ecological Modelling*, **160**, 115–130.

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.

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

Franklin, J. (2009) *Mapping species distributions: spatial inference and prediction*. Cambridge University Press, Cambridge, UK.

Freeman, E.A. & Moisen, G.G. (2008) A comparison of the performance of threshold criteria for binary classification in terms of predicted prevalence and kappa. *Ecological Modelling*, **217**, 48–58.

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

Guisan, A. & Rahbek, C. (2011) SESAM – a new framework integrating macroecological and species distribution models for predicting spatio-temporal patterns of species assemblages. *Journal of Biogeography*, **38**, 1433–1444.

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

Hernandez, P.A., Graham, C.H., Master, L.L. & Albert, D.L. (2006) The effect of sample size and species characteristics on performance of different species distribution modeling methods. *Ecography*, **29**, 773–785.

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

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.

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

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.

Kadmon, R., Farber, O. & Danin, A. (2004) Effect of roadside bias on the accuracy of predictive maps produced by bioclimatic models. *Ecological Applications*, **14**, 401–413.

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

Liu, C., White, M., Newell, G. & Griffioen, P. (2012) Species distribution modelling for conservation planning in Victoria, Australia. *Ecological Modelling* (in press). doi:10.1016/j.ecolmodel.2012.07.003.

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.

Nenzén, H.K. & Araújo, M.B. (2011) Choice of threshold alters projections of species range shifts under climate change. *Ecological Modelling*, **222**, 3346–3354.

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.

Papeş, M. & Gaubert, P. (2007) Modelling ecological niches from low numbers of occurrences: assessment of the conservation status of poorly known viverrids (Mammalia, Carnivora) across two continents. *Diversity and Distributions*, **13**, 890–902.

Pearson, R.G. (2007) *Species' distribution modeling for conservation educators and practitioners*. Available at: http://biodiversityinformatics.amnh.org (accessed 23 September 2011).

Pearson, R.G., Dawson, T.P. & Liu, C. (2004) Modelling species distributions in Britain: a hierarchical integration of climate and land-cover data. *Ecography*, **27**, 285–298.

Pearson, R.G., Raxworthy, C.J., Nakamura, M. & Peterson, A. T. (2007) Predicting species' distributions from small numbers of occurrence records: a test case using cryptic geckos in Madagascar. *Journal of Biogeography*, **34**, 102–117.

Peterson, A.T., Papeş, M. & Soberón, J. (2008) Rethinking receiver operating characteristic analysis applications in ecological niche modelling. *Ecological Modelling*, **213**, 63–72.

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

R Development Core Team (2010) *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. Available at: http://www.r-project.org.

Raes, N. & ter Steege, H. (2007) A null-model for significance testing of presence-only species distribution models. *Ecography*, **30**, 727–736.

Rebelo, H. & Jones, G. (2010) Ground validation of presence-only modelling with rare species: a case study on barbastelles *Barbastella barbastellus* (Chiroptera: Vespertilionidae). *Journal of Applied Ecology*, **47**, 410–420.

Segurado, P. & Araújo, M.B. (2004) An evaluation of methods for modelling species distributions. *Journal of Biogeography*, **31**, 1555–1568.

Stockwell, D.R.B. & Peterson, A.T. (2002) Effects of sample size on accuracy of species distribution models. *Ecological Modelling*, **148**, 1–13.

Thuiller, W., Lafourcade, B., Engler, R. & Araújo, M.B. (2009) BIOMOD – a platform for ensemble forecasting of species distributions. *Ecography*, **32**, 369–373.

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

Wisz, M.S., Hijmans, R.J., Li, J., Peterson, A.T., Graham, C. H., Guisan, A. & NCEAS Predicting Species Distributions Working Group (2008) Effects of sample size on the performance of species distribution models. *Diversity and Distributions*, **14**, 763–773.

## SUPPORTING INFORMATION

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

**Appendix S1** Theoretical explanation for some threshold selection methods

**Appendix S2** Examples of simulated species distributions.

**Appendix S3** Difference of threshold using independent data from Group 1 simulations.

## BIOSKETCHES

**Canran Liu** is a quantitative ecologist with interests in species distribution modelling, diversity measurement, and spatial pattern analysis at various scales.

**Matt White** maintains interests in plant ecology, spatial modelling, and climate change issues.

**Graeme Newell** is interested in spatial modelling, vegetation condition and climate change issues.

Editor: Richard Pearson