

**Plotting receiver operating characteristic and precision-recall curves from presence and background data**

**Wenkai Li<sup>a\*</sup> and Qinghua Guo<sup>b</sup>**

<sup>a\*</sup>. School of Geography and Planning, Sun Yat-Sen University, Guangzhou 510275, China. (correspondence: liwenk3@mail.sysu.edu.cn)

<sup>b</sup>. Institute of Ecology, College of Urban and Environmental Sciences, Peking University, Beijing 100871, China.

**Running headline: ROC/PR plots without absence data**

## Abstract

1. The receiver operating characteristic (ROC) and precision-recall (PR) plots have been widely used to evaluate the performances of species distribution models. Plotting ROC/PR curves requires a traditional test set with both presence and absence data (namely PA approach), but species absence data are usually not available in reality. Plotting ROC/PR curves from presence-only data while treating background data as pseudo absence data (namely PO approach) may provide misleading results.
2. In this study we propose a new approach to calibrate the ROC/PR curves from presence and background data with user-provided information on a constant  $c$ , namely PB approach. An estimate of  $c$  can also be derived from the PB-based ROC/PR plots given that a model with good ability of discrimination is available. We used three virtual species and a real aerial photography to test the effectiveness of the proposed PB-based ROC/PR plots. Different models (or classifiers) were trained from presence and background data with various samples sizes. The ROC/PR curves plotted by PA approach were used to benchmark the curves plotted by PO and PB approaches.
3. Experimental results show that the curves and areas under curves by PB approach are more similar to that by PA approach as compared with PO approach. The PB-based ROC/PR plots also provide highly accurate estimations of  $c$  in our experiment.
4. We conclude that the proposed PB-based ROC/PR plots can provide valuable complements to existing model assessment methods, and they also provide an

additional way to estimate the constant  $c$  (or species prevalence) from presence and background data.

## KEYWORDS

area under the curve, model evaluation, precision-recall curve, presence and background data, receiver operating characteristic curve, species distribution modelling

## 1. INTRODUCTION

Species distribution modeling (SDM) is an important tool to understand the statistical relationship between occurrences of species and environmental variables, and it has been applied in a variety of fields (Peterson & Holt, 2003; Guisan & Thuiller, 2005; Elith et al., 2006). For example, Kueppers et al. (2005) used discriminant analysis to study the potential ranges of two California endemic oaks in response to regional climate change. Hagar et al. (2020) used maximum entropy (MAXENT) to predict the habitat suitability of northern spotted owl in Oregon with forest structural attributes derived from airborne light detection and ranging data. When both observed presence and absence data are available, it is straightforward to apply standard binary classifiers such as logistic regression and neural network to predict the conditional probability of species occurrence at given locations (Guisan et al., 2002; Marmion et al., 2009; Li et al., 2011). However, reliable species absence data are usually not available in practice, which is referred to as the presence-only problem (Elith et al., 2006). With presence-only data, it is difficult to estimate the probability of species occurrence, so researchers

usually estimate a relative index of habitat suitability instead (Elith et al., 2006; Hastie & Fithian, 2013; Phillips & Elith, 2013). One category of methods for presence-only data is to train models using only presence data, such as ecological niche factor analysis (Hirzel et al., 2002), BIOCLIM (Busby, 1986), and DOMAIN (Carpenter et al., 1993). Another category of presence-only methods involves generating pseudo absence or background data and combining them with observed presence data to train models, such as MAXENT, maximum likelihood analysis (MAXLIKE), inhomogeneous Poisson point process, naive logistic regression, and presence and background learning (Keating & Cherry, 2004; Phillips et al., 2006; Ward et al., 2009; Li et al., 2011; Aarts et al., 2012; Royle et al., 2012).

Model performance can be evaluated from two different aspects, namely calibration and discrimination (Lobo et al., 2008; Phillips & Elith, 2010; Jiménez-Valverde et al., 2013). Calibration measures the agreement between predicted and true probabilities of species occurrence, whereas discrimination measures the ability to distinguish between presence and absence data (Phillips & Elith, 2010). In this study we only focus on the aspect of discrimination. Using an independent test set consisting of both presence and absence data, we can generate a  $2 \times 2$  confusion matrix to cross-tabulate binary predictions and observations, from which a variety of accuracy measures can be derived, such as overall accuracy, kappa statistic, true skill statistic (TSS), and *F*-measure (Congalton, 1991; Fielding & Bell, 1997; Liu et al., 2011; Li & Guo, 2013). These accuracy measures consider both commission and omission errors, and they are threshold-dependent, so a single threshold is required to convert

continuous outputs to binary outputs. Without absence data, however, commission error cannot be calculated, making model evaluation problematic with these traditional accuracy measures. To solve this problem, absolute validation index (AVI) and contrast validation index (CVI) were proposed to evaluate binary predictions without considering commission error (Hirzel et al., 2006). Li and Guo (2013) proposed two new statistics, namely  $F_{cpb}$  and  $F_{pb}$ , to evaluate the predictive accuracy of binary predictions from presence and background data.  $F_{cpb}$  is an unbiased estimate of  $F$ -measure, but it requires prior information of species prevalence. When species prevalence is not available,  $F_{pb}$  can be applied as a proxy of  $F$ -measure, but it is only applicable to rank models for the same species because its upper bound is affected by the unknown prevalence. Liu et al. (2013) proved that maximizing TSS from presence and pseudo absence data is equivalent to maximizing TSS from presence and absence data in terms of threshold selection.

The receiver operating characteristic (ROC) curve and area under the ROC curve (AUCROC) have also been commonly used for model evaluation in SDM (Fielding & Bell, 1997). Unlike the threshold-dependent measures that rely on a single threshold, ROC curve and AUCROC evaluate model performance by considering all possible thresholds, so they are applicable to continuous outputs without requiring thresholding. Alternatively, users can plot precision-recall (PR) curve and calculate area under the PR curve (AUCPR) to evaluate model performance (Davis & Goadrich, 2006). Please note that ROC curve incorporates correctly predicted absence sites (true negative), and hence AUCROC value is influenced by total geographic extent (Lobo et al., 2008).

When species prevalence is very small or the geographic extent is very large, AUCROC value may be inflated unrealistically (Sofaer et al., 2019). By contrast, the PR curve ignores true negative, so it is more robust to geographic extent and suitable for species with small prevalence (Leroy et al., 2018; Sofaer et al., 2019). In other words, ROC curve is more suitable for balanced datasets whereas PR curve is more suitable for imbalanced datasets (Davis & Goadrich, 2006; Saito & Rehmsmeier, 2015; Sofaer et al., 2019).

Essentially, ROC and PR curves are based on both commission and omission errors, so they also suffer from the presence-only problem. Currently, it is a common practice to plot ROC/PR curves and calculate area under the curve (AUC) by treating the background data as absence data in the literature, but researchers have pointed out that this approach can make the results misleading and difficult to interpret because background data are actually contaminated by presence data (Phillips et al., 2006; Peterson et al., 2008; Jiménez-Valverde, 2012). Li and Guo (2013) have proved that both recall (inversely related to omission error) and precision (inversely related to commission error) can be unbiasedly estimated from presence and background data given that species prevalence is available, thus making it possible to plot the correct ROC/PR curves without absence data. However, this approach has not yet been applied to correct the ROC/PR curves from presence and background data in the field of SDM. In this study, therefore, we aim to investigate the following two questions. Given true species prevalence, can we plot the correct ROC/PR curves from presence and background data? Without true species prevalence, can we estimate prevalence from

presence and background data?

## 2. MATERIALS AND METHODS

### 2.1 Model evaluation with a non-traditional test set

An independent test set with random samples drawn from the population is required for model evaluation. Let  $y = 1$  denote presence data and  $y = 0$  denote absence data;  $s = 1$  denote labeled data and  $s = 0$  denote unlabeled data. A traditional test set contains fully labeled presence-absence data randomly sampled from the population. A non-traditional test set contains labeled and unlabeled data, in which only presence data are labeled and unlabeled data are a mixture of presence and absence data whose labels are unknown. In other words, the labeled data ( $s = 1$ ) must be presence data ( $y = 1$ ), but unlabeled data ( $s = 0$ ) may be presence ( $y = 1$ ) or absence ( $y = 0$ ) data in a non-traditional test set. By comparing the true labels and binary predictions on a traditional test set, we can generate a confusion matrix with four quadrants: true positive ( $TP$ ), false positive ( $FP$ ), false negative ( $FN$ ), and true negative ( $TN$ ). If we simply treat the unlabeled data as absence data in a non-traditional test set, we also create a confusion matrix with the four quadrants denoted differently: true positive ( $TP'$ ), false positive ( $FP'$ ), false negative ( $FN'$ ), and true negative ( $TN'$ ) (see Table 1).

From a traditional confusion matrix, we can calculate precision ( $p$ ), recall ( $r$ ), and false positive rate ( $FPR$ ) using the following equations:

$$p = \frac{TP}{TP+FP} \quad (1)$$

$$r = \frac{TP}{TP+FN} \quad (2)$$

$$FPR = \frac{FP}{FP+TN} \quad (3)$$

The species prevalence  $P(y = 1)$  and the proportion of predicted presences  $P(y' = 1)$  can be calculated through the following equations:

$$P(y = 1) = \frac{TP+FN}{TP+FN+FP+TN} \quad (4)$$

$$P(y' = 1) = \frac{TP+FP}{TP+FN+FP+TN} \quad (5)$$

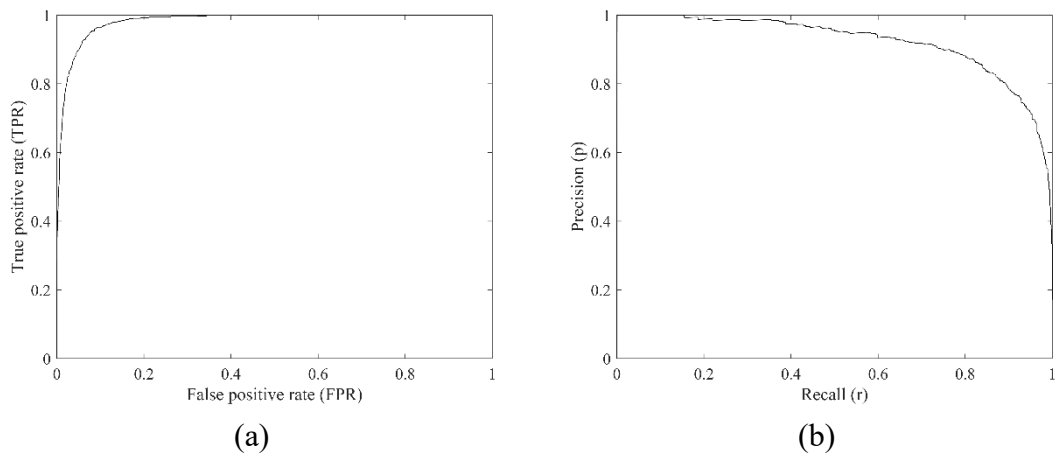
Therefore, we can rewrite Equation 3 as:

$$FPR = \frac{FP}{FP+TN} = \frac{P(y' = 1) \times (1-p)}{1-P(y=1)} \quad (6)$$

By considering all possible thresholds, the ROC curve plots true positive rate ( $TPR$ ) versus  $FPR$ , whereas the PR curve plots  $p$  versus  $r$  (Fig. 1).  $TPR$  is exactly the same as  $r$  that is related omission error ( $FN$ ), and both  $FPR$  and  $p$  are related to commission error ( $FP$ ), so we can connect ROC and PR curves through Equation 6.

**TABLE 1** Confusion matrices from traditional and non-traditional test sets.

Prediction	Reference			
	Traditional test set		Non-traditional test set	
	$y = 1$	$y = 0$	$s = 1$	$s = 0$
$y' = 1$	$TP$	$FP$	$TP'$	$FP'$
$y' = 0$	$FN$	$TN$	$FN'$	$TN'$



**FIGURE 1** Examples of ROC (a) and PR (b) curves.



On a non-traditional test set, only a proportion of presence data are labeled and the labels of absence data are unknown, so the traditional confusion matrix cannot be completely determined. Here we consider two common sampling scenarios: single-training-set (Elkan & Noto, 2008) and case-control (Lancaster & Imbens, 1996). In the single-training-set scenario, we visit a number of sites randomly distributed within the study area, and a site is labeled as presence if species occurrence is observed or unlabeled otherwise. In the case-control scenario, the labeled presence data are randomly sampled from the presence subset, and unlabeled data are randomly sampled from the population. Let  $c = P(s = 1|y = 1)$  define the probability that species occurrence is detected (labeled), i.e. the ratio of labeled presence data to the total number of presence data in a test set (Li et al., 2011). The value of  $c$  is usually smaller than one, so unlabeled data actually contain both presence and absence data.

With a non-traditional test set, we can define the following measures:

$$p' = \frac{TP'}{TP' + FP'} \quad (7)$$

$$r' = \frac{TP'}{TP' + FN'} \quad (8)$$

Because  $r'$  is calculated from the observed (labeled) presence data, we have  $r' = r$ .

However,  $p'$  is not equal to  $p$  because it is calculated from unlabeled data. According to

Li and Guo (2013),  $p'$  and  $p$  have the following relationship:

$$p = \frac{1-c}{c} \times \frac{p'}{1-p'} \quad (9)$$

in the case-control scenario. In the single-training-set scenario, their relationship is

slightly different, which should be:

$$p = p' / c \quad (10)$$

Here we use Table 2 to illustrate the derivations of Equations 9 and 10. In Table 2,  $m_1$  and  $m_4$  can be calculated from labeled data, but  $m_2$ ,  $m_3$ ,  $m_5$ , and  $m_6$  cannot be calculated because the true labels of unlabeled data are not known. Meanwhile, the total number of labeled data  $n_1$ , total number of unlabeled data  $n_0$ , total number of predicted presences  $k_1$ , total number of predicted absences  $k_0$ , and total number of test data  $t$  are known. According to the definitions of  $c$ ,  $p'$  and  $r'$ , we have the following equations:

$$c = \frac{m_1 + m_4}{m_1 + m_2 + m_4 + m_5} \quad (11)$$

$$p' = \frac{m_1}{m_1 + m_2 + m_3} \quad (12)$$

$$r' = \frac{m_1}{m_1 + m_4} \quad (13)$$

In the case-control scenario, the unlabeled data are randomly sampled from the population, so  $p$ ,  $r$ ,  $P(y = 1)$ , and  $P(y' = 1)$  can be calculated as:

$$p = \frac{m_2}{m_2 + m_3} \quad (14)$$

$$r = \frac{m_2}{m_2 + m_5} \quad (15)$$

$$P(y = 1) = (m_2 + m_5) / n_0 \quad (16)$$

$$P(y' = 1) = (k_1 - m_1) / n_0 \quad (17)$$

According to Equations 11 and 12, we have:

$$\frac{1-c}{c} \times \frac{p'}{1-p'} = \frac{m_2 + m_5}{m_1 + m_4} \times \frac{m_1}{m_2 + m_3} = \frac{m_1}{m_1 + m_4} \times \frac{m_2 + m_5}{m_2 + m_3} \quad (18)$$

Because  $r = r'$ , substituting Equations 13-15 to 18, we have:

$$\frac{1-c}{c} \times \frac{p'}{1-p'} = \frac{m_2}{m_2 + m_5} \times \frac{m_2 + m_5}{m_2 + m_3} = p \quad (19)$$

which proves the relationship between  $p$  and  $p'$  in Equation 9 under the case-control scenario. Please note that  $(1 - c) / c$  here is equal to the reciprocal of the constant term  $c$  in Li and Guo (2013), so Equation 9 of this article is equivalent to Equation 9 in Li

and Guo (2013).

Unlike the case-control scenario where unlabeled data alone are random samples of the population, the combined labeled and unlabeled data together constitute random samples of the population in the single-training-set scenario, so  $p$ ,  $r$ ,  $P(y = 1)$ , and  $P(y' = 1)$  are calculated differently:

$$p = \frac{m_1 + m_2}{m_1 + m_2 + m_3} \quad (20)$$

$$r = \frac{m_1 + m_2}{m_1 + m_2 + m_4 + m_5} \quad (21)$$

$$P(y = 1) = (m_1 + m_2 + m_4 + m_5) / t \quad (22)$$

$$P(y' = 1) = k_1 / t \quad (23)$$

According to Equations 11 and 12, we have:

$$\frac{p'}{c} = \frac{m_1}{m_1 + m_2 + m_3} \times \frac{m_1 + m_2 + m_4 + m_5}{m_1 + m_4} = \frac{m_1}{m_1 + m_4} \times \frac{m_1 + m_2 + m_4 + m_5}{m_1 + m_2 + m_3} \quad (24)$$

Since  $r = r'$ , substituting Equations 13, 20, and 21 to 24, we have:

$$\frac{p'}{c} = \frac{m_1 + m_2}{m_1 + m_2 + m_4 + m_5} \times \frac{m_1 + m_2 + m_4 + m_5}{m_1 + m_2 + m_3} = p \quad (25)$$

which proves the relationship between  $p$  and  $p'$  in Equation 10 under the single-training-set scenario.

Please note that  $c$  has the same definition in both scenarios, i.e.  $c = P(s = 1 | y = 1)$ , but its relationship with species prevalence is different in two scenarios:

$$c = n_1 / [n_1 + n_0 \times P(y = 1)] \quad (26)$$

according to Equations 11 and 16 in the case-control scenario;

$$c = n_1 / [t \times P(y = 1)] \quad (27)$$

according to Equations 11 and 22 in the single-training-set scenario. Given a non-traditional test set,  $c$  is a fixed constant whose value is affected by the number of labeled

data ( $n_1$ ), number of unlabeled data ( $n_1$ ), and species prevalence. Meanwhile,  $P(y' = 1)$  is equal to the proportion of predicted presences among the unlabeled set in the case-control scenario, or equal to the proportion of predicted presences among the whole test set in the single-training-set scenario. According to Equation 6,  $FPR$  can be determined if  $p$ ,  $P(y' = 1)$ , and  $P(y = 1)$  are known. Therefore, if species prevalence is available, we can calculate  $p$ ,  $r$ , and  $FPR$  from a non-traditional test set, and then plot the corrected ROC/PR curves.

**TABLE 2** A confusion matrix from a non-traditional test set.

	Reference			
	$s = 1$	$s = 0$		
Prediction	$y = 1$	$y = 1$	$y = 0$	Total
$y' = 1$	$m_1$	$m_2$	$m_3$	$k_1 = m_1 + m_2 + m_3$
$y' = 0$	$m_4$	$m_5$	$m_6$	$k_0 = m_4 + m_5 + m_6$
Total	$n_1 = m_1 + m_4$	$n_0 = m_2 + m_3 + m_5 + m_6$		$t = n_1 + n_0$

Note: numbers with shade are known, and numbers without shade are not known.

Here the key information is the species prevalence or the constant  $c$ . If one of them is known, the other one can be determined as well. In real-world applications, however, species prevalence and hence the constant  $c$  are usually unknown. Although species prevalence is normally unidentifiable without absence data, it can be estimated under certain assumptions or conditions (Lancaster & Imbens, 1996; Ward et al., 2009; Li et al., 2011; Royle et al., 2012; Hastie & Fithian, 2013; Phillips & Elith, 2013). Here we propose to estimate  $c$  from the ROC/PR curves under the condition that a model with good discrimination ability exists. When we increase the threshold to produce binary

prediction, the omission error will increase but commission error will decrease. If a model has a good ability to separate presence from absence data, we can set a high threshold to minimize commission error, so  $p$  will reach its maximum value of one and  $FPR$  will reach its minimum value of zero. The ROC/PR curves in Fig. 1 are produced by a model that satisfies the condition of good discrimination ability, from which we can observe that the ROC curve passes the positions with minimum  $FPR$  in the lower left corner and PR curve passes the positions with maximum  $p$  in the upper left corner.

According to Equations 9 and 10,  $p$  is a monotonically increasing function of  $p'$ , so we can find the maximum value of  $p'$  (i.e. the highest point in the PR curve or the most left point in the ROC curve) to infer the constant  $c$ . Because the maximum value of  $p'$  (denoted as  $p'_{\max}$ ) is the position where  $p = 1$ , we obtain  $c = p'_{\max}$  according to Equation 9 or 10. However, estimating  $c$  using a single point in the ROC/PR curve may result in a large variance, so we propose to select multiple points whose values of  $p'$  are relatively high to estimate  $c$ . Let  $PP$  be a subset of points in the ROC/PR curve whose values of  $p'$  fall within a range of user-specified percentiles. We have the following estimator:

$$c = \frac{1}{j} \sum_{i \in PP} p'_i \quad (28)$$

where  $j$  is the cardinality of  $PP$ . For example, we can select those points where  $p'$  falls between 90<sup>th</sup> and 99<sup>th</sup> percentiles across all possible thresholds. Once  $c$  is estimated, species prevalence can be estimated as well according to Equation 26 or 27.

## 2.2 Experimental design

In this section we investigate the effectiveness of the proposed method to correct ROC/PR curves from presence and background data, which is the case-control scenario

commonly used in SDM. We trained different models from presence and background data, and model performances were evaluated using a traditional test set with presence-absence data and a non-traditional test set with presence-background data, respectively. ROC/PR curves were plotted using three different approaches: standard presence-absence (PA) approach, presence-only (PO) approach by simply treating background data as absence data, and presence-background (PB) approach using the proposed method to calibrate the curves. Curves produced by PO and PB approaches were compared to the benchmark curves produced by PA approach. Because it is difficult to obtain reliable species absence data in reality, we used virtual species in our experiment, which has become a common approach to test models from different aspects (Hirzel et al., 2001; Li et al., 2011; Meynard & Kaplan, 2013; Duan et al., 2015). One-class classification of remote sensing imagery is similar to SDM in that classifiers are evaluated using the same approach, and it is possible to collect reliable absence data in image classification, so we also used a real aerial photograph to test the proposed method.

## 2.3 Dataset

We simulated three virtual species with different prevalence values following the procedure of Li et al. (2011). The conditional probability of species occurrence  $P(y = 1|x)$  was modeled using the following equation:

$$P(y = 1|x) = \frac{e^{b_0 + b_1x_1 + b_2x_2 + b_3x_3}}{1 + e^{b_0 + b_1x_1 + b_2x_2 + b_3x_3}} \quad (29)$$

where  $b_i$  is a coefficient (see Table 3) and  $x_i$  is an environmental variable. We considered three environmental variables in California, including annually average precipitation,

annually average temperature, and elevation, all of which were extracted from the WorldClim database (<https://worldclim.org/>) with a spatial resolution of 1 km (Fick & Hijmans, 2017). At each pixel we used a random number ( $0 \leq q < 1$ ) to generate realized binary labels, i.e. presence ( $y = 1$ ) if  $q < P(y = 1|x)$  or absence ( $y = 0$ ) if  $q \geq P(y = 1|x)$ . From the realized binary map, we drew a non-traditional training set and a non-traditional test set, separately, both of which contained case-control presence-background data. The number of presence data in the test set was 1000, whereas the number of presence data in the training set varied, including 10, 50, 100, 500, and 1000. The number of background data in the training/test set was five times of presence data. With virtual species, we actually know the true labels of random background data, so we also used them to constitute a traditional test set. The training and test sets were randomly realized ten different times, and the experimental results were averaged in our analysis.

**TABLE 3** Prevalence and coefficients of three virtual species.

Species	Prevalence	Coefficients			
		$b_0$	$b_1$	$b_2$	$b_3$
Spec1	0.1638	-10	-0.15	-2.5	28
Spec2	0.3298	0.5	-1.5	-8.5	18
Spec3	0.4471	-0.4	-0.8	-5	18

The aerial photograph in Li et al. (2021) was also used to test the proposed method. The image covers an extent of 500 m  $\times$  500 m in the city of El Cerrito in California, with a spatial resolution of 0.3 m. The total number of pixels is 2778889, and the prevalence values of urban, tree, and grass are 0.2292, 0.2106, and 0.1880, respectively.

We performed different one-class classifications to map different land types (i.e. urban, tree, and grass), separately, treating them as three different species. For each land type, we drew a non-traditional training set and a non-traditional test set, respectively, following the case-control sampling scheme. The number of presence data in the test set was 2000, and the number of presence data in the training set was set as 200, 1000, and 5000, respectively. The number of background data was five times of presence data in both training and test sets. Again, the true labels of background data in the non-traditional test set can be obtained through manual interpretation, so we also used these background data to constitute a traditional test set. Both the training and test sets were randomly realized ten different times, and the experimental results were averaged in our analysis.

We trained different classifiers using different sample sizes to produce different model performances. For convenience we refer to sample size as the number of labeled presence data in a training set throughout this paper. For the virtual species, we trained DOMAIN (Carpenter et al., 1993), generalized linear model (GLM) (Guisan et al., 2002), and artificial neural network (ANN) (Hecht-Nielsen, 1989) using five sample sizes (i.e. 10, 50, 100, 500, and 1000); for the aerial photograph, we trained GLM and ANN using three sample sizes (i.e. 200, 1000, and 5000) and convolutional neural network (CNN) (Lecun et al., 1998) with only one sample size (i.e. 5000). DOMAIN was trained from presence-only data whereas other classifiers were trained from presence-background data. All of the models were evaluated by a traditional test set and a non-traditional test set, respectively. We plotted ROC/PR curves and calculated AUC



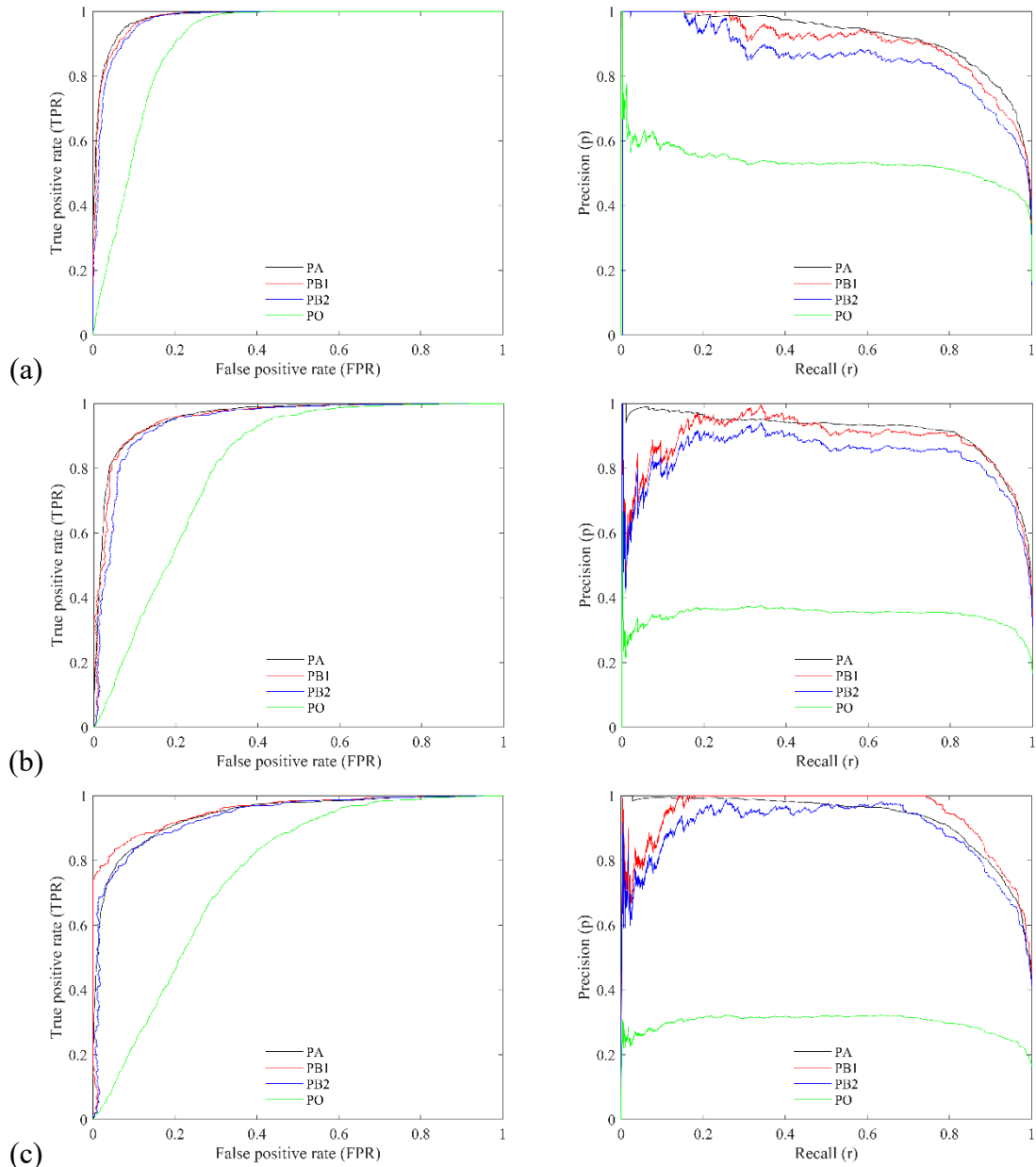
values using PA, PO, and PB approaches. For the PB approach, we tested two different scenarios: true species prevalence was given; and species prevalence was estimated from ROC/PR curves.

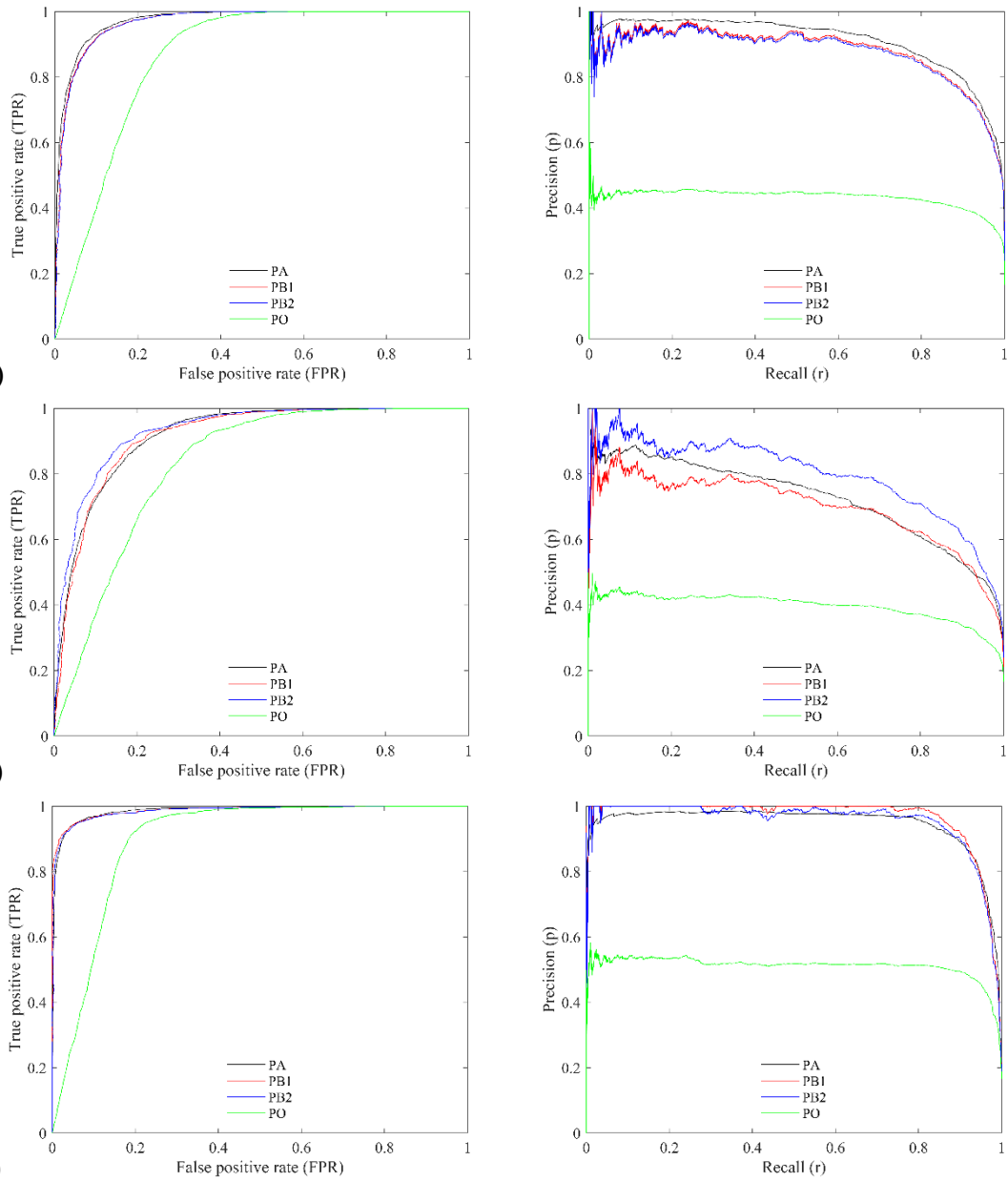
### 3. RESULTS

In Fig. 2 we present part of the ROC/PR curves by different approaches. Generally, the curves produced by PB approach are quite similar to the benchmark curves by PA approach, whereas the curves produced by PO approach are obviously lower than that by PA approach for all species. Meanwhile, the discrepancies between PR curves are obviously larger than that between ROC curves. According to Fig. 3, the rankings of models by AUC values are similar for PA, PB, and PO approaches, and the correlations of AUC values between different approaches are strong in general. The correlation coefficient of AUC between PA and PB is slightly higher than that between PA and PO approaches. For example, the correlation coefficient of AUCPR between PA and PB is 0.9789 when true prevalence is given or 0.9731 when prevalence is estimated, whereas the correlation coefficient of AUCPR between PA and PO is 0.9689 for the virtual species spec1.

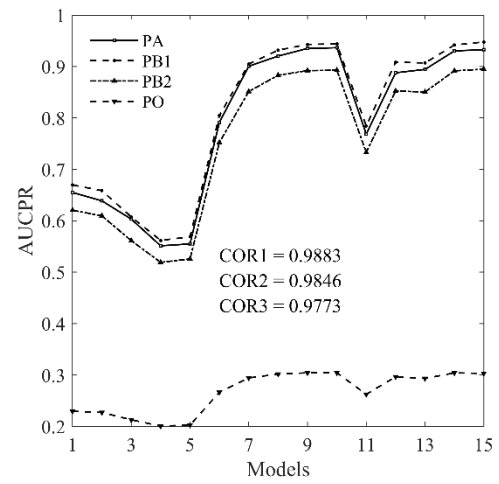
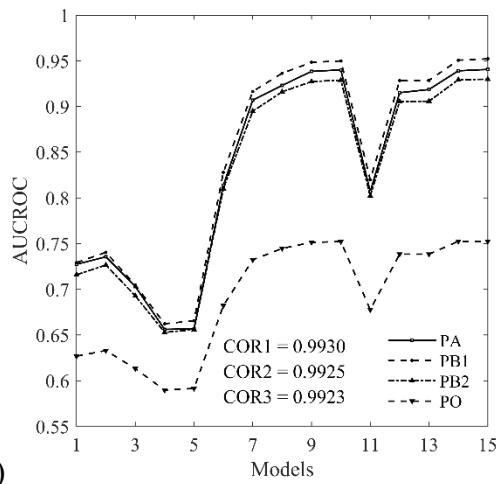
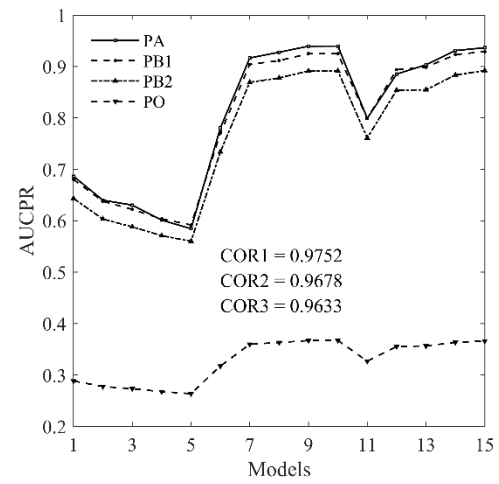
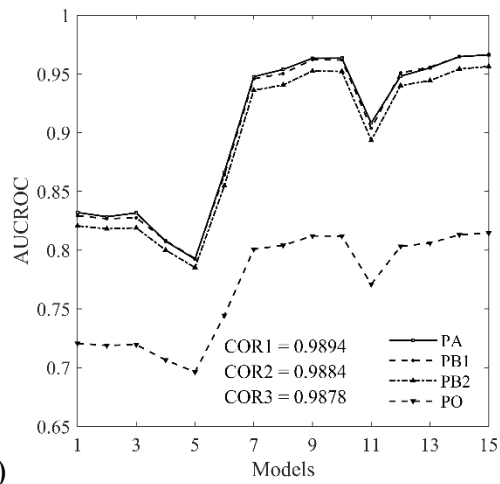
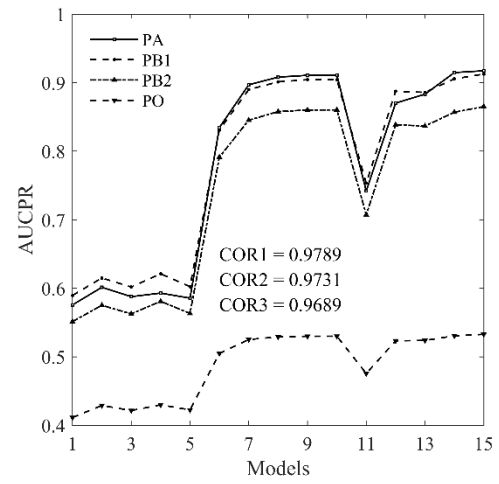
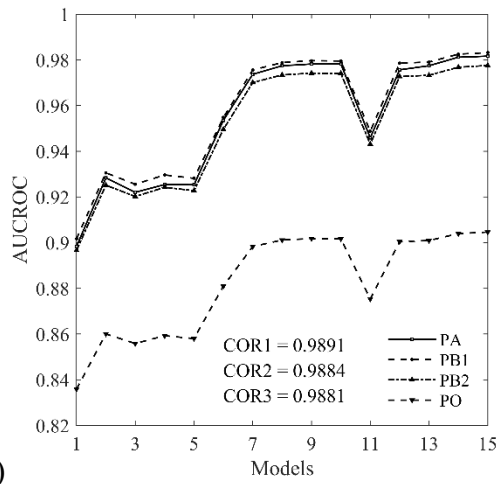
Based on the ranking of models by PO approach, we used the ROC (or PR) curve produced by the best model with the highest AUCROC (or AUCPR) value to estimate the constant  $c$  and prevalence, which are shown in Table 4. As can be seen, the accuracies of estimated prevalence and  $c$  are relatively high in most cases. For example, the true values of prevalence and  $c$  for urban are 0.2292 and 0.4660, respectively, and the estimated values are 0.2275 and 0.4678, respectively. For the virtual species, the

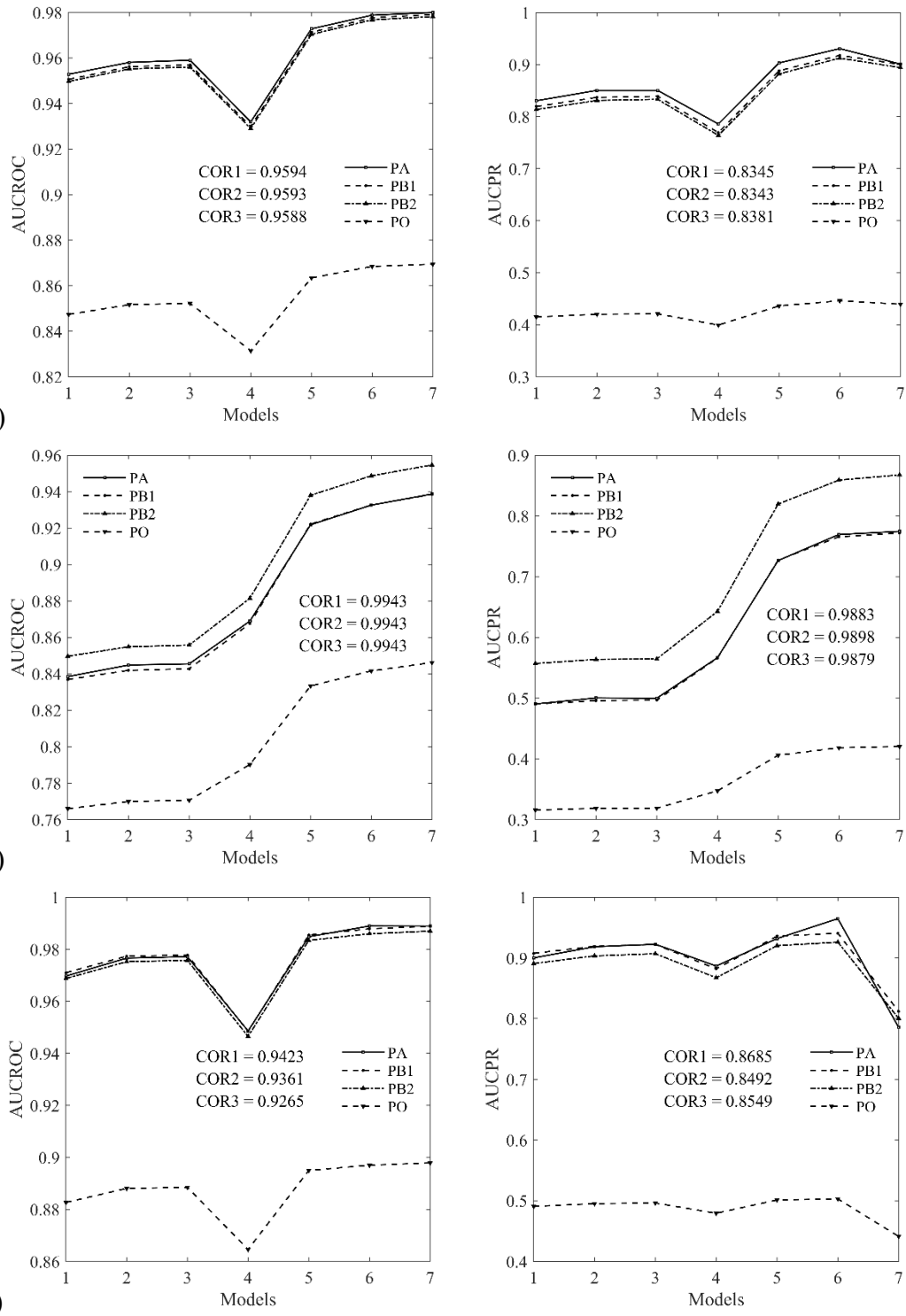
absolute value of relative error of prevalence ranges from 5%–8%, whereas the absolute value of relative error of  $c$  ranges from 3%–6%. For the aerial photograph, the absolute value of relative error of prevalence ranges from 0%–14%, whereas the absolute value of relative error of  $c$  ranges from 0%–7%. The highest accuracy is produced by classification of urban, and the largest error is produced by classification of tree from the real aerial photograph.





**Figure 2** The ROC (left) and PR (right) curves by PA, PB, and PO approaches. PB1: prevalence is given; PB2: prevalence is estimated. Model: ANN trained with a sample size of 1000. Virtual species: spec1 (a); spec2 (b); spec3 (c). Aerial photograph: urban (d); tree (e); grass (f).





**Figure 3** The average values of AUCROC (left) and AUCPR (right) over ten random realizations by PA, PB, and PO approaches. PB1: prevalence is given; PB2: prevalence is estimated. Virtual species: spec1 (a); spec2 (b); spec3 (c). Aerial photograph: urban (d); tree (e); grass (f). For virtual species, models 1–15 refer to DOMAIN with five

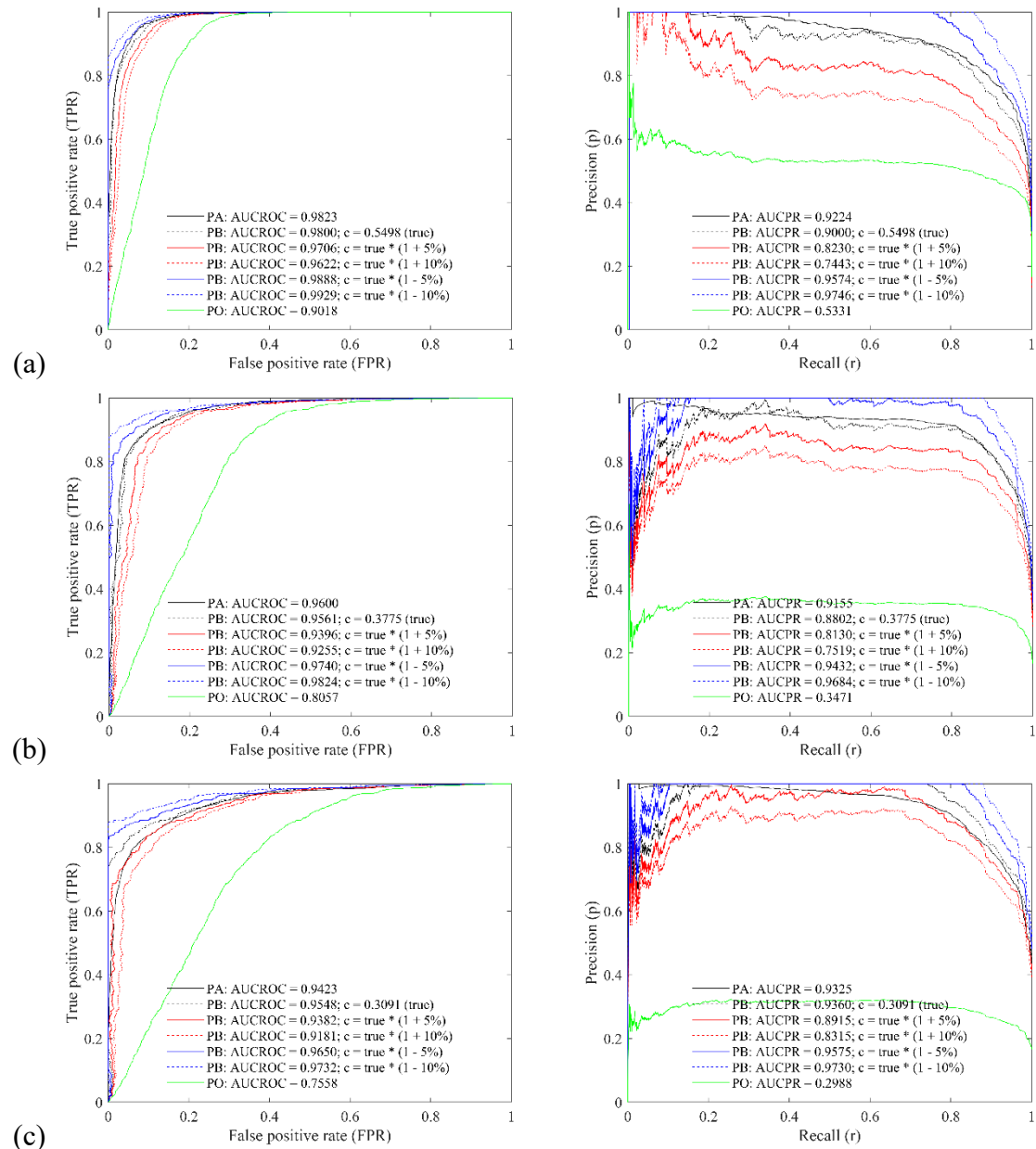
sample sizes, GLM with five samples, and ANN with five samples, respectively. For aerial photograph, models 1–7 refer to GLM with three sample sizes, ANN with three sample sizes, and CNN with one sample size, respectively.

**TABLE 4** The true and estimated values of prevalence and  $c$ .

Type	True value		Estimated value		Relative error (%)	
	$P(y = 1 x)$	$c$	$P(y = 1 x)$	$c$	$P(y = 1 x)$	$c$
Spec1	0.1638	0.5498	0.1531	0.5664	-6.53	3.02
Spec2	0.3298	0.3775	0.3117	0.3908	-5.48	3.54
Spec3	0.4471	0.3091	0.4128	0.3263	-7.66	5.59
Urban	0.2292	0.4660	0.2275	0.4678	-0.74	0.39
Tree	0.2106	0.4871	0.2394	0.4552	13.68	-6.56
Grass	0.1880	0.5154	0.1839	0.5210	-2.18	1.08

According to Table 4, the largest absolute value of relative error of estimated  $c$  in our experiment is 6.56%. The sensitivity of the calibrated curves by PB to the constant  $c$  is shown in Fig. 4, in which the ROC/PR curves are plotted using the true value of  $c$  with additive relative errors of  $\pm 10\%$ . We can see that the ROC curve moves rightward and PR curve moves downward when  $c$  is overestimated, and this trend switches to the opposite direction correspondently when  $c$  is underestimated. Consequently, AUCROC and AUCPR values are underestimated when  $c$  is overestimated, whereas they are overestimated when  $c$  is underestimated. Take virtual species spec1 as an example, the values of AUCROC and AUCPR for PA approach are 0.9823 and 0.9224, respectively; the values of AUCROC and AUCPR for PB approach are 0.9800 and 0.9000, respectively, when true value of  $c$  is given; the values of AUCROC and AUCPR for PB

approach are 0.9622 and 0.7443, respectively, when  $c$  with +10% relative error is given; the values of AUCROC and AUCPR for PB approach are 0.9929 and 0.9746, respectively, when  $c$  with -10% relative error is given. Meanwhile, the variation of PR curve by PB approach with different values of  $c$  is larger than the variation of ROC curve.



**Figure 4** The sensitivity of ROC (left) and PR (right) curves to constant  $c$  by PB

approach for the virtual species: spec1 (a); spec2 (b); spec3 (c). Model: ANN trained with a sample size of 1000.

#### 4. DISCUSSION

Developing novel methods to evaluate the performance of models without absence data is important in SDM since reliable absence data are usually not available in real-world applications. Currently, one of the most commonly used approaches to address the presence-only problem in SDM is to train models using presence and background data, which belongs to the case-control sampling scenario, and models are usually evaluated using ROC/PR plots by simply treating the background data as absence data (Phillips et al., 2006; Lobo et al., 2008; Peterson et al., 2008; Jiménez-Valverde, 2012; Sofaer et al., 2019). This PO approach can rank the models by the relative value of AUC, but the absolute value of AUC may be quite different from its true value and hence is misleading (Lobo et al., 2008; Sofaer et al., 2019). In this study, both the AUCROC and AUCPR are underestimated by PO approach in all cases. Take the classification of urban as an example, the AUCROC and AUCPR values by GLM trained with a sample size of 200 are 0.9529 and 0.8303 for the PA approach, but the AUC values become 0.8474 and 0.4148 for the PO approach. Please note that the ROC curve of a null model is a straight line connecting the points (0, 0) and (1, 1) in the ROC space, showing that  $TPR$  is equal to  $FPR$ . By contrast, the ROC curve of a trained model is usually higher than that of a null model, showing that  $TPR$  is larger than  $FPR$ . In other words,  $m_2 / (m_2 + m_5)$  is larger than  $m_3 / (m_3 + m_6)$  for a trained model according to Table 2. Consequently,  $(m_2 + m_3) / (m_2 + m_5 + m_3 + m_6)$  is larger than  $m_3 / (m_3 + m_6)$ , i.e.  $FPR'$



of PO approach is larger than  $FPR$  of PA approach. Meanwhile,  $TPR'$  (equivalent to  $r'$ ) of PO approach is equal to  $TPR$  (equivalent to  $r$ ) of PA approach. As a result, the ROC curve and AUCROC of a trained model by PO approach are usually lower than that by PA approach. Meanwhile, we can infer that  $p' = p / [p + (1 - c) / c]$  according to Equation 9. In this study  $c$  ranges from 0.3 to 0.55 and  $p$  ranges from zero to one, so  $p$  is larger than  $p'$  in most cases, which is the reason why the PR curve and AUCPR of a trained model by PO approach are also lower than that by PA approach in our experiment.

Unlike the PO approach that treats all of the background data as absence data, the PB approach acknowledges that background data contain both presence and absence data, and it infers the true performance of a model based on a constant  $c$ . According to our experimental results, the PB approach is effective in calibrating ROC/PR curves given that the true value of  $c$  is known. The curves and AUC values by PB approach are very similar to that by PA approach. In reality, however, the true value of  $c$  is usually unknown and hence it has to be estimated. According to Equation 9, an overestimate of  $c$  will result in an underestimate of  $p$ . Because  $p$  is negatively related to commission error whereas  $FPR$  is positively related to commission error, an underestimate of  $p$  will result in an overestimate of  $FPR$ . These are the reasons why the ROC/PR curves and AUC values are underestimated when  $c$  is overestimated. Since the largest absolute value of relative error of  $c$  is 6.56% in this study, we tested the sensitivity of the calibrated curves by PB to  $c$  with additive relative errors of  $\pm 10\%$ , and the results show that the curves and AUC values by PB with the largest relative error of  $c$  are still better than that by PO approach. Previous research has indicated that ROC curve and

AUCROC value may be inflated when a large number of  $TN$  data exist in a confusion matrix (Lobo et al., 2008). By contrast, PR curve does not consider  $TN$  data in a confusion matrix so it is more robust to geographic extent and species prevalence, but it has a more variable shape than ROC curve especially at the positions with low values of  $r$  (Boyd et al., 2012; Sofaer et al., 2019). Consistently, we can observe that AUCROC values are larger than AUCPR values, and ROC curves are generally more stable than PR curves. In the sensitivity analysis of  $c$ , the variation of ROC curve is smaller than that of PR curve probably because the effect of  $c$  is offset by a large number of  $TN$  data.

The proposed PB method to calibrate ROC/PR curves is based on the work of Li and Guo (2013). The omission error is related to  $r$  whereas commission error is related to  $p$ , both of which are quantified in ROC/PR plots. The relationship between  $r$  and  $r'$  and the relationship between  $p$  and  $p'$  derived in Li and Guo (2013) are used to reconstruct the true ROC/PR curves from presence and background data. The key information of this PB method is the constant  $c$  or species prevalence. Although true species prevalence is regarded as unidentifiable without absence data, an estimation of prevalence is helpful and possible under certain conditions (Li et al., 2011; Royle et al., 2012; Hastie & Fithian, 2013; Phillips & Elith, 2013). Please note that  $r$  is equal to  $r'$ , and the relationship between  $p$  and  $p'$  is similar to the relationship between probability of species occurrence and a naive model fitted from presence-background data. Let  $f = P(y = 1|x)$  denote probability of species occurrence and  $f' = P(s = 1|x)$  denote a naive model. We have  $f = f' / c$  and  $p = p' / c$  in the single-training-set scenario, or  $f = (1 - c) / c \times f' / (1 - f')$  and  $p = (1 - c) / c \times p' / (1 - p')$  in the case-control scenario (Elkan

& Noto, 2008; Li et al., 2011). Therefore, an estimation of  $c$  (or prevalence) not only enables us to model the probability of species occurrence, but also helps us to assess the model performance without requiring labeled absence data.

There are several ways to estimate the constant  $c$  (or prevalence). Li et al. (2011) proved that the average predicted values of  $f'$  at prototypical presence sites where the habitats are maximally suitable for a species can be used to estimate  $c$ , but this approach may lead to an underestimate of  $c$  because the probability of species occurrence at a selected prototypical presence site may be smaller than one. Royle et al. (2012) proposed the MAXLIKE that can be used to infer prevalence, but the linear logistic model assumption may be violated in reality (Hastie & Fithian, 2013; Phillips & Elith, 2013). Li and Guo (2013) showed that thresholding a naive model based on maximizing the measure  $F_{pb}$  on a validation set can also estimate prevalence, but Liu et al. (2016) argued that it is difficult to estimate prevalence using threshold-based approach. In this study, we propose to estimate  $c$  from ROC/PR plots based on the fact that a model of good discrimination ability can produce the maximum value of  $p$  (or minimum value of  $FPR$ ) with a high value of threshold. In other words, we can adjust the PR curve so that its highest point reaches the maximum value of one (equivalent to adjusting the ROC curve so that its most left point reaches the minimum value of zero), and the relationship between  $p$  and  $p'$  yields an estimate of  $c$ . Because there could be multiple threshold values that can produce maximum value of  $p$ , we use multiple points rather than the highest point in the PR curve to obtain a more robust estimate of  $c$ , e.g. the higher threshold values between 90<sup>th</sup> to 99<sup>th</sup> percentiles. In our experiment, this

percentile range consistently produces high accuracies of  $c$  for different species, and it can be adjusted in other situations. Meanwhile, the discrimination ability of a model can affect the accuracy of  $c$ . The largest relative error of  $c$  is observed for the classification of tree because the model cannot perform well in discriminating tree from other land types, with the lowest value of AUCPR compared with other species. If possible, users can also derive  $c$  (or prevalence) from other sources such as limited presence–absence surveys or expert knowledge (Phillips & Elith, 2013). However, the uncertainty of  $c$  is almost inevitable no matter it is derived from models or surveys.

In this study we only focus on calibrating the ROC/PR curves from presence and background data, but model evaluation may involve multiple aspects and multiple measures. The strengths and drawbacks of ROC/PR plots have been well investigated in the literature (Fielding & Bell, 1997; Davis & Goadrich, 2006; Lobo et al., 2008; Boyd et al., 2012; Sofaer et al., 2019). For example, current ROC/PR plots have been criticized to equally weigh the commission and omission errors, but these two types of errors may not be of the same importance (Lobo et al., 2008; Peterson et al., 2008). In spite of the limitations of ROC/PR plots, the proposed method can be used as a complement to other model assessment methods. For example, the presence-only calibration (POC) plot by Phillips and Elith (2010) can be used to measure the ability of calibration for continuous outputs. The AVI and CVI in Hirzel et al. (2006),  $F_{pb}$  and  $F_{cpb}$  in Li and Guo (2013), and TSS in Liu et al. (2013) can assess the accuracy of binary outputs without requiring absence data. Other methods such as Boyce index and the compositional and multinomial procedure can also be considered to quantify model

performance from different aspects when absence data are not available (Boyce et al., 2002; Ottaviani et al., 2004).

Similar to the proposed PB-based ROC/PR plots, both the POC plot and  $F_{cpb}$  also require additional information on the constant  $c$  (or species prevalence) (Phillips & Elith, 2010; Li & Guo, 2013). Actually, the four quadrants of a confusion matrix can be fully determined from presence and background data if  $c$  is known, and all of the accuracy measures derived from a confusion matrix can be calculated, which should be investigated in future research. Although different approaches to estimate  $c$  still have their limitations, such attempts are necessary. In practice, users can consider applying multiple approaches to reduce the uncertainty of  $c$ . Since  $c = n_1 / [n_1 + n_0 \times P(y = 1)]$  in the case-control scenario,  $n_1 / (n_1 + n_0) \leq c \leq 1$  because  $0 \leq P(y = 1) \leq 1$ . If we can provide a rough estimate of prevalence such as from survey or expert knowledge, then the range of  $c$  can be refined. Meanwhile, the estimator derived from prototypical presences in Li et al. (2011) usually underestimates  $c$ , so it can be used as the lower bound of  $c$ .

In this study the ROC/PR curves plotted from PA approach are used to benchmark the curves from PO and PB approaches, so a test set with presence-absence data are required. Because we do not have real species datasets with reliable absence data, we only tested the proposed method using virtual species datasets. In addition, we used a real aerial photograph since model evaluation of binary classification in remote sensing is mathematically similar to model evaluation in SDM, and both presence and absence data can be obtained through manual interpretation of the aerial photography. However,

there are still some differences between remote sensing classification and SDM due to the complicated biological processes like biotic interactions (Warren et al., 2020). Therefore, the proposed PB method should be further investigated using real species datasets in the future when reliable species absence data are available.

## 5. CONCLUSION

In this study we propose a new PB approach to plot ROC/PR curves from presence-background data with additional information of a constant  $c$  (or species prevalence). The accuracy measures  $r$  and  $p$  derived from presence-absence data can be connected to  $r'$  and  $p'$  derived from presence-background data through the constant  $c$ , which enables reconstructing the true ROC/PR curves from presence-background data. Meanwhile,  $c$  can be estimated from ROC/PR plots under the condition that a model of good discrimination ability exists. Our experimental results demonstrate that the proposed PB approach is effective both in plotting ROC/PR curves and estimating  $c$  from presence-background data in the case-control sampling scenario.

## ACKNOWLEDGEMENTS

## AUTHOURS CONTRIBUTIONS

W.L. and Q.G. conceived the ideas and designed methodology; W.L. collected data, performed statistical analyses, and led the writing of the manuscript; Both authors contributed critically to the drafts and gave final approval for publication.

## DATA AVAILABILITY

The data used to simulated virtual species are available at WorldClim database (<https://worldclim.org/>). Data will be archived in Dryad upon acceptance of this manuscript.

## REFERENCES

- Aarts, G., Fieberg, J., & Matthiopoulos, J. (2012). Comparative interpretation of count, presence–absence and point methods for species distribution models. *Methods in Ecology and Evolution*, 3(1), 177–187. <https://doi.org/10.1111/j.2041-210X.2011.00141.x>
- Boyce, M. S. et al. (2002). Evaluating resource selection functions. *Ecological Modelling*, 157(2), 281–300. [https://doi.org/10.1016/S0304-3800\(02\)00200-4](https://doi.org/10.1016/S0304-3800(02)00200-4)
- Boyd, K. et al. (2012). Unachievable region in precision-recall space and its effect on empirical evaluation.
- Busby, J. R. (1986). A biogeoclimatic analysis of *Nothofagus cunninghamii* (Hook.) Oerst. in southeastern Australia. *Austral Ecology*, 11(1), 1–7.
- Carpenter, G., Gillison, A. N., & Winter, J. (1993). DOMAIN: a flexible modelling procedure for mapping potential distributions of plants and animals. *Biodiversity and Conservation*, 2(6), 667–680.
- Congalton, R. G. (1991). A review of assessing the accuracy of classifications of remotely sensed data. *Remote Sensing of Environment*, 37(1), 35–46.
- Davis, J., & Goadrich, M. (2006). The relationship between Precision-Recall and ROC curves. *Proceedings of the 23rd international conference on Machine learning*,

577 pp. 233–240. Association for Computing Machinery, Pittsburgh, Pennsylvania,  
 578 USA.

579 Duan, R.-Y. et al. (2015). SDMvspecies: a software for creating virtual species for  
 580 species distribution modelling. *Ecography*, 38(1), 108–110.  
 581 <https://doi.org/10.1111/ecog.01080>

582 Elith, J. et al. (2006). Novel methods improve prediction of species' distributions from  
 583 occurrence data. *Ecography*, 29(2), 129–151.

584 Elkan, C., & Noto, K. (2008). Learning classifiers from only positive and unlabeled  
 585 data. *Proceeding of the 14th ACM SIGKDD international conference on*  
 586 *Knowledge discovery and data mining*, pp. 213–220. ACM, Las Vegas, Nevada,  
 587 USA.

588 Fick, S. E., & Hijmans, R. J. (2017). WorldClim 2: new 1-km spatial resolution climate  
 589 surfaces for global land areas. *International Journal of Climatology*, 37(12),  
 590 4302–4315. <https://doi.org/10.1002/joc.5086>

591 Fielding, A. H., & Bell, J. F. (1997). A review of methods for the assessment of  
 592 prediction errors in conservation presence/absence models. *Environmental*  
 593 *Conservation*, 24(1), 38–49. doi:10.1017/S0376892997000088

594 Guisan, A., Edwards, T. C., & Hastie, T. (2002). Generalized linear and generalized  
 595 additive models in studies of species distributions: setting the scene. *Ecological*  
 596 *Modelling*, 157(2-3), 89–100.

597 Guisan, A., & Thuiller, W. (2005). Predicting species distribution: offering more than  
 598 simple habitat models. *Ecology Letters*, 8(9), 993–1009. 10.1111/j.1461-



0248.2005.00792.x

Hagar, J. C., Yost, A., & Haggerty, P. K. (2020). Incorporating LiDAR metrics into a structure-based habitat model for a canopy-dwelling species. *Remote Sensing of Environment*, 236. <https://doi.org/10.1016/j.rse.2019.111499>

Hastie, T., & Fithian, W. (2013). Inference from presence-only data; the ongoing controversy. *Ecography*, 36(8), 864–867. <https://doi.org/10.1111/j.1600-0587.2013.00321.x>

Hecht-Nielsen, R. (1989) Theory of the backpropagation neural network. *International Joint Conference on Neural Networks, 1989. IJCNN.*, pp. 593–605 vol.591.

Hirzel, A. H. et al. (2002). Ecological-niche factor analysis: how to compute habitat-suitability maps without absence data? *Ecology*, 83(7), 2027–2036.

Hirzel, A. H., Helfer, V., & Metral, F. (2001). Assessing habitat-suitability models with a virtual species. *Ecological Modelling*, 145(2-3), 111–121.

Hirzel, A. H. et al. (2006). Evaluating the ability of habitat suitability models to predict species presences. *Ecological Modelling*, 199(2), 142–152. <https://doi.org/10.1016/j.ecolmodel.2006.05.017>

Jiménez-Valverde, A. (2012). Insights into the area under the receiver operating characteristic curve (AUC) as a discrimination measure in species distribution modelling. *Global Ecology and Biogeography*, 21(3/4), 498–507. [10.1111/j.1466-8238.2011.00683.x](https://doi.org/10.1111/j.1466-8238.2011.00683.x)

Jiménez-Valverde, A. et al. (2013). Discrimination capacity in species distribution models depends on the representativeness of the environmental domain. *Global*

621 *Ecology and Biogeography*, 22(4), 508–516. <https://doi.org/10.1111/geb.12007>

622 Keating, K. A., & Cherry, S. (2004). Use and interpretation of logistic regression in  
623 habitat-selection studies. *Journal of Wildlife Management*, 68(4), 774–789.

624 Kueppers, L. M. et al. (2005). Modeled regional climate change and california endemic  
625 oak ranges. *Proceedings of the National Academy of Sciences of the United*  
626 *States of America*, 102, 16281–16286.

627 Lancaster, T., & Imbens, G. (1996). Case-control studies with contaminated controls.  
628 *Journal of Econometrics*, 71(1-2), 145–160.

629 Lecun, Y. et al. (1998). Gradient-based learning applied to document recognition.  
630 *Proceedings of the IEEE*, 86(11), 2278–2324. [10.1109/5.726791](https://doi.org/10.1109/5.726791)

631 Leroy, B. et al. (2018). Without quality presence–absence data, discrimination metrics  
632 such as TSS can be misleading measures of model performance. *Journal of*  
633 *Biogeography*, 45(9), 1994–2002. <https://doi.org/10.1111/jbi.13402>

634 Li, W., & Guo, Q. (2013). How to assess the prediction accuracy of species presence–  
635 absence models without absence data? *Ecography*, 36, 788–799.  
636 [10.1111/j.1600-0587.2013.07585.x](https://doi.org/10.1111/j.1600-0587.2013.07585.x)

637 Li, W., Guo, Q., & Elkan, C. (2011). Can we model the probability of presence of  
638 species without absence data? *Ecography*, 34(6), 1096–1105. [10.1111/j.1600-](https://doi.org/10.1111/j.1600-0587.2011.06888.x)  
639 [0587.2011.06888.x](https://doi.org/10.1111/j.1600-0587.2011.06888.x)

640 Li, W., Guo, Q., & Elkan, C. (2021). One-class remote sensing classification from  
641 positive and unlabeled background Data. *IEEE Journal of Selected Topics in*  
642 *Applied Earth Observations and Remote Sensing*, 14, 730–746.

10.1109/JSTARS.2020.3025451

Liu, C., Newell, G., & White, M. (2016). On the selection of thresholds for predicting species occurrence with presence-only data. *Ecology and Evolution*, 6(1), 337–348. <https://doi.org/10.1002/ece3.1878>

Liu, C., White, M., & Newell, G. (2011). Measuring and comparing the accuracy of species distribution models with presence–absence data. *Ecography*, 34(2), 232–243. 10.1111/j.1600-0587.2010.06354.x

Liu, C., White, M., & Newell, G. (2013). Selecting thresholds for the prediction of species occurrence with presence-only data. *Journal of Biogeography*, 40(4), 778–789. <https://doi.org/10.1111/jbi.12058>

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(2), 145–151.

Marmion, M. et al. (2009). The performance of state-of-the-art modelling techniques depends on geographical distribution of species. *Ecological Modelling*, 220(24), 3512–3520. <https://doi.org/10.1016/j.ecolmodel.2008.10.019>

Meynard, C. N., & Kaplan, D. M. (2013). Using virtual species to study species distributions and model performance. *Journal of Biogeography*, 40(1), 1–8. <https://doi.org/10.1111/jbi.12006>

Ottaviani, D., Lasinio, G. J., & Boitani, L. (2004). Two statistical methods to validate habitat suitability models using presence-only data. *Ecological Modelling*, 179, 417–443.

665 Peterson, A. T., & Holt, R. D. (2003). Niche differentiation in Mexican birds: using  
 666 point occurrences to detect ecological innovation. *Ecology Letters*, 6(8), 774–  
 667 782. 10.1046/j.1461-0248.2003.00502.x  
 668 Peterson, A. T., Papeş, M., & Soberón, J. (2008). Rethinking receiver operating  
 669 characteristic analysis applications in ecological niche modeling. *Ecological*  
 670 *Modelling*, 213(1), 63–72. <https://doi.org/10.1016/j.ecolmodel.2007.11.008>  
 671 Phillips, S. J., Anderson, R. P., & Schapire, R. E. (2006). Maximum entropy modeling  
 672 of species geographic distributions. *Ecological Modelling*, 190(3-4), 231–259.  
 673 Phillips, S. J., & Elith, J. (2010). POC plots: calibrating species distribution models  
 674 with presence-only data. *Ecology*, 91(8), 2476–2484.  
 675 Phillips, S. J., & Elith, J. (2013). On estimating probability of presence from use–  
 676 availability or presence–background data. *Ecology*, 94(6), 1409–1419.  
 677 <https://doi.org/10.1890/12-1520.1>  
 678 Royle, J. A. et al. (2012). Likelihood analysis of species occurrence probability from  
 679 presence-only data for modelling species distributions. *Methods in Ecology and*  
 680 *Evolution*, 3(3), 545–554. <https://doi.org/10.1111/j.2041-210X.2011.00182.x>  
 681 Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than  
 682 the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLOS*  
 683 *ONE*, 10(3), e0118432. 10.1371/journal.pone.0118432  
 684 Sofaer, H. R., Hoeting, J. A., & Jarnevich, C. S. (2019). The area under the precision–  
 685 recall curve as a performance metric for rare binary events. *Methods in Ecology*  
 686 *and Evolution*, 10(4), 565–577. <https://doi.org/10.1111/2041-210X.13140>

687 Ward, G. et al. (2009). Presence-only data and the EM algorithm. *Biometrics*, 65(2),  
688 554–563.

689 Warren, D. L., Matzke, N. J., & Iglesias, T. L. (2020). Evaluating presence-only species  
690 distribution models with discrimination accuracy is uninformative for many  
691 applications. *Journal of Biogeography*, 47(1), 167–180.  
692 <https://doi.org/10.1111/jbi.13705>

693

694