Environmental Systems Research

RESEARCH Open Access

Performance of inhomogeneous Poisson point process models under different scenarios of uncertainty in species presence-only data

Yannick Mugumaarhahama[1,2*], Adandé Belarmain Fandohan[1,3] and Romain L. Glèlè Kakaï[1]

Abstract

Haphazard and opportunistic species occurrence (PO) data are widely used in species distribution models (SDMs) instead of high-quality species data gathered using appropriate and structured sampling methods, which is expensive and often spatially limited. Despite their widespread use in ecology, PO data are prone to errors and uncertainties, such as imperfect detectability, positional imprecision, and spatial niche truncation, which make their use analytically challenging for effective and adaptive biodiversity management and conservation. Using simulated data, this study investigates the effects of these uncertainties on the performance of spatial point process based presence-only and integrated SDMs. We investigated three SDMs in this study, one that ignores imperfect detectability: the presence-only model (PO model), and two that account for it: the thinned presence-only model (THINPO model) and the integrated model (PBPC model). The ability of these SDMs to produce accurate maximum likelihood estimates of intensity model coefficients and reliable predictions of species distributions under different data quality scenarios was investigated. The results show that SDMs that account for imperfect detectability (THINPO or PBPC models) are not applicable in situations of high detectability. In this situation, the PO model produces the most accurate maximum likelihood estimates of the models' coefficients ($\hat{\beta}_k$), and consequently the most accurate predictions of species distributions ($\hat{\lambda}(s)$). The effects of positional uncertainty and spatial niche truncation on this SDM output are minimal. However, in situations of low detectability, it is preferable to use the PBPC model. Positional uncertainty and spatial niche truncation have negligible effects on the output of this SDM, except when positionally uncertain PO data are analyzed along with truncated PC data. These minimal effects of spatial niche truncation on SDM outputs demonstrate the transferability of SDMs. However, the effects of all these uncertainties may depend on the characteristics of the species. Prior to modeling species distributions, a multivariate environmental similarity surface analysis should be performed to test the similarity between data from the restricted region to be used for model calibration and data from the entire range. If this analysis reveals dissimilarities, larger spatial and ecological scales should be considered to address the issue of spatial niche truncation. Further efforts could address the effects of species characteristics on SDMs performance and assess the effects of species-specific uncertainties.

Keywords Imperfect detection, Sampling bias, Positional uncertainty, Species distribution models, Poisson point process, Data integration, Data simulation

*Correspondence:
Yannick Mugumaarhahama
lesmas2020@gmail.com
Full list of author information is available at the end of the article

Introduction

In recent decades, biodiversity and ecosystems have come under intense pressure from global changes that are significantly altering species composition and distribution (Dawson et al. 2011). Thus, understanding the spatial distribution of species and the underlying environmental factors is a fundamental question for a wide range of ecological, evolutionary and conservation applications (Guillera-Arroita 2017; Inman et al. 2021). Then, tools that quantify changes in species distributions are of great importance (Rosenberg et al. 2019; Inman et al. 2021).

Species distribution models (SDMs) are quantitative tools and statistical modeling approaches widely used in ecology to map out suitable habitat for species and to assess the potential impact of climate change on their ecological niche (Guisan et al. 2013; Inman et al. 2021). SDMs are widely used to delineate priority areas for effective management and conservation of species (Franklin 2013). Accurate species distribution maps are essential for the implementation of sustainable conservation plans (Jarnevich et al. 2015). This requires using high quality species data collected using appropriate survey methods and structured sampling methods and tools (Osborne and Leitão 2009; Duputié et al. 2014; Moudrý et al. 2017).

Unfortunately, high quality data are very costly and often spatially limited (truncated) (Osborne and Leitão 2009; Duputié et al. 2014; Moudrý et al. 2017; Inman et al. 2021). This precludes their use, resulting in reliance on the widely available presence-only (PO) data that are collected haphazardly and opportunistically (Inman et al. 2021; Suhaimi et al. 2021). Although PO data are widely available, the way they are collected introduces errors and uncertainties that make their use analytically challenging (Graham et al. 2008; Inman et al. 2021; Suhaimi et al. 2021). Consequently, they rarely meet the assumptions of SDMs.

First, while SDMs assume that sampling effort is uniform across the landscape, and that the species niche is sampled across the full range of environmental conditions in which it occurs (Phillips et al. 2009; Hastie and Fithian 2013). PO data are susceptible to sampling bias resulting from unsystematic field surveys, biased data collection from relatively accessible areas, or biased sampling effort (Graham et al. 2004; Hortal et al. 2007; Syfert et al. 2013). Even in areas where occurrence data are collected, individuals of the species may be present but undetected introducing the bias of imperfect detection (Yoccoz et al. 2001; Dorazio 2012; Chen et al. 2013).

Second, SDMs assume that the species data encompass the entire realized niche of the species (covering broad environmental gradients) (Elith and Leathwick 2009; Phillips et al. 2009; Chevalier et al. 2021). In many ecological applications, this assumption is violated because study areas are primarily defined by geographic or political boundaries that cover only a subset of a species' realized niche (Hannemann et al. 2016; El-Gabbas and Dormann 2018). Thus, the realized niche is said to be truncated, and this can significantly degrade SDM predictions (Thuiller et al. 2004; Chevalier et al. 2021). Surprisingly, only a handful of studies have examined the effects of spatial niche truncation (see, Pearson et al. 2004; Thuiller et al. 2004; Barbet-Massin et al. 2010; Mateo et al. 2019).

Third, positional measurement inaccuracies, digitization errors, georeferencing problems, and operator error all contribute to high positional uncertainty in PO data (Graham et al. 2004, 2008; Naimi et al. 2011; Rocchini et al. 2011). While some studies have concluded that SDMs are generally insensitive to variations in the level of positional uncertainty (Graham et al. 2008; Fernandez et al. 2009; Mitchell et al. 2017), others have reached the opposite conclusion (Visscher 2006; Johnson and Gillingham 2008; Naimi et al. 2011, 2014). As a result, there is no consensus on its impact on SDM predictions.

Despite widespread criticism of the use of PO data in SDM, they are still widely used in ecology. This makes the use of integrated SDMs to be on the rise (Suhaimi et al. 2021). Integrated SDMs are a recent innovation that combine the spatial point process model approach (which includes inhomogeneous Poisson point process models, see Warton and Shepherd (2010) and Renner et al. (2015) for further details) with the hierarchical model approaches. They incorporate PO data and higher quality data (point count data or site occupancy data) into the same model (Dorazio 2014; Koshkina et al. 2017) to model species distributions, taking advantage of each data source while accounting for their respective limitations (Koshkina et al. 2017). However, another challenge lies in the appropriate use of these modeling approaches (Schank et al. 2019).

To assist ecologists in appropriately modeling species distribution based on the PO data, we must examine the performance of existing modeling approaches under a variety of scenarios of data quality (Suhaimi et al. 2021). Mugumaarhahama et al. (2022) show that in the context of sampling bias and imperfect detection in PO data, the best results are obtained by analysing PO data in conjunction with point count (PC) data using the approach introduced by Dorazio (2014). However, the extent to which the PO data used, which are susceptible to the aforementioned uncertainties, may affect the effectiveness of PO models and the integrated SDM is not yet well understood.

The focus of this study is on the use of PO data in SDM through the use of spatial point process models. This research aims to assess the impacts of aforementioned uncertainties in data on performance of SDMs. We use a virtual ecologist approach (Zurell et al. 2010; Suhaimi et al. 2021), in which we simulate the distribution of a virtual species and sample it under different conditions of data quality. In our simulations, we consider different scenarios in which a modeler has both PO data of varying quality and PC data that fully or partially cover the range of the virtual species, and must decide whether or not to use the two datasets. Specifically, we assess the marginal and combined effects of both positional uncertainty in PO data and data truncation in PO and/or PC data on the performance of PO and integrated SDMs under conditions of low and high species detectability.

Methods

To assess SDMs for factors that might influence their performance, the use of real data could lead to erroneous conclusions (Meynard and Kaplan 2013; Miller 2014; Leroy et al. 2016). However, the use of simulated virtual species has the advantage that the “true” distribution of the species is completely known and the variables that influence this distribution are all known (Hirzel et al. 2001; Zurell et al. 2010; Meynard and Kaplan 2013; Leroy et al. 2016). The main strength of this approach is the ability to compare the model output to a known (virtual) “truth”.

Simulation study

This study assesses the performance of PO models and integrated models under different scenarios of data quality. We explored the performance of these SDMs by examining their ability to estimate key parameters from simulated data whose characteristics are known. The use of virtual species, whose distributions are uniquely determined by a set of simulated environmental factors, ensures that the suitability of all species at each site is strictly determined by these factors, without additional biotic or dispersal restrictions. By simulating the distribution of the virtual species and introducing various biases into the data, and then refitting the models with these data, the resulting parameter estimates can be compared with the initial parameters of the “true” distribution and thus determine the effects of the factors under study (Hirzel et al. 2001; Zurell et al. 2010; Meynard and Kaplan 2013; Leroy et al. 2016). Figure 1 depicts the simulation framework.

Generating virtual species range

For the data generation process, the simulation design was similar to that described in Dorazio (2014) and Koshkina et al. (2017). We assumed that individuals of the virtual species reside within a 2D grid B which is assumed to be a square divided into 1000×1000 grid cells.

Two environmental covariates, $x(s)$ and $w(s)$, were generated using bivariate distributions that vary spatially and being independent of each other. The bivariate distributions were chosen so that both environmental covariates are defined at every point s of B [(Dorazio (2014) and Koshkina et al. (2017) are recommended reading for a more in-depth understanding].

Considering that n individuals of the virtual species reside within B, PO data are a set $s = s_1, s_2, s_3, \ldots, s_n$ of point locations in B, where these individuals are recorded. It is assumed that the activity centers of observed individuals are a realization of a Poisson point process parameterized by a first-order intensity function $\lambda(s)$ (Dorazio 2014). The process that characterizes the presence-only data is inhomogeneous because the intensity $\lambda(s)$ varies with location depending on environmental covariates hypothesized to influence or define potential habitat of species (Franklin 2010). In this study, we used a log-linear function that depends on a single covariate $x(s)$ to generate the intensity surface over B, which represents the “true” spatial patterns of the virtual species distributions as follows:

$$\log(\lambda(s)) = \beta_0 + \beta_1 x(s) \tag{1}$$

We considered $\beta_0 = \log(1000) \approx 6.91$ and $\beta_1 = 0.5$. With these arbitrarily chosen values, the lower the values of $x(s)$, the lower the virtual species intensity and vice-versa.

Sampling PO data of the virtual species with different scenarios of uncertainty

To simulate different scenarios of data quality, we introduced errors and uncertainties in simulated PO data. As species occurrence data are prone to imperfect detectability that includes sampling bias and imperfect detection, only m of the n individuals are observed. Due to imperfect detectability, PO data are $s = s_1, s_2, s_3, \ldots, s_m$ of point locations of observed individuals in B $(m < n)$. m, the number of observed individuals depends on the thinned intensity $\lambda(s)b(s)$ (Dorazio 2014; Koshkina et al. 2017). $b(s)$ was simulated depending on covariate $w(s)$ using the following equation:

$$\text{logit}\,(b(s)) = \alpha_0 + \alpha_1 w(s) \tag{2}$$

In this study, we considered arbitrarily $\alpha_1 = -1.5$ and $\alpha_0 = -1.1$ for the low detectability scenario and $\alpha_0 = 4.6$ for the high detectability scenario. With these arbitrarily chosen values, the mean detectability, $\bar{b}(s) \approx 0.33$ for the low detectability scenario while the mean detectability,

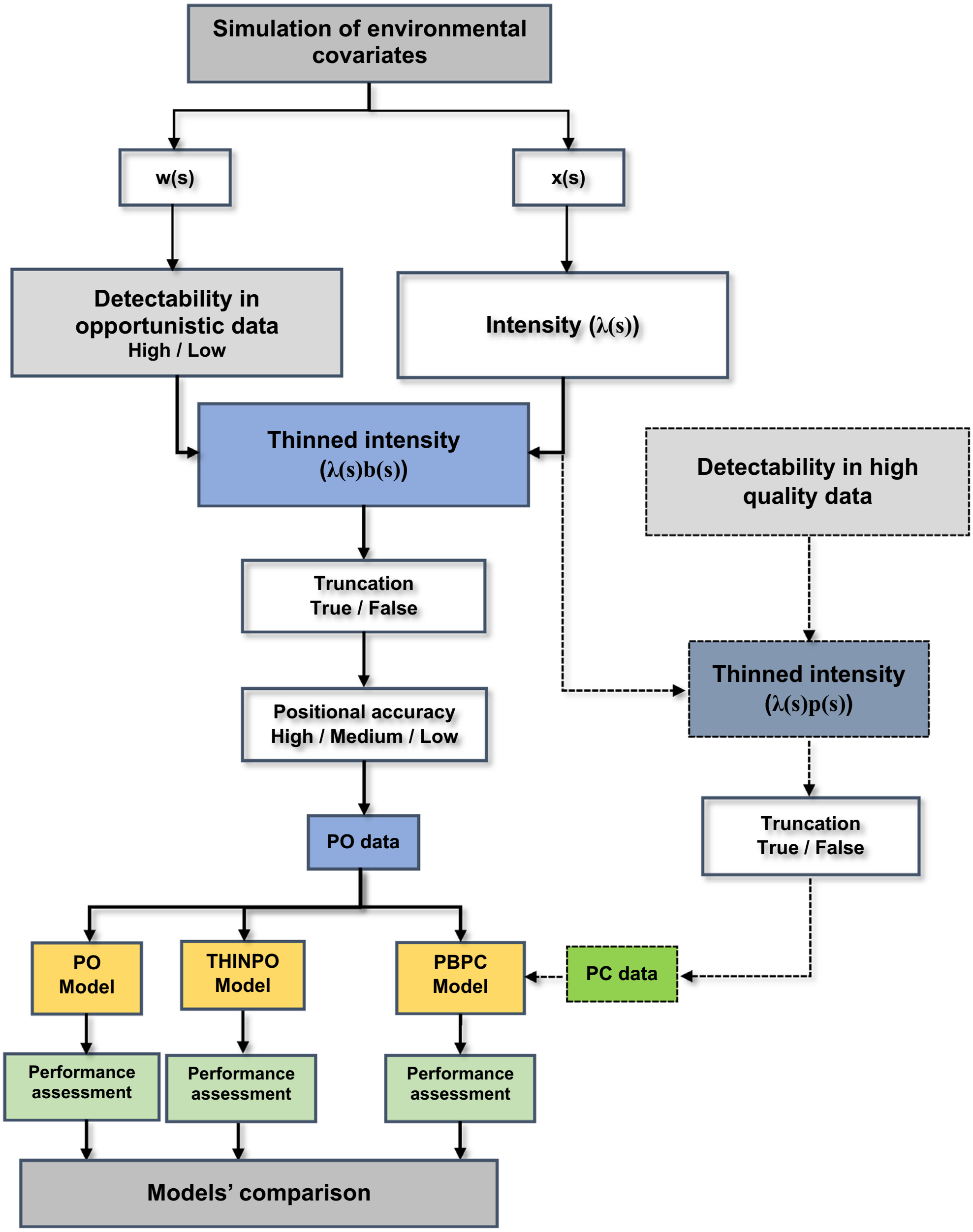

Fig. 1 General framework of the simulation process

$\bar{b}(s) \approx 0.98$ for the high detectability scenario. Note that $b(s)$ includes both, sampling bias and imperfect detection. As a result, observed individuals were simulated following $\lambda(s)b(s)$, the thinned intensity (Dorazio 2014; Koshkina et al. 2017). Figure 2 illustrates the simulated intensity $\lambda(s)$ and the thinned intensity $\lambda(s)b(s)$.

In addition to issues of detectability, PO data do not cover the full extent of species ranges. These PO data are said to be spatially truncated because they cover only a subset of the realized niche of species that follow geographic or political boundaries, such as country or continental borders (Thuiller et al. 2004; Hannemann et al. 2016; El-Gabbas and Dormann 2018; Mateo et al. 2019). In this study, we arbitrarily considered B', a subset of B, a rectangle area divided in 320×600 grid cells, to simulate the spatial niche truncation of the virtual species (see Fig. 2). To assess the effects of niche truncation on SDMs, two cases were considered for PO data in this study: the case where occurrence data cover the

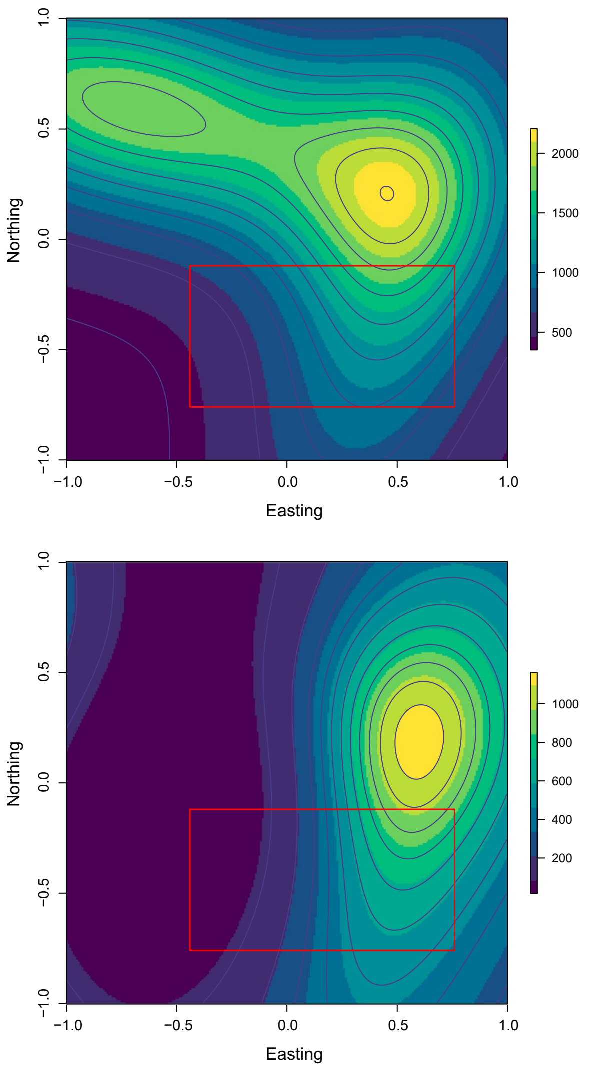

Fig. 2 Simulated intensity ($\lambda(s)$) and thinned intensity ($\lambda(s)b(s)$). The top panel corresponds to the high detectability scenario, and the bottom panel corresponds to the low detectability scenario. The red box represents subregion B' where truncated data are located

entire B (no truncation), and the case where occurrence data cover only B' (truncated niche).

Furthermore, among uncertainties associated with PO data is the uncertainty about where the occurrence was located. Positional uncertainty in PO data leads to a shift in point's position in the longitudinal and latitudinal directions (Heuvelink et al. 2007; Graham et al. 2008; Naimi et al. 2011; Osborne and Leitão 2009; Hefley et al. 2014). Let's denote E_i and N_i the coordinates (easting and northing) of where each individual species i was observed and recorded. In this study, we introduced positional error in PO data by introducing a positional error, ε in E_i and N_i using a probabilistic approach. The same approach was used in Hamm et al. (2004) and Naimi et al. (2011). This resulted in shifting the sampled species occurrences in random directions. We introduced the positional error, ε, in PO data as follows:

$$\varepsilon \sim \mathcal{N}(0, \vartheta)$$

$$E_i = E_i + k\varepsilon_{E_i} \tag{3}$$

$$N_i = N_i + k\varepsilon_{N_i} \tag{4}$$

With k, the resolution of $\lambda(s)$. Taking $\varepsilon \sim \mathcal{N}(0, \vartheta)$ gives a normally distributed unbiased error with the standard deviation ϑ that defines the positional uncertainty. The lower the ϑ, the higher the positional accuracy and vice versa. In this study, three levels of positional uncertainty were introduced by varying the values of ϑ. The values of ϑ were chosen so that the corresponding positional accuracy was high (uncertainty = no pixel shift), medium (uncertainty = shift of 4 pixels), and low (uncertainty = shift of 8 pixels).

Point-count data

To fit the integrated SDMs, additional point count (PC) are required. Collection of PC data i s expensive. Thus, these data are often collected in a small contiguous subregion of the study area (Koshkina et al. 2017). In this situation, PC data do not cover the full extent of the species' ranges. They are said to be spatially truncated. In this study, we examine how gathering PC data from B', a subset of B, could impact the performance of integrated SDM (PBPC Model). Two situations were considered in the process of simulating PC data: full range data and truncated data. Thus, we simulated the PC data by first dividing B into 50 × 50 square quadrats of equal size. Each quadrat corresponds to 20 × 20 grid cells of B. Usually, the PO observations far exceed the number of sites visited in the PC surveys (Dorazio 2014). In this study, the quadrats sampled in the planned surveys represent 4% of the total number of quadrats. Therefore, 100 quadrats were randomly selected from B for the full PC data, while 19 quadrats were randomly selected from B' for the truncated PC data. For each quadrat, the corresponding intensity values $\lambda(s)$ or the "true" number of individuals present in it was equal to the sum of the intensities corresponding to the grid cells that fall in it.

PC data are not prone to sampling bias because they are collected following structured sampling methods. However, imperfect detection can occur even during planned surveys. Hence, in contrast with PO data, detectability (detection probability) in PC data include only imperfection detection. In this study, we simulated each quadrat being visited by J repeated surveys, and as in Dorazio (2014) and (Koshkina et al. 2017), we assumed that the detectability, $p(s)$ and $b(s)$ at any site s are influenced by the same covariate $w(s)$. For repeated planned surveys, the detection probability was simulated depending only on a single covariate $w(s)$ as follows:

$$\text{logit}\,(p_j(s)) = \gamma_0 + \gamma_1 w(s) \tag{5}$$

For J repeated planned surveys, the detectability was assumed to be the same for all j surveys. In other words, for any site s in B, $p_1(s) = p_2(s) = \cdots = p_j(s)$. In this study, we arbitrarily considered $\gamma_0 = 2.5$ and $\gamma_1 = -1.0$.

Simulated PC data were obtained by conducting $J = 4$ independent binomial draws from individuals of each quadrat. In other words, each simulated set of PC was computed by aggregating the realized locations of individuals in the study area into quadrats, by selecting a random sample of these quadrats, and by taking $J = 4$ independent binomial draws from the individuals present in each sampled quadrat (see Dorazio 2014).

With all simulated data used to fit integrated SDMs, three cases of niche truncation were obtained: (i) no truncation: all data (PO and PC data) cover the whole B, (ii) partial truncation: PO data cover the whole B while PC data cover only B', (iii) full truncation: all data cover only B'.

Data analysis

Three SDMs were tested in this work:

1. PO Model: The spatial point process model that analyzes PO data ignoring the effect of $b(s)$. This model was fitted using simulated PO data sets solely (see Warton and Shepherd 2010);
2. THINPO Model: The spatial point process model that analyzes PO data as a thinned point process. This model account for $b(s)$ based on PO data solely (see Dorazio 2014);
3. PBPC Model: The integrated SDM that accounts for $b(s)$ by analyzing PO data in conjunction with PC data. This model was fitted using simulated PO data sets and PC data sets (see Dorazio 2014).

In the first stage, we tested the ability of these SDMs to estimate β_0 and β_1 parameters that determine the intensity $\lambda(\mathbf{s})$ in Eq. 1. The estimates of fitted models were compared to the "true" values used in the simulation process.

In all experiments, a total of 500 data sets containing PO observations and PC were simulated, with the SDMs then fitted to each realization of the data. β_0 and β_1 parameters were estimated using the BFGS (Broyden–Fletcher–Goldfarb–Shanno) optimization algorithm implemented in the *optim* function in R software (version 4.0.5) from the likelihood of the SDMs (Schank et al. 2019). Sometimes, the *optim* function failed to return an optimized set of parameters. If estimated parameters were returned from this function, we determined whether they were identifiable using the reciprocal of the condition number which is the ratio of the smallest to the largest eigenvalues of the Fisher information matrix (Dorazio 2014). The parameters of the species distribution models were considered identifiable if the reciprocal of the condition number had a value greater than 10^{-6}. Indeed, values of the reciprocal of the condition number close to 0 indicate poor conditioning (poor optimization) while values close to 1 indicate good conditioning (good optimization) (Golub and Loan 2013; Schank et al. 2019). Only estimates from models with identifiable parameters were considered for further analysis.

Performance assessment

Model evaluation is an essential step in model selection and determining the accuracy of the prediction. In general, model precision is measured primarily via evaluation and agreement metrics (Liu et al. 2011; Soultan and Safi 2017). In this study, the performance of each SDM was assessed at two levels: the ability of models to produce accurate operating characteristics of maximum likelihood estimates of β_k (namely $\hat{\beta}_k, k = 0, 1$), and their ability to predict accurately the species distribution $\lambda(s)$.

Measuring performance in estimating β_k

The utilized performance measures to assess the performance of SDMs in estimating β_k are presented in Table 1.

For $\hat{\beta}_k$, the relative bias (%Bias) was calculated for each replication while the standard deviation of $\hat{\beta}_k$ and the root mean squared error (RMSE) of $\hat{\beta}_k$ were calculated over N = 500 replications (runs) of the simulation process.

Measuring performance in predicting $\lambda(s)$

The Root mean squared error (RMSE) and two agreement metrics, namely the Schoener's D index and the overall concordance correlation coefficient (OCCC),

Table 1 Measures used to assess the performance of SDMs in estimating βk

Measure	Formula	Role
Relative bias in $\hat{\beta}_{ki}$	$\frac{\hat{\beta}_{ki}-\beta_{ki}}{\beta_{ki}} \times 100$	Unbiasedness of $\hat{\beta}_{ki}$
Standard deviation of $\hat{\beta}_k$	$\sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(\hat{\beta}_{ki} - \bar{\hat{\beta}}_k)^2}$	Precision of $\hat{\beta}_k$
RMSE of $\hat{\beta}_k$	$\sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{\beta}_{ki} - \beta_k)^2}$	Accuracy of $\hat{\beta}_k$

With $\hat{\beta}_k (k = 0, 1)$ the maximum likelihood estimates of the coefficients of $\log(\lambda(s))$ (see Eq. 1); N is the number of the replications (runs) of the simulation process. In this study, N = 500 were used to assess respectively the statistical performance (accuracy) of SDMs in predicting $\lambda(s)$ and the reliability of their predictions. The RMSE measures the unbiasedness (accuracy) of $\hat{\lambda}(s)$, whereas agreement metrics assess the spatial agreement between the "true" and the predicted ranges to determine prediction reliability. In other words, reliability can be used to determine the distance between predicted ranges and the "true" ranges (Soultan and Safi 2017). In this study, we determined the degree of agreement between "true" and predicted ranges by calculating the overlap of their geographical niches. We determined Schoener's D index using the "nicheOverlap" function from the "dismo" R package. The niche overlap value ranges from 0 to 1, where 0 denotes no overlap and 1 denotes complete overlap (Warren et al. 2008; Soultan and Safi 2017). In addition, we measured the absolute agreement between the "true" and modeled ranges via a pixel-by-pixel comparison using the OCCC, a measure of agreement between two continuous datasets generated using two distinct methodologies (Warren et al. 2008). The OCCC was calculated using the "epiR" R package. The OCCC value ranges from 0 to 1, with 0 indicating 100% disagreement and 1 indicating 100% agreement between the "true" and predicted ranges. These metrics were computed for each replication (run). M = 1000 random points were selected over B and used to extract $\lambda(s_1), \lambda(s_2), \lambda(s_3) \ldots \lambda(s_{1000})$, the "true" values of $\lambda(s)$ and $\hat{\lambda}(s_1), \hat{\lambda}(s_2), \hat{\lambda}(s_3) \ldots \hat{\lambda}(s_{1000})$. The extracted values of $\lambda(s)$ and $\hat{\lambda}(s)$ were then used to calculate the RMSE, the Schoener's D and the OCCC.

Results

Obtained $\hat{\beta}_0$ and $\hat{\beta}_1$ with different SDMs using PO data prone to different sources of uncertainty

Figure 3 shows the 95% confidence ellipses for the intensity coefficients (β_0 and β_1) obtained by fitting different SDMs under different types of uncertainty in data (PO and PC data). The plot illustrates the precision and accuracy with which the coefficients are estimated by each SDM. To highlight the marginal effects of each type of uncertainty, the confidence ellipses are determined using data that are not prone to the other two types of uncertainty. First, the results in Fig. 3A show that failure to account for imperfect detectability can lead to highly biased coefficients estimates ($\hat{\beta}_0$ and $\hat{\beta}_1$), altering the estimated geographic distribution of species ($\hat{\lambda}(s)$). The $\hat{\beta}_0$ coefficient is the most affected by the imperfect detectability. In that case, we are effectively estimating the presence-only intensity ($\lambda(s)b(s)$) instead of the species intensity ($\lambda(s)$) (see Fig. 3A). On the other hand, alternatives that attempt to account for imperfect detectability (THINPO and PBPC) give results that are more or less close to reality. In fact, the real values of the β_0 and β_1 coefficients fall within the confidence ellipses resulting from the THINPO and PBPC models, regardless of the detectability scenario. However, it is worth noting the widening of the confidence ellipses in low detectability situations. Second, Fig. 3B also shows a weak effect of positional uncertainty on the $\hat{\beta}_0$ and $\hat{\beta}_1$ coefficients for all three studied SDMs. We can see that the confidence ellipses contain the real values of the β_0 and β_1 coefficients despite the increase in positional uncertainty. With the variation of this factor, the accuracy and precision of $\hat{\beta}_0$ and $\hat{\beta}_1$ vary only slightly. However, with higher levels of imprecision than those considered in this study, it is not guaranteed that the effects will remain as small. Finally, regarding the spatial niche truncation, the same trend is observed for the effects of data truncation. For this source of uncertainty, the obtained confidence ellipses also contain points that represent the real values of the β_0 and β_1 coefficients, regardless of the SDM (see Fig. 3C). However, it is worth noting the widening of the confidence ellipses in the situation of full truncation (all data do not cover the full range of the species). It is mainly the loss of precision of $\hat{\beta}_1$. The results presented in the rest of this section illustrate the performance of the SDMs under different combinations of these three factors.

Effects of uncertainties on maximum likelihood of β_0 and β_1

In situation of low detectability, $\hat{\beta}_0$ obtained with PO Model are strongly biased. In doing so based on PO data solely, the THINPO Model has shown to improve $\hat{\beta}_0$. This approach alleviate bias in $\hat{\beta}_0$ but with high variance, which reflects a low precision. To obtain much better estimates (with low bias and high precision), the integrated SDMs are the best alternatives. The use PC data in conjunction with PO data through the PBPC Model did improve $\hat{\beta}_0$ over the THINPO Model by increasing their precision. Regarding niche truncation, the precision of $\hat{\beta}_0$ decreases slightly in the situation of full spatial niche truncation. Partial truncation becomes challenging for $\hat{\beta}_0$ when in addition the PO data are subject to position imprecision (low and medium precision). This behavior is observed for all detectability scenarios except that the higher the detectability, the higher the precision of $\hat{\beta}_0$. In the situation of high detectability, the PO model outperformed the others by giving unbiased and the most precise $\hat{\beta}_0$, whatever the positional uncertainty or the spatial niche truncation (see Fig. 4 and Table 2).

We notice that globally for all SDMs $\hat{\beta}_1$ are relatively unbiased whatever the detectability scenario and niche truncation. The only exception is for the PO Model under low detectability when PO data are truncated. However, the precision of $\hat{\beta}_1$ is impaired by the decrease in detectability and the spatial niche truncation. For this parameter, the low the detectability, the

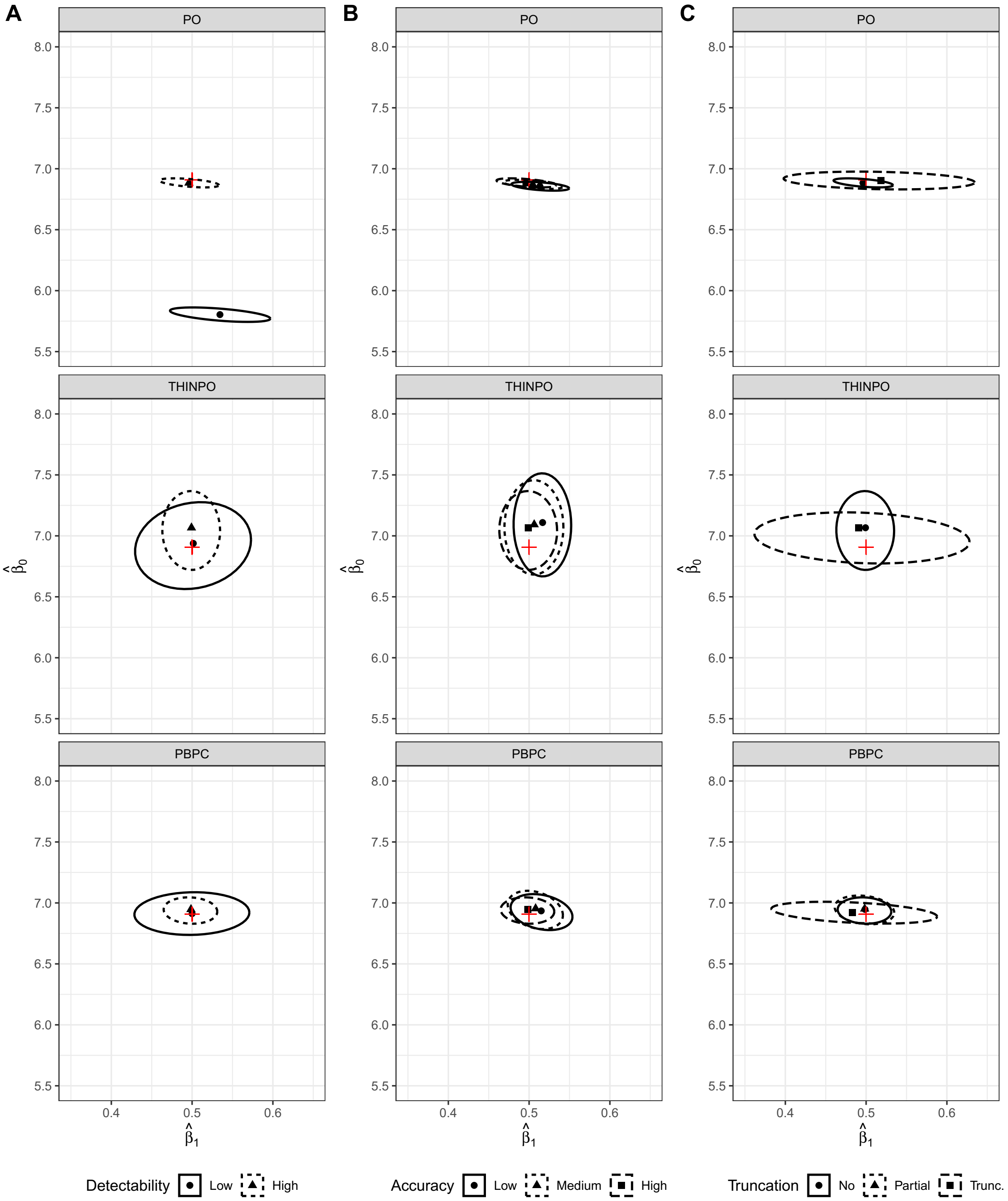

Fig. 3 95% confidence ellipses for $\hat{\beta}_0$ and $\hat{\beta}_1$ obtained by fitting SDMs under varying detectability (**A**), positional accuracy (**B**), and niche truncation (**C**). The red plus denotes the "true" values of the parameters of interest ($\beta_0 \approx 6.91$and $\beta_1 = 0.5$).

lower the precision. With partial or full niche truncation, the precision $\hat{\beta}_1$ is lower. As for $\hat{\beta}_1$, the partial niche truncation becomes challenging for $\hat{\beta}_1$ when PO data are prone to positional imprecision (see Fig. 5 and Table 2).

Effects of uncertainties on the accuracy and reliability of SDMs' predictions ($\hat{\lambda}(s)$)

All the effects of the uncertainties under study on $\hat{\beta}_0$ and $\hat{\beta}_1$ are expected to affect the estimates of the species distribution (intensity) and thus $\hat{\lambda}(s)$ obtained of the used SDMs. In this study we have assessed the

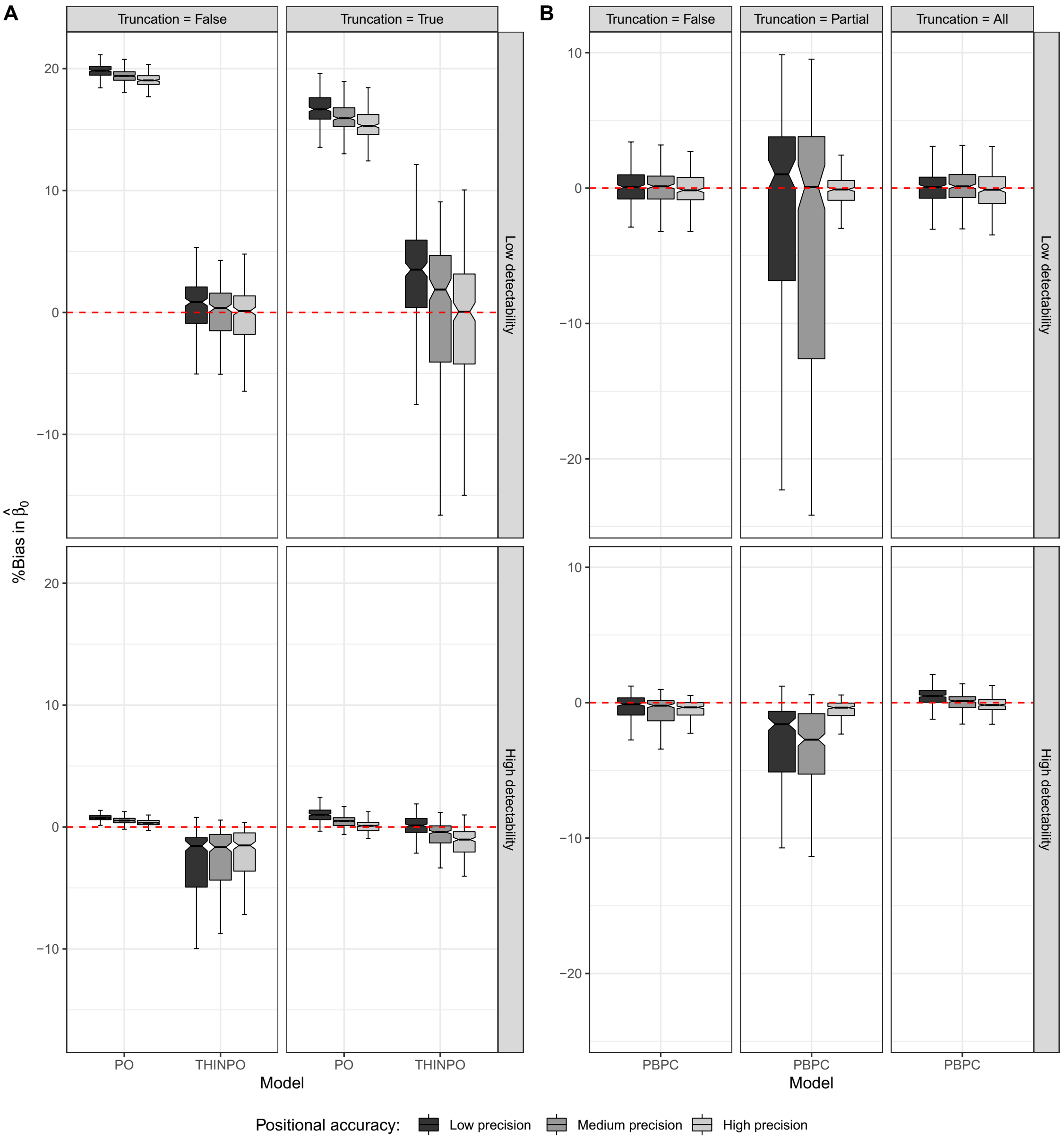

Fig. 4 Relative bias in maximum likelihood estimates of β_0 obtained by fitting PO SDMs (**A**) and Integrated SDM (**B**). The dashed red line indicates the ideal situation where the bias in β_0 is equal to 0

statistical performance of SDMs through RMSE which measures the accuracy of $\hat{\lambda}(s)$. In addition, we measured the reliability of $\hat{\lambda}(s)$ through the Schoener's D index and the OCCC which measure the spatial agreement between "true" and predicted species distribution. The Schoener's D measures the relative agreement while the OCCC measures the absolute agreement. By doing so, we consider that a model is well performing when the its corresponding RMSE is low (near zero) and high Schoener's D and OCCC (near 1). The results obtained are summarized in Figs. 6, 7 and 8.

Imperfect detection leads to a significant loss of accuracy (significant increase of RMSE) in species distribution estimates ($\hat{\lambda}(s)$) when not accounted for (see Fig. 6A). The THINPO Model and PBPC Model were able to estimate the distribution of species ($\hat{\lambda}(s)$) with high accuracy (lower RMSE) compared to the PO Model. For integrated SDM (PBPC Model), in addition to being less sensitive to imperfect detectability, the positional uncertainty does not induce considerable effects on the accuracy of this model (it induces less variation in RMSE). Furthermore, the spatial niche truncation does not induce significant

Table 2 Standard deviation (and RMSE) of maximum likelihood estimates of β_0 and β_1 under varying positional uncertainty in the situation of low and high detectability

SDM	Truncation	Detectability	$\hat{\beta}_0$			$\hat{\beta}_1$		
			Low	Medium	High	Low	Medium	High
PO	False	Low	0.06 (1.14)	0.06 (1.12)	0.05 (1.10)	0.06 (0.07)	0.06 (0.06)	0.06 (0.04)
PO	False	High	0.03 (0.05)	0.04 (0.04)	0.04 (0.03)	0.03 (0.02)	0.03 (0.02)	0.03 (0.02)
PO	True	Low	0.13 (0.99)	0.13 (0.96)	0.13 (0.93)	0.18 (0.55)	0.18 (0.55)	0.18 (0.55)
PO	True	High	0.08 (0.08)	0.07 (0.05)	0.07 (0.04)	0.11 (0.06)	0.11 (0.06)	0.11 (0.06)
THINPO	False	Low	0.36 (0.18)	0.36 (0.18)	0.36 (0.18)	0.07 (0.05)	0.07 (0.04)	0.07 (0.03)
THINPO	False	High	0.38 (0.28)	0.36 (0.26)	0.30 (0.22)	0.03 (0.02)	0.03 (0.02)	0.03 (0.02)
THINPO	True	Low	1.46 (0.73)	1.87 (0.95)	2.09 (1.09)	0.24 (0.12)	0.24 (0.12)	0.26 (0.13)
THINPO	True	High	0.23 (0.11)	0.52 (0.28)	0.68 (0.37)	0.17 (0.08)	0.16 (0.08)	0.14 (0.07)
PBPC	False	Low	0.18 (0.09)	0.18 (0.09)	0.16 (0.08)	0.06 (0.04)	0.06 (0.03)	0.07 (0.03)
PBPC	False	High	0.14 (0.08)	0.16 (0.09)	0.12 (0.07)	0.04 (0.02)	0.03 (0.02)	0.03 (0.02)
PBPC	Partial	Low	1.43 (0.75)	1.54 (0.84)	0.15 (0.07)	0.25 (0.13)	0.26 (0.13)	0.07 (0.03)
PBPC	Partial	High	0.80 (0.48)	0.60 (0.41)	0.12 (0.07)	0.18 (0.10)	0.18 (0.10)	0.03 (0.02)
PBPC	All	Low	0.17 (0.09)	0.20 (0.10)	0.19 (0.09)	0.18 (0.09)	0.17 (0.08)	0.18 (0.09)
PBPC	All	High	0.12 (0.06)	0.11 (0.05)	0.09 (0.05)	0.12 (0.06)	0.11 (0.06)	0.09 (0.05)

loss of accuracy in $\hat{\lambda}(s)$, except when it occurs together with the issues of positional uncertainty in PO data. In this situation, the PBPC model shows a non-negligible loss of accuracy, as expressed by RMSE values (see Fig. 6B). For the THINPO model, the effects of spatial truncation are not as negligible as for the PBPC model. And the behavior of this model becomes erratic in situations of near perfect detectability (see Fig. 6A). It should be noted that when there is no problem of imperfect detectability in the PO data, the PO model outperforms the other SDMs. In fact, under these circumstances, this model gives the best accuracy (low RMSE) (see Fig. 6).

The results of Schoener's D index show that none of the factors studied (imperfect detectability, positional uncertainty and spatial niche truncation) induce significant effects on the spatial niche overlap of the predictions ($\hat{\lambda}(s)$) with the real distribution of the species of interest ($\lambda(s)$), except in the case of low detectability and spatial truncation for the PO model. It should be noted that obtained Schoener's D index values seem to minimize the effects of the studied uncertainties in data on the reliability of SDMs. Surprisingly, the results of Schoener's D index show that, despite the effects observed in the previous results, the predictions obtained remain reliable.

On the other hand, the OCCC results are more or less in line with those of the RMSE. Imperfect detection leads to a significant loss of spatial agreement between $\hat{\lambda}(s)$ and $\lambda(s)$ when not accounted for (see Fig. 8). The integrated SDM (PBPC Model), in addition to being less sensitive to imperfect detectability, the positional uncertainty does not induce considerable effects on the spatial agreement between $\lambda(s)$ and $\hat{\lambda}(s)$ it gives. Furthermore, the spatial niche truncation does not induce significant loss of spatial agreement, except when it occurs together with the issues of positional uncertainty in PO data as shown in Fig. 6B. As for results of RMSE, it should be noted that when there is no problem of imperfect detectability in the PO data, the PO model outperforms the other SDMs, even in situation spatial niche truncation. In fact, under these circumstances, this model gives OCCC almost equal to 1.

Discussion

SDMs are based on a number of assumptions to guarantee their performance and the reliability of their predictions. However, it is not readily available to find PO data that meet these assumptions, which raises doubts about the reliability of the conclusions drawn. This study illustrates the effect of multi-source uncertainties in species PO data on SDMs performance. The aim is to assess the (marginal and combined) effects of these uncertainties on the ability of the SDMs to estimate the parameters of the species distribution model (intensity, $\lambda(s)$) and on the predictive performance of these models.

In this study, the poor performance of the PO model is evidence that imperfect detectability leads to a serious loss of predictive performance in SDMs, leading to erroneous conclusions about species ranges. SDMs predictions obtained without accounting for imperfect detectability, instead of estimating $\lambda(s)$, the "true" species distribution, they reflect $\lambda(s)b(s)$, the sampling efforts (Phillips et al. 2009; Fithian et al. 2015). In this context,

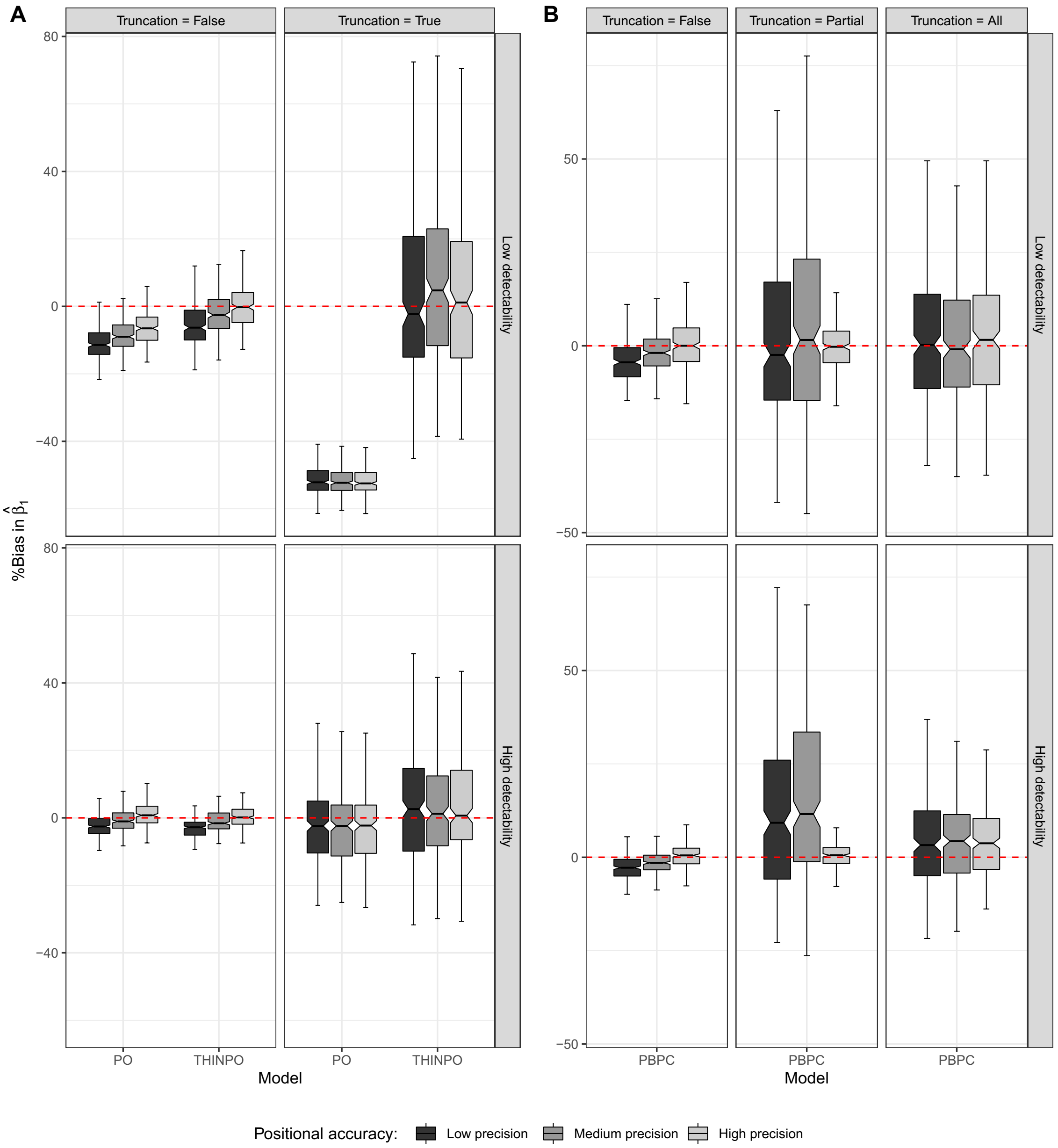

Fig. 5 Relative bias in maximum likelihood estimates of β_1 obtained by fitting PO SDMs (**A**) and Integrated SDM (**B**). The dashed red line indicates the ideal situation where the bias in β_1 is equal to 0

it is challenging to distinguish between predictions that accurately reflect ecological processes that influence the spatial distribution of a species and those that are linked to detectability effects or sampling effect (Dorazio 2012; Fithian et al. 2015; Guillera-Arroita 2017). Therefore, our findings emphasize on the importance of accounting for imperfect detectability in SDMs. They corroborate findings of other works that insist on the risk of ignoring this type of bias in species occurrence data. As for our study, Phillips et al. (2009), Yackulic et al. (2013) and Guillera-Arroita (2017) showed that ignoring imperfect detectability is not inconsequential to the reliability of SDMs predictions. It leads to erroneous conclusions regarding the distribution of species, erroneous inferences regarding the determinants of species distribution, incorrect quantifications of biodiversity, and incorrect conclusions regarding environmental change (Guillera-Arroita 2017). However, our findings are not in line with some studies that recommend ignoring the effects of imperfect detectability. Indeed, there are differing

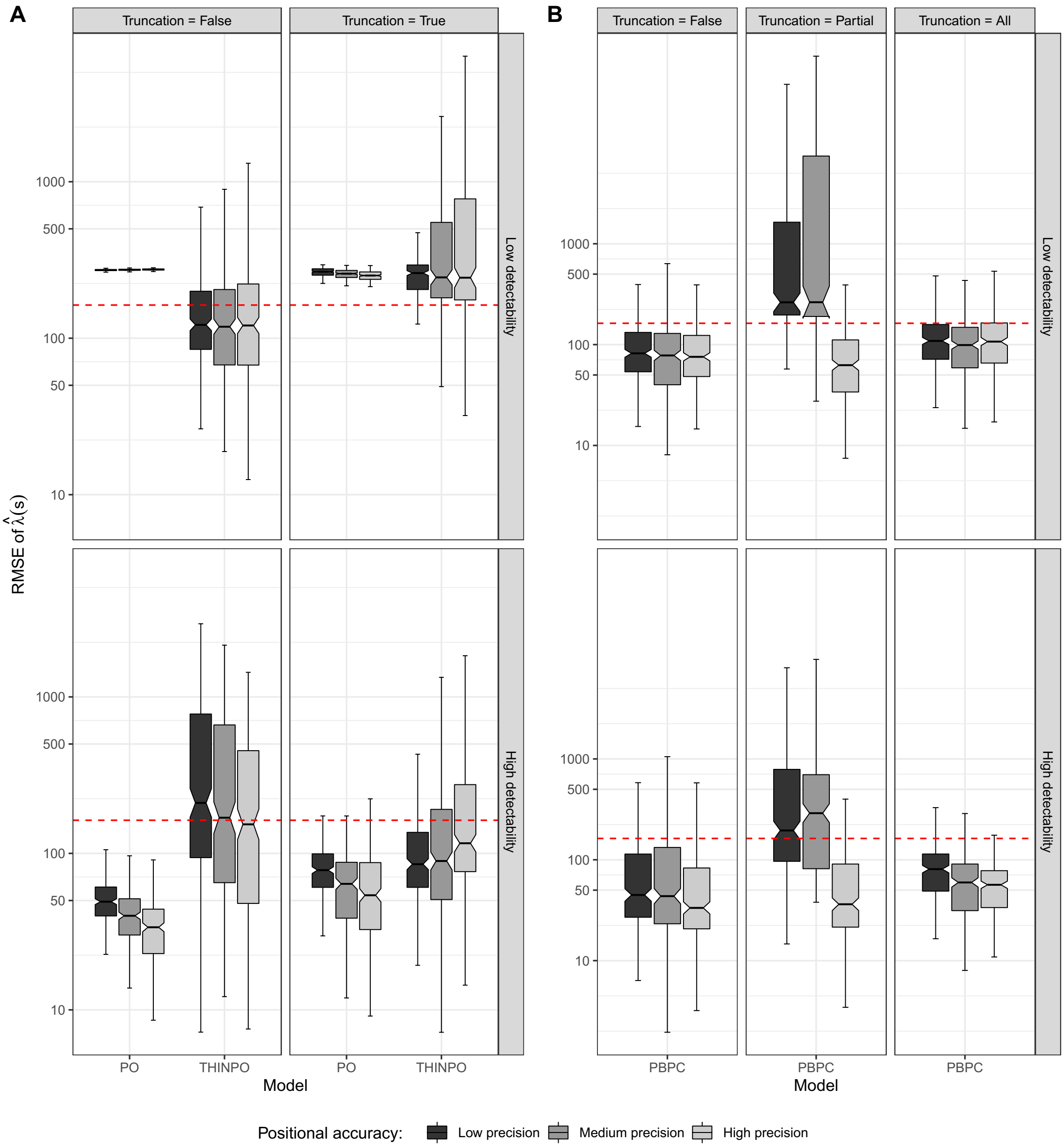

Fig. 6 Accuracy of $\hat{\lambda}(s)$ as shown by the RMSE obtained with different SDMs fitted using different scenario of uncertainty in data: PO SDMs (**A**) and Integrated SDM (**B**). The dashed red line indicates the median of RMSE. In this figure, the higher the RMSE, the lower the accuracy and vice versa

opinions regarding the effects of imperfect detectability on SDMs performance (Guélat and Kéry 2018). Some studies concluded that the effects of imperfect detectability are negligible and recommended ignoring them (e.g., Banks-Leite et al. 2014; Johnson and Gillingham 2008; Stephens et al. 2015). We believe that this difference of opinion may be explained by the fact that the effects of imperfect detectability may vary according to the species eco-geographic characteristics. For example, for generalist species, although the data are prone to geographic sampling bias, they may be sufficiently representative of the environmental conditions across the full species range and then allow the SDMs to capture the favorable environmental conditions for the species. In contrast, this is not necessarily true for specialized species. The effects of imperfect detectability would also vary with the variance and importance of the underlying factors. For example, Fithian et al. (2015); Thibaud et al. (2014), and Fletcher et al. (2016) and other more recent studies such as Chevalier et al. (2021) suggested using covariates such

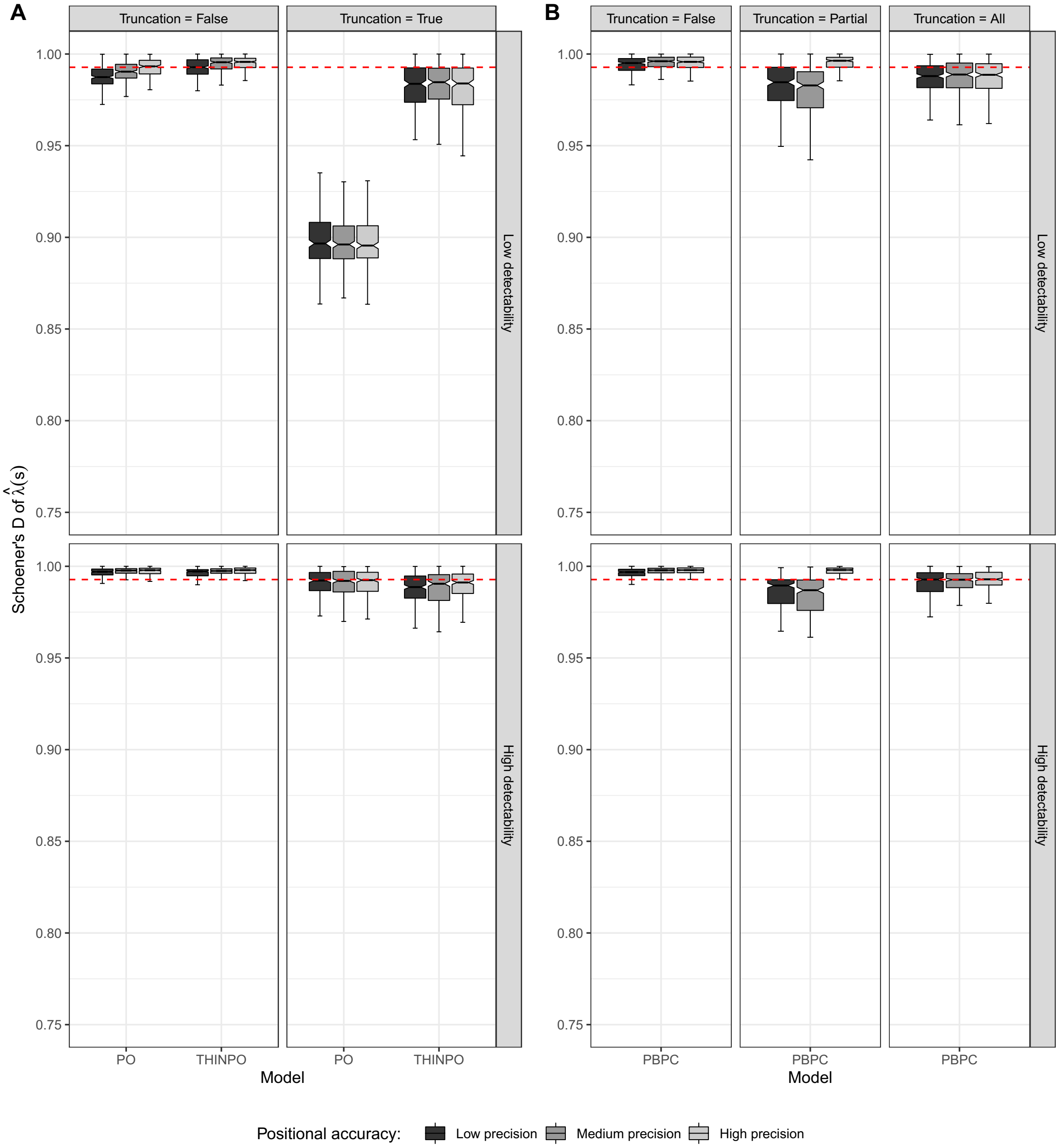

Fig. 7 The spatial niche overlap between the predicted species distribution ($\hat{\lambda}(s)$) with the "true" species distribution ($\lambda(s)$) according to the Schoener's D index obtained with different SDMs fitted using different scenario of uncertainty in data: PO SDMs (**A**) and Integrated SDM (**B**). The dashed red line indicates the median of Schoener's D index

as the distance to a roads or distance to cities as predictors of imperfect detectability. If distance to roads and/or distance to large cities are the main factors underlying imperfect detectability, then if the road network is sufficiently developed and the cities sufficiently numerous and scattered, thus covering a large part of the environmental conditions of the species, the effects of this factor may be sufficiently reduced and thus induce minor losses in model performance.

Findings of this study show that in the context of low detectability all studied SDMs produce unbiased estimates of β_0 and β_1, but they differ mainly in the precision of these estimates. The best accuracy is obtained with the PBPC Model. However, in the context of high detectability, PO Model outperformed the SDMs that account for imperfect detectability (THINPO and PBPC Models). It is therefore important to make a careful selection of the model to be used, taking into account the characteristics

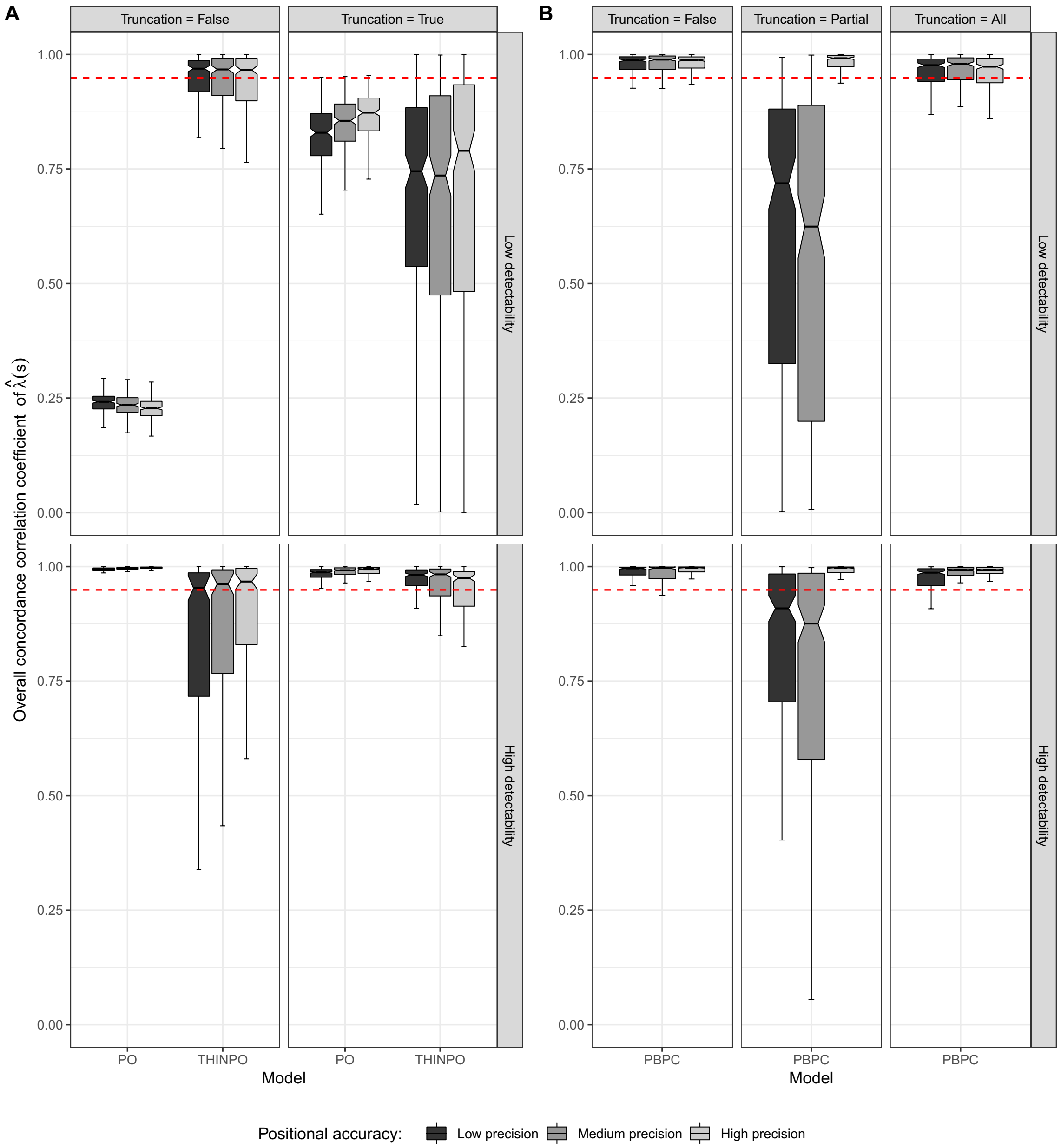

Fig. 8 The spatial agreement between the predicted species distribution ($\hat{\lambda}(s)$) with the "true" species distribution ($\lambda(s)$) according to the OCCC obtained with different SDMs fitted using different scenario of uncertainty in data: PO SDMs (**A**) and Integrated SDM (**B**). The dashed red line indicates the median of OCCC

of the data. To account for imperfect detectability in PO data, several authors have proposed modeling species distribution as a thinned Poisson point process (THINPO Model) (Chakraborty et al. 2011; Fithian and Hastie 2013; Hefley et al. 2013; Warton et al. 2013; Dorazio 2014; Fithian et al. 2015). The results of this study show that this may not be enough depending on data characteristics. It may be necessary to use additional data in some context as demonstrated in this study. The results of this study are in accordance with those of some authors that argue that it is impossible to accurately estimate species distribution using PO data alone because PO data are not informative about species detectability. Additional data that are informative about species detectability are required to improve the estimation of the parameters of species distribution models (Fithian et al. 2015; Dorazio 2014; Koshkina et al. 2017). Although the use of these additional data is crucial, it is also necessary to make a reasoned choice of the covariates to be included in the model, especially the detectability variables. This choice is not trivial. It must be done as thoroughly as possible (Koshkina et al. 2017). A poor choice of these covariates could also have negative effects on the performance of the SDMs. However, our results are silent on this issue. What would be the bias in the parameter estimates if the key determinants of detectability are omitted? Nolan et al. (2022) proposed an approach to identify sources of sampling bias in PO data through the use of zero-inflated models. We must therefore try to list all the potential sampling bias drivers and then use the approach proposed by Nolan et al. (2022) to identify those that should be retained for use in the THINPO Model or integrated SDM.

In addition, the assumption regarding the data representativeness of the environmental conditions across the full species range is often violated because in most of the SDMs' studies, PO data are typically collected in areas defined by geographical or political borders (e.g., national monitoring programs) that only encompass a subset of a species' realized niche (Hannemann et al. 2016; El-Gabbas and Dormann 2018; Chevalier et al. 2021). These data are therefore not necessarily representative of the species range (Chevalier et al. 2021). Surprisingly, the results obtained in this study show little effect of spatial niche truncation on the maximum likelihood estimates of β_0 and β_1, and thus fail to induce strong effects on the predicted species distribution ($\hat{\lambda}(s)$), especially for the integrated SDM. Indeed, in this study, $\hat{\beta}_0$ and $\hat{\beta}_1$ of the integrated SDM (PBPC model) are unbiased regardless of the spatial niche truncation scenario. Furthermore, small effects of spatial niche truncation on the precision of $\hat{\beta}_0$ and $\hat{\beta}_1$ are observed, which do not significantly affect the predictions ($\hat{\lambda}(s)$) of this model, except when PO data are positionally imprecise in addition to being truncated. Consequently, this study suggests that the use of spatially constrained data, has little effect on the results of SDMs. And since high quality data are often very spatially constrained, there is no reason to fear that this will affect the results of integrated models. In terms of transferability, if the data used to calibrate the models are sufficiently representative of the full range of the species, the results of SDMs can be generalized to other geographic areas and time periods without concern. This is not consistent with a number of previous studies that have found evidence of severe effects of spatial niche truncation on the quality of SDM outputs. Indeed, these research stated that if species occurrence data fail to capture the full species realized niche, they can not adequately characterize the environmental conditions tolerated by species. Thus, it may not be possible to obtain reliable outputs from models built using such data (Pearson and Dawson 2003; Thuiller et al. 2004; Titeux et al. 2017). Consequently, it is difficult to estimate the function linking species distributions and environmental variables, leading to inaccurate predictions of species distributions (Chevalier et al. 2021; Thuiller et al. 2004), which can result in wasting resources on ineffective and expensive restoration plans or losing populations of conservation concern (Guisan et al. 2013; Araújo et al. 2019; Chevalier et al. 2021). We suspect that the discrepancy between our results and those of previous research may be explained by the fact that the ecological conditions of the area chosen for our spatial niche truncation simulations may have been sufficiently representative of the environmental tolerance of the virtual species. Elith et al. (2010) recommend that the similarity between the restricted data used for model calibration and the full range data (extrapolation data or projection data) be tested by multivariate environmental similarity surface analysis. If this analysis shows dissimilarities, one should pay attention to spatial niche truncation effects (Barbet-Massin et al. 2010; Bálint et al. 2011; Edman et al. 2011; Keenan et al. 2011; Bertrand et al. 2012; Raes 2012; Chevalier et al. 2021). An alternative to address this issue is to consider data from larger spatial and (especially) ecological scales (Hannemann et al. 2016; Chevalier et al. 2021). It is then preferable to use the full range of species and environmental data to calibrate SDMs rather than considering only a subset of them (Araújo and Guisan 2006). However, this alternative may not be satisfactory if SDMs are to be fitted at fine spatial scales using local predictors such as land cover and fine environmental details such as local microclimate (Pearson et al. 2004; Zellweger et al. 2019; Chevalier et al. 2021). Furthermore, we suggest that the effects of spatial niche truncation are likely to vary with respect to the eco-geographic characteristics of the species. Spatial niche truncation may have stronger effects on specialized (narrowly distributed) species than on generalist species. Since our virtual species is not highly specialized, we expect it to be less sensitive to spatial niche truncation.

In addition to the aforementioned uncertainties, the positional uncertainty of PO data is an additional concern (Naimi et al. 2011; Rocchini et al. 2011; Soultan and Safi 2017). The use of such prone-to-uncertainty PO data is hypothesized to result in inaccurate predictions of species distributions, and then misguide biodiversity management and conservation efforts. The results of this study indicate that the effects of positional uncertainty on maximum likelihood estimates of the β_0 and β_1 coefficients are not as severe as one might expect. These results are consistent with previous research indicating that the effect of positional uncertainty of species occurrences on the performance of SDM is relatively small (Graham et al. 2008; Osborne and Leitão 2009; Soultan and Safi

2017; Hayes et al. 2015; Fernandez et al. 2009; Mitchell et al. 2017). However, there are other studies that contradict our findings. They found that positional precision can lead to inaccurate predictions of species distributions (Visscher 2006; Johnson and Gillingham 2008; Naimi et al. 2011, 2014). We suspect that this disagreement may be related to the fact that positional uncertainty effects would vary with species characteristics, which may vary with responses to environmental covariates (Soultan and Safi 2017). Indeed, Soultan and Safi (2017) found that species specialization affects the sensitivity of SDMs to the positional uncertainty in species occurrence data. For generalist species, positional precision has a relatively small effect, whereas specialist species are more sensitive to positional uncertainty. The sensitivity of specialist species may be due to an increased probability of assigning imprecise species occurrences to inappropriate areas, whereas this probability is inherently reduced for generalist species. Species characteristics are also likely to influence the effectiveness of SDMs. The results of this study may not be applicable to all species. Therefore, future research should assess the effects of species characteristics on the performance of SDMs and the uncertainty in PO data as a function of species characteristics to determine if these SDMs continue to exhibit the same performance.

Conclusion

In this study, we used simulated data to investigate the effects of positional uncertainty and spatial niche truncation on the performance of species distribution models (SDMs) under low and high species detectability. We show that SDMs that account for imperfect detectability (THINPO or PBPC models) are not applicable in high detectability situations. In this situation, PO model produces the most accurate maximum likelihood estimates of β_0 and β_1, and consequently the most accurate predictions of species distributions ($\hat{\lambda}(s)$). The effects of positional uncertainty and spatial niche truncation on this SDM output are minimal. However, in situations of low detectability, it is preferable to analyze PO data alongside PC data. It has been demonstrated that positional uncertainty and spatial niche truncation have negligible effects on the output of this SDM, except when positionally uncertain PO data are analyzed alongside truncated PC data. However, depending on species characteristics, the effects of positional uncertainty and spatial niche truncation may vary. They can have a significant impact on the outputs of SDMs for specialized species. Multivariate environmental similarity surface analysis is proposed to test the similarity between data from the restricted region to be used for model calibration and data from the entire range. If this analysis reveals dissimilarities, spatial niche truncation effects should be considered. Data from larger spatial and ecological scales should be considered as an alternative to address this issue. It is therefore preferable to use the full range of species and environmental data to calibrate SDMs, rather than just a subset of them. Assessing the effects of species characteristics on model performance and the effects of uncertainties on model performance as a function of species characteristics could be the subject of future research.

Acknowledgements
Support for this research was made possible through a capacity building competitive grant Training Next Generation of Scientists (Grant #RU/2020/GTA/DRG/036) provided by Carnegie Cooperation of New-York through the Regional Universities Forum for Capacity Building in Agriculture (RUFORUM). The first author gratefully acknowledges both the Université Evangélique en Afrique (UEA), through the university's project to improve the quality of research and teaching funded by Pain pour le Monde (Project A-COD-2023-0035), and the Centre d'Excellence d'Afrique en Sciences Mathématiques, Informatique et Applications (CEA-SMIA), by providing travel and living expenses grants during his Ph.D. training.

Author contributions
Conceptualization: YM and ABF; methodology: YM, ABF, and RGK; writing-original draft preparation: YM, ABF, and RGK; writing-review and editing: YM, ABF, and RGK; supervision: ABF and RGK. All authors read and approved the final manuscript.

Funding
Not applicable.

Availability of data and materials
The data supporting the findings of this study are available from the corresponding author upon reasonable request.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1] Laboratoire de Biomathématiques et d'Estimations Forestières (LABEF), Faculty of Agronomic Sciences, University of Abomey-Calavi, 04 BP 1525, Cotonou, Benin. [2] Unit of Applied Biostatistics, Faculty of Agriculture and Environmental Sciences, Université Evangélique en Afrique, PO. Box: 3323, Bukavu, Democratic Republic of Congo. [3] Unité de Recherche en Foresterie et Conservation des Bioressources, Ecole de Foresterie Tropicale, Université Nationale d'Agriculture, PO. Box: 45, Kétou, Benin.

Received: 9 May 2023 Accepted: 10 July 2023
Published online: 25 July 2023

References

Araújo MB, Guisan A (2006) Five (or so) challenges for species distribution modelling. J Biogeogr 33(10):1677–1688. https://doi.org/10.1111/j.1365-2699.2006.01584.x

Araújo MB, Anderson RP, Barbosa AM et al (2019) Standards for distribution models in biodiversity assessments. Sci Adv 5(1):eaat4858. https://doi.org/10.1126/sciadv.aat4858

Banks-Leite C, Pardini R, Boscolo D et al (2014) Assessing the utility of statistical adjustments for imperfect detection in tropical conservation science. J Appl Ecol 51(4):849–859. https://doi.org/10.1111/1365-2664.12272

Barbet-Massin M, Thuiller W, Jiguet F (2010) How much do we overestimate future local extinction rates when restricting the range of occurrence data in climate suitability models? Ecography 33:878–886. https://doi.org/10.1111/j.1600-0587.2010.06181.x

Bertrand R, Perez V, Gégout JC (2012) Disregarding the edaphic dimension in species distribution models leads to the omission of crucial spatial information under climate change: the case of *Quercus pubescens* in France. Glob Change Biol 18:2648–2660. https://doi.org/10.1111/j.1365-2486.2012.02679.x

Bálint M, Domisch S, Engelhardt C et al (2011) Cryptic biodiversity loss linked to global climate change. Nat Clim Change 1:313–318. https://doi.org/10.1038/nclimate1191

Chakraborty A, Gelfand AE, Wilson AM et al (2011) Point pattern modelling for degraded presence-only data over large regions. J R Stat Soc Ser C (Appl Stat) 60(5):757–776. https://doi.org/10.1111/j.1467-9876.2011.00769.x

Chen G, Kéry M, Plattner M et al (2013) Imperfect detection is the rule rather than the exception in plant distribution studies. J Ecol 101(1):183–191. https://doi.org/10.1111/1365-2745.12021

Chevalier M, Broennimann O, Cornuault J et al (2021) Data integration methods to account for spatial niche truncation effects in regional projections of species distribution. Ecol Appl 31(7):e02427. https://doi.org/10.1002/eap.2427

Dawson TP, Jackson ST, House JI et al (2011) Beyond predictions: biodiversity conservation in a changing climate. Science 332(6025):53–58. https://doi.org/10.1126/SCIENCE.1200303

Dorazio RM (2012) Predicting the geographic distribution of a species from presence-only data subject to detection errors. Biometrics 68(4):1303–1312. https://doi.org/10.1111/j.1541-0420.2012.01779.x

Dorazio RM (2014) Accounting for imperfect detection and survey bias in statistical analysis of presence-only data. Glob Ecol Biogeogr 23(12):1472–1484. https://doi.org/10.1111/geb.12216

Duputié A, Zimmermann NE, Chuine I (2014) Where are the wild things? Why we need better data on species distribution. Glob Ecol Biogeogr 23(4):457–467. https://doi.org/10.1111/GEB.12118

Edman T, Angelstam P, Mikusinski G et al (2011) Spatial planning for biodiversity conservation: assessment of forest landscapes' conservation value using umbrella species requirements in Poland. Landsc Urban Plan 102:16–23. https://doi.org/10.1016/j.landurbplan.2011.03.004

El-Gabbas A, Dormann C (2018) Wrong, but useful: regional species distribution models may not be improved by range-wide data under biased sampling. Ecol Evol 8(2196–2206):1. https://doi.org/10.1002/ece3.3834

Elith J, Leathwick J (2009) Species distribution models: ecological explanation and prediction across space and time. Annu Rev Ecol Evol Syst 40:677–697. https://doi.org/10.1146/annurev.ecolsys.110308.120159

Elith J, Kearney M, Phillips S (2010) The art of modelling range-shifting species. Methods Ecol Evol 1(4):330–342. https://doi.org/10.1111/j.2041-210X.2010.00036.x

Fernandez M, Blum S, Reichle S et al (2009) Locality uncertainty and the differential performance of four common niche-based modeling techniques. Biodivers Inform 6:36–52. https://doi.org/10.17161/bi.v6i1.3314

Fithian W, Hastie T (2013) Finite-sample equivalence in statistical models for presence-only data. Ann Appl Stat 7(4):1917–1939. https://doi.org/10.1214/13-AOAS667

Fithian W, Elith J, Hastie T et al (2015) Bias correction in species distribution models: pooling survey and collection data for multiple species. Methods Ecol Evol 6(4):424–438. https://doi.org/10.1111/2041-210X.12242

Fletcher RJ, McCleery RA, Greene DU et al (2016) Integrated models that unite local and regional data reveal larger-scale environmental relationships and improve predictions of species distributions. Landsc Ecol 31(6):1369–1382. https://doi.org/10.1007/s10980-015-0327-9

Franklin J (2010) Mapping species distributions: spatial inference and prediction. Ecology, biodiversity and conservation. Cambridge University Press, Cambridge. https://doi.org/10.1017/CBO9780511810602

Franklin J (2013) Species distribution models in conservation biogeography: developments and challenges. Divers Distrib. https://doi.org/10.1111/ddi.12125

Golub G, Loan C (2013) Matrix computations, 4th edn. Johns Hopkins University Press, Baltimore

Graham CH, Ferrier S, Huettman F et al (2004) New developments in museum-based informatics and applications in biodiversity analysis. Trends Ecol Evol 19(9):497–503. https://doi.org/10.1016/j.tree.2004.07.006

Graham C, Elith J, Hijmans R et al (2008) The influence of spatial errors in species occurrence data used in distribution models. J Appl Ecol 45:239–247. https://doi.org/10.1111/j.1365-2664.2007.01408.x

Guillera-Arroita G (2017) Modelling of species distributions, range dynamics and communities under imperfect detection: advances, challenges and opportunities. Ecography 40(2):281–295. https://doi.org/10.1111/ecog.02445

Guisan A, Tingley R, Baumgartner JB et al (2013) Predicting species distributions for conservation decisions. Ecol Lett 16(12):1424–1435. https://doi.org/10.1111/ele.12189

Guélat J, Kéry M (2018) Effects of spatial autocorrelation and imperfect detection on species distribution models. Methods Ecol Evol 9(6):1614–1625. https://doi.org/10.1111/2041-210X.12983

Hamm N, Atkinson PM, Milton EJ (2004) On the effect of positional uncertainty in field measurements on the atmospheric correction of remotely sensed imagery. In: Sanchez-Vila X, Carrera J, Gómez-Hernández JJ (eds) geoENV IV—geostatistics for environmental applications. Springer, Netherlands, pp 91–102

Hannemann H, Willis KJ, Marc MF et al (2016) Wrong, but useful: regional species distribution models may not be improved by range-wide data under biased sampling. Glob Ecol Biogeogr 25(1):26–35. https://doi.org/10.1111/geb.12381

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

Hayes M, Ozenberger K, Cryan P et al (2015) Not to put too fine a point on it—does increasing precision of geographic referencing improve species distribution models for a wide-ranging migratory bat? Acta Chiropterol 17:159–169. https://doi.org/10.3161/15081109ACC2015.17.1.013

Hefley TJ, Tyre AJ, Baasch DM et al (2013) Nondetection sampling bias in marked presence-only data. Ecol Evol 3(16):5225–5236. https://doi.org/10.1002/ece3.887

Hefley T, Baasch D, Tyre AJ et al (2014) Correction of location errors for presence-only species distribution models. Methods Ecol Evol 5:207–214. https://doi.org/10.1111/2041-210X.12144

Heuvelink GBM, Brown JD, van Loon EE (2007) A probabilistic framework for representing and simulating uncertain environmental variables. Int J Geogr Inf Sci 21(5):497–513. https://doi.org/10.1080/13658810601063951

Hirzel A, Helfer V, Metral F (2001) Assessing habitat-suitability models with a virtual species. Ecol Model 145(2–3):111–1121. https://doi.org/10.1016/S0304-3800(01)00396-9

Hortal J, Lobo JM, Jiménez-Valverde A (2007) Limitations of biodiversity databases: case study on seed-plant diversity in Tenerife, Canary islands. Conserv Biol 21(3):853–863. https://doi.org/10.1111/J.1523-1739.2007.00686.X

Inman R, Franklin J, Esque T et al (2021) Comparing sample bias correction methods for species distribution modeling using virtual species. Ecosphere 12(3):e03422. https://doi.org/10.1002/ecs2.3422

Jarnevich CS, Stohlgren TJ, Kumar S et al (2015) Caveats for correlative species distribution modeling. Ecol Inform 29(1):6–15. https://doi.org/10.1016/j.ecoinf.2015.06.007

Johnson CJ, Gillingham MP (2008) Sensitivity of species-distribution models to error, bias, and model design: an application to resource selection functions for woodland caribou. Ecol Model 213(2):143–155. https://doi.org/10.1016/j.ecolmodel.2007.11.013

Keenan T, Maria Serra J, Lloret F et al (2011) Predicting the future of forests in the Mediterranean under climate change, with niche- and process-based models: CO$_2$ matters! Glob Change Biol 17:565–579. https://doi.org/10.1111/j.1365-2486.2010.02254.x

Koshkina V, Wang Y, Gordon A et al (2017) Integrated species distribution models: combining presence-background data and site-occupany data with imperfect detection. Methods Ecol Evol 8(4):420–430. https://doi.org/10.1111/2041-210X.12738

Leroy B, Meynard C, Bellard C et al (2016) virtualspecies, an R package to generate virtual species distributions. Ecography 39:599–607. https://doi.org/10.1111/ecog.01388

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

Mateo RG, Gastón A, Aroca-Fernández MJ et al (2019) Hierarchical species distribution models in support of vegetation conservation at the landscape scale. J Veg Sci 30:386–396. https://doi.org/10.1111/jvs.12726

Meynard CN, Kaplan DM (2013) Using virtual species to study species distributions and model performance. J Biogeogr 40(1):1–8. https://doi.org/10.1111/jbi.12006

Miller JA (2014) Virtual species distribution models: using simulated data to evaluate aspects of model performance. Progr Phys Geogr Earth Environ 38(1):117–128. https://doi.org/10.1177/0309133314521448

Mitchell P, Monk J, Laurenson L (2017) Sensitivity of fine-scale species distribution models to locational uncertainty in occurrence data across multiple sample sizes. Methods Ecol Evol 8:12–21. https://doi.org/10.1111/2041-210X.12645

Moudrý V, Komarek J, Šímová P (2017) Which breeding bird categories should we use in models of species distribution? Ecol Indic 74(March):526–529. https://doi.org/10.1016/J.ECOLIND.2016.11.006

Mugumaarhahama Y, Fandohan AB, Mushagalusa AC et al (2022) Inhomogeneous Poisson point process for species distribution modelling: relative performance of methods accounting for sampling bias and imperfect detection. Model Earth Syst Environ 8(4):5419–5432. https://doi.org/10.1007/s40808-022-01417-3

Naimi B, Skidmore A, Groen T et al (2011) Spatial autocorrelation in predictors reduces the impact of positional uncertainty in occurrence data on species distribution modelling. J Biogeogr 38:1497–1509. https://doi.org/10.1111/j.1365-2699.2011.02523.x

Naimi B, Hamm N, Groen T et al (2014) Where is positional uncertainty a problem for species distribution modelling? Ecography 37:191–203. https://doi.org/10.1111/j.1600-0587.2013.00205.x

Nolan V, Gilbert F, Reader T (2022) Solving sampling bias problems in presence-absence or presence-only species data using zero-inflated models. J Biogeogr 49(1):215–232. https://doi.org/10.1111/jbi.14268

Osborne PE, Leitão PJ (2009) Effects of species and habitat positional errors on the performance and interpretation of species distribution models. Divers Distrib 15(4):671–681. https://doi.org/10.1111/j.1472-4642.2009.00572.x

Pearson RG, Dawson TE (2003) Predicting the impacts of climate change on the distribution of species: are bioclimate envelope models useful? Glob Ecol Biogeogr 12:361–372. https://doi.org/10.1046/j.1466-822X.2003.00042.x

Pearson RG, Dawson TP, Liu C (2004) Modelling species distributions in Britain: a hierarchical integration of climate and land-cover data. Ecography 27:285–298. https://doi.org/10.1111/j.0906-7590.2004.03740.x

Phillips SJ, Dudík M, Elith J et al (2009) Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. Ecol Appl 19(1):181–197. https://doi.org/10.1890/07-2153.1

Raes N (2012) Partial versus full species distribution models. Natureza & Conservação 10:127–138. https://doi.org/10.4322/natcon.2012.020

Renner IW, Elith J, Baddeley A et al (2015) Point process models for presence-only analysis. Methods Ecol Evol 6(4):366–379. https://doi.org/10.1111/2041-210X.12352

Rocchini D, Hortal J, Lengyel S et al (2011) Accounting for uncertainty when mapping species distributions: the need for maps of ignorance. Progr Phys Geogr Earth Environ 35(2):211–226. https://doi.org/10.1177/0309133311399491

Rosenberg KV, Dokter AM, Blancher PJ et al (2019) Decline of the North American avifauna. Science 366(6461):120–124. https://doi.org/10.1126/science.aaw1313

Schank CJ, Cove MV, Kelly MJ et al (2019) A sensitivity analysis of the application of integrated species distribution models to mobile species: a case study with the endangered Baird's Tapir. Environ Conserv 46(3):184–192. https://doi.org/10.1017/S0376892919000055

Soultan A, Safi K (2017) The interplay of various sources of noise on reliability of species distribution models hinges on ecological specialisation. PLoS ONE 12(11):0187906. https://doi.org/10.1371/JOURNAL.PONE.0187906

Stephens PA, Pettorelli N, Barlow J et al (2015) Management by proxy? The use of indices in applied ecology. J Appl Ecol 52(1):1–6. https://doi.org/10.1111/1365-2664.12383

Suhaimi SSA, Blair GS, Jarvis SG (2021) Integrated species distribution models: a comparison of approaches under different data quality scenarios. Divers Distrib 27(6):1066–1075. https://doi.org/10.1111/DDI.13255

Syfert MM, Smith MJ, Coomes DA (2013) The effects of sampling bias and model complexity on the predictive performance of maxent species distribution models. PLoS ONE 8(2):e55158. https://doi.org/10.1371/journal.pone.0055158

Thibaud E, Petitpierre B, Broennimann O et al (2014) Measuring the relative effect of factors affecting species distribution model predictions. Methods Ecol Evol 5(9):947–955. https://doi.org/10.1111/2041-210X.12203

Thuiller W, Brotons L, Araújo M et al (2004) Effects of restricting environmental range of data to project current and future species distributions. Ecography 27:165–172. https://doi.org/10.1111/j.0906-7590.2004.03673.x

Titeux N, Maes D, van Daele T et al (2017) The need for large-scale distribution data to estimate regional changes in species richness under future climate change. Divers Distrib 23:1393–1407. https://doi.org/10.1111/ddi.12634

Visscher D (2006) GPS measurement error and resource selection functions in a fragmented landscape. Ecography 29:458–464. https://doi.org/10.1111/j.0906-7590.2006.04648.x

Warren DL, Glor RE, Turelli M (2008) Environmental niche equivalency versus conservatism: quantitative approaches to niche evolution. Evolution 62(11):2868–2883. https://doi.org/10.1111/j.1558-5646.2008.00482.x

Warton DI, Shepherd LC (2010) Poisson point process models solve the 'pseudo-absence problem' for presence-only data in ecology. Ann Appl Stat 4(3):1383–1402. https://doi.org/10.1214/10-AOAS331

Warton D, Renner I, Ramp D (2013) Model-based control of observer bias for the analysis of presence-only data in ecology. PLoS ONE 8(11):e79168. https://doi.org/10.1371/journal.pone.0079168

Yackulic CB, Chandler R, Zipkin EF et al (2013) Presence-only modelling using maxent: when can we trust the inferences? Methods Ecol Evol 4(3):236–243. https://doi.org/10.1111/2041-210x.12004

Yoccoz NG, Nichols JD, Boulinier T (2001) Monitoring of biological diversity in space and time. Trends Ecol Evol 16(8):446–453. https://doi.org/10.1016/S0169-5347(01)02205-4

Zellweger F, De Frenne P, Lenoir J et al (2019) Advances in microclimate ecology arising from remote sensing. Trends Ecol Evol 34(4):327–341. https://doi.org/10.1016/j.tree.2018.12.012

Zurell D, Berger U, Cabral JS et al (2010) The virtual ecologist approach: simulating data and observers. Oikos 119(4):622–635. https://doi.org/10.1111/J.1600-0706.2009.18284.X

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.