A joint distribution framework to improve presence-only species distribution models by exploiting opportunistic surveys

Juan M. Escamilla Molgora1,2 | Luigi Sedda3 | Peter Diggle2 | Peter M. Atkinson4

Abstract

Aim: The availability of data related to species occurrences has favoured the development of species distribution models using only observations of presence. These data are intrinsically biased by the sampling effort. Presence-only (PO) species distribution models (SDM) typically account for this effect by introducing additional data considered to be related with the sampling. This approach, however, does not allow the characterisation of the sampling effort and hinders the interpretation of the model. Here, we propose a Bayesian framework for PO SDMs that can explicitly model the sampling effect.

Location: Mexico.

Taxon: Pines, flycatchers (family Tyranidae), birds and plants.

Methods: The framework defines a bivariate process separable into ecological and sampling effort processes. PO data are conceived of incomplete observations where some presences have been filtered out. A choosing principle is used to separate out presences, missing data and absences relative to the species of interest and the sampling observations. The framework provides three modelling alternatives to account for a spatial autocorrelation structure: independent latent variables (model I); common latent spatial random effect (model II) and correlated latent spatial random effects (model III). The framework was compared against the Maximum Entropy (MaxEnt) algorithm in two case studies: one for the prediction of pines and another for the prediction of flycatchers.

Results: In both case studies, at least one of the proposed models achieved higher predictive accuracy than MaxEnt. The model III fit best when the sampling effort was informative, while model II was more suitable in cases with a high proportion of non-sampled sites.

Main Conclusions: Our approach provides a flexible framework for PO SDMs aided by a sampling effort process informed by the accumulated observations of independent and heterogeneous surveys. For the two case studies, the framework provided a model with a higher predictive accuracy than an optimised version of MaxEnt.

KEYWORDS
aggregated areal data, conditional autoregressive models, maximum entropy benchmark, presence-only data, sampling bias, sampling effort, species distribution models
Species distribution models (SDMs) are statistical and computational methods for characterising the distribution of organisms across space (Elith & Leathwick, 2009; Guisan & Zimmermann, 2000). SDMs have helped to develop strategies for management, adaptation and mitigation of human-induced impacts to the biosphere (Cheung et al., 2016; Foden & Young, 2016; Intergovernmental Panel on Climate Change, 2014) as well as forecasting changes in species distributions under different environmental scenarios, providing meaningful insights in which to assess biodiversity loss (Pereira et al., 2010), adaptation to climate change (Wiens et al., 2009), ecosystem management and conservation (Navarro et al., 2017) and risk of invasive species (Jiménez-Valverde et al., 2011).

SDMs are, essentially, regression models with varying methodological frameworks adapted to specific problems, particularly involving the availability of data types (e.g. presence-absence, abundance). As such, SDMs use observations of species occurrences as response variables and environmental features (hereafter called covariates) as explanatory variables (Elith & Leathwick, 2009). Framing SDMs in a regression-based context allows statistical inference of ecological concepts (e.g. ecological niche or environmental equilibrium) and provides the mathematical structure to incorporate multiple data sources. Particularly, hierarchical Bayesian models provide a clearer characterisation of uncertainties in parameter estimation, allowing the specification of flexible random effects (Cressie et al., 2009; Gelfand et al., 2006). For example, in cases where spatial autocorrelation is present, the use of Gaussian Processes (Golding & Purse, 2016) or Gaussian Markov Random Fields (GMRF) (Illian et al., 2013) can increase predictive accuracy while improving the model interpretability. However, despite their clear mathematical specification, the efficacy of regression-based SDMs depends on the quality and availability of data. Ideally, these data should represent presences and absences with a common sampling design to guarantee unbiased estimations. In practice, when the research question involves the estimation of distributions across large geographical regions, collecting unbiased presence-absence records requires a careful (and expensive) sampling design with possibly hundreds of experts deployed in the field. Surveys of this kind are atypical and usually are developed by governments or similar sized institutions that can afford the creation of full inventories or census data (e.g. Forest Inventory and Analysis (Smith, 2002) and the Mexican Inventario Nacional Forestal (CONAFOR, 2018)).

When data of presences and absences are not available across the whole study area, the assumptions on species absences should be considered carefully. Attempting to fit models with classic logistic regression using only occurrence records could yield to non-identifiable models (Ward et al., 2009). That is, with the exception of some unrealistic assumptions about absences (e.g. assuming that the absence of evidence is equivalent to evidence of absence), estimating the probability for species occurrence using solely presence-only observations can result in models with highly biased estimations (Keating & Cherry, 2004). Despite this, presence-only observations contain valuable information about species distributions and, therefore, several modelling frameworks for presence-only data have been proposed for such purposes.

One of the most popular presence-only methods is Maximum Entropy (MaxEnt) (Phillips et al., 2006), an algorithm for predicting occurrences based on the density of environmental covariates conditional to the known species presences using background data. In this method, the background data are sampled from the available area and can include presences or absence of observations. The algorithm lacks accountability for uncertainties related to an optimised abstract distribution and does not specify variation through random effects. Despite this, it can perform well in practice (Elith et al., 2006) and is still one of the most widely used methods for predicting species distributions (> 2600 articles in Web of Science at the time of writing). To account for the sampling bias in MaxEnt-based presence-only SDMs, Phillips et al. (2009) proposed the use of an informative background set (called “target group”). Typically, this informative background is obtained from records of species that potentially share the same characteristics of the target species during sampling. The use of this set showed to significantly increased the models’ accuracy (Phillips et al., 2009). Accounting the sampling bias in this manner has been addressed by Royle and Kéry (2007) in an occupancy model to account for imperfect observations (using abundance and occurrence records from census data) and by Ward et al. (2009) through the expectation maximisation (EM) algorithm to estimate the underlying presence-absence logistic signal used as presence-only observations. Both proposals contributed significantly to the development of presence-only SDMs, and were stepping stones for creating better models to account for sampling bias.

Recently, presence-only SDMs have shifted to conceive of occurrences as spatial point processes (SPP), that is, when the number of occurrences and their locations are assumed to be random. This modelling approach has been claimed to be the natural way to analyse presence-only species distributions, when data arise as point events (Renner et al., 2015; Warton & Shepherd, 2010). The SPP methodology proposes a unifying framework for species distribution informed only by presence observations, as it allows to test spatial patterns (clustered, random or inhibited) using area interaction processes (Baddeley & van Lieshout, 1995) as well as allowing prediction across space in regression-based models, such as latent Gaussian processes, to describe spatial random effects (e.g. log-Gaussian Cox processes (LGCP, Møller et al. [1998]). The focus of attention of this paper is on regression-based models. Methods related to area-interaction processes are reviewed in Diggle (2013).

While SPP models do not use background or pseudo-absence data to inform the latent distribution (i.e. intensity surface), in practice they still need to define a set of locations (named quadrature points) on which to calculate the intensity function, an equivalence issue to that of background selection (Fithian et al., 2015; Renner et al., 2015). Fitting the latent intensity process involves intense computational operations. As such, applications involving large geographical areas with large datasets are often unfeasible when the inference and uncertainty assessments are ought to be precise.
approximation methods allow scaling the model to large datasets, at the cost of (possibly) not fitting the asymptotically exact posterior distribution. Nevertheless, these methodologies have shown relevant advantages. A distinguished example is the specification of a Gaussian predicted process (GPP; Banerjee et al., 2008) to approximate a latent spatial process to model the intensity surface of the species presence. In this setting, the sampling effort is conceived as a degradation of the intensity surface. This joint modelling approach (between the sampling effort and the latent intensity surface) produced greater predictive accuracy than MaxEnt (Chakraborty et al., 2011). Although this demonstration was performed in a simulated setting (i.e. using a log-GP Cox point process), it showed the potential of jointly modelling the environmental suitability and sampling effort processes.

We agree with the view that modelling jointly both processes (i.e. environmental suitability and sampling effort) is an integrative and statistically sound approach for modelling species presences under preferential sampling using presence-only records. However, SPP models are not a silver bullet for all types of opportunistic data on occurrences nor are applicable to all case studies. For example, some observation records are not appropriate for fitting SPPs, as they need to satisfy data assumptions that are not always available, like unique geographical coordinates or, where a point representation is not enough to abstract the nature of the species of interest (discussions around this issue are in section 5). Consequently, several applications may benefit from specifying the joint approach not by focusing on the precise location of the occurrences but, rather, by aggregating them into spatial units (e.g. on a grid) and, subsequently, modelling the presences conditioned to that spatial structure. A spatial statistical model that permits a specification like this is the conditional autoregressive model (CAR, Besag, 1974). Similar to the hierarchical models of Chakraborty et al. (2011); Golding and Purse (2016), the likelihood for presence is conditioned upon a Gaussian-based spatial process (i.e. a multivariate normal prior). The key difference is that, in the case of CAR models, the latent spatial process is specified as a Gaussian Markov random field (GMRF, Rue & Held, 2005), which greatly reduces the computational complexity during inference by representing the spatial correlations as a sparse precision matrix. We consider that there is a gap for contributing with an areal-based spatial statistical framework for presence-only species distributions that can explicitly model the influence of the sampling effort. By areal based, we mean data aggregated over a spatial lattice (i.e. grid).

Here, we present a statistical framework of this kind for modelling species distributions using presence-only data (with a flexible design for defining missing data). We assume that the registered occurrences of a taxon of interest (ToI) are incomplete observations of a bivariate process that includes information about the environmental suitability (i.e. where the ToI can live) and complementary occurrence data that serve as a proxy for sampling effort, providing information on how the observations were recorded. The framework specifies three hierarchical Bayesian models that jointly specify the ecological and sampling processes. The approach provides a full description of the data generating process, giving a more direct interpretation of the parameters as well as giving explicit estimates of their uncertainties. The presented model assumes that the species populations are static in time, and in equilibrium with the environment (in the sense of Guisan & Zimmermann, 2000). Therefore, this model does not differentiate between sink populations or populations with sustained growth.

The paper is structured as follows. Section 2 describes the general specification of the framework. Here, we develop a logistic hierarchical model defined as a bivariate process that accounts for spatial random effects. In section 3, we propose two study cases for predicting presences of Pines (class: Pinopsida) and Flycatchers (family: Tyrannidae). The prediction analysis is described in sections 4.1 and 4.2, respectively. We compared the model of the framework against the MaxEnt algorithm. Finally, section 5 discusses the methodology, caveats and future research. The supplementary materials include the full framework specification and an users guide for implementing the framework in R.

## 2 | MATERIALS AND METHODS

As presence-only data lack real absences, there exists no knowledge on whether the absence of data is due to the inaccessibility of a potential sampling location or the real absence of the taxon of interest (ToI). This ambiguity suggests that presence-only data provide incomplete evidence of two underlying processes acting together. A process \( P_y \) that generates the ecological phenomenon of a taxon's occurrence, and a process \( P_x \) associated with the sampling effort or survey. As such, locations with no records of the ecological phenomenon or sampling effort indicate incomplete or missing information. Our proposal is an attempt to model these two processes using a hierarchical Bayesian framework with the aim of predicting the probability of occurrence for a ToI using presence-only data under different configurations of the spatial autocorrelation of \( X \) and \( Y \).

### 2.1 | Model summary

In general, the framework specifies a Bayesian hierarchical model that accounts for the joint effect of two components: an ecological process \( P_y \), that drives the occurrence of species of interest in the study region, and a sampling effort process \( P_x \) that models how the occurrence data were sampled. Each stochastic process includes a structural component (fixed effect) and a random effect that includes the specification of spatial autocorrelation. The model is defined on a discrete spatial lattice. Consequently, the estimations are also discrete and are defined in each area element of the lattice. The support of the model is the area element.

The presence-only data are assumed to represent realisations of a bivariate stochastic binary process (Bernoulli) separable in two components: one relative to an ecological process \( P_y \) that drives the
environmental suitability for the Tol, and another process $P_X$ related to the sampling effort. $P_Y$ and $P_X$ are modelled according to the following equations:

$$\log\left(\frac{p_y}{1-p_y}\right) = d_y^\top \beta_y + r_y$$ (1)

$$\log\left(\frac{p_x}{1-p_x}\right) = d_x^\top \beta_x + r_x$$ (2)

where $d_y$ and $d_x$ represent vectors of explanatory variables and $r_y$ and $r_x$ the random effects for $Y$ and $X$, respectively. Specifically, $d_y$ is suited for environmental variables of ecological importance, while $d_x$ should account for variables that help explain the sampling process.

The data used to fit both processes include information on known occurrences of the Tol, the sampling effort and missing observations. To predict the probability for sites with missing data, we use the data augmentation scheme proposed by Tanner and Wong (1987) and implemented by Lee (2013) in the R-Cran package CARBayes. The approach generates posterior samples of $Y$ and $X$ as well as the latent variables related to processes $P_Y$ and $P_X$ in all locations, including the ones with missing observations (i.e. $Y$ and $X$). For the sake of clarity, we had included a brief description of the employed terms in Table 1. The complete description of the model specification is explained in Supplementary Materials.

### Table 1: Definitions of the used terms and symbols

| Symbol/term                | Definition                                           |
|----------------------------|------------------------------------------------------|
| Response vector            | Input vector, each entry could be a presence, absence or missing data |
| Relative absence           | An entry for presence (1) in a list of observations (i.e. target or background) |
| Missing observation (or missing data) | An entry (N.A) in a response vector with no information about presence or relative absence |
| Target observations ($t$)  | Input (presence-only) data, used by the choosing principle, in conjunction with ($b$), to derive a response vector. |
| Background observations ($b$) | Input (presence-only) data used by the choosing principle, in conjunction with ($t$) to define a response vector. |
| Choosing principle         | A function that receives two binary observations and returns a ternary value representing presence (1), relative absence (0) or missing data (N.A) |
| $Y$                       | Response vector of the taxon of interest |
| $X$                       | Response vector of representing a sample considered to be relevant to characterise the sampling effort associated with the taxon of interest. This sample is also referred as informative sample |
| $\tilde{Y}$               | Missing observations contained in the response vector ($Y$). These values are parameters and are sampled by the MCMC procedure |
| $\tilde{X}$               | Missing observations contained in the response vector ($X$). These values are parameters and are sampled by the MCMC procedure |
| $P_Y$                     | Latent variable for ecological process |
| $P_X$                     | Latent variable for sampling effort process |
| $r_Y$ or ($R_Y$)           | Random effect (latent process) for the ecological process |
| $r_X$ or ($R_X$)           | Random effect (latent process) for the sampling process |
| $S$                       | Spatial process, a component of the random effect |
| $Z$                       | Unstructured random effect, normal distributed |

### 2.1.1 Three models for spatial variation

The proposed framework assumes that the ecological process $P_Y$ and the anthropogenic sampling process $P_X$ are conditionally independent given the random effects $R_Y$ and $R_X$. Figure 1 shows the model structure while a detailed description of the framework specification is given in Supplementary Materials. The spatial random effect is described by components $S_Y$ (Tol) and $S_X$ (sampling effort). The only source of dependency between $R_Y$ and $R_X$ is the dependency between these spatial components. In addition, each random effect incorporates an independent component for modelling unstructured variation, namely variables $Z_Y$ and $Z_X$, corresponding to $R_Y$ and $R_X$, respectively. The framework assumes that the observations of presence for the Tol and the existence of the survey (sampling) are independent when conditioned to the spatial effect. As such, the spatial autocorrelation structure is responsible for informing both processes and it acts as information transfer between the residuals of $P_Y$ and $P_X$. To test for this effect, we designed three possible models in which the spatial processes $S_Y$ and $S_X$ inform $R_Y$ and $R_X$: Model I where $S_Y$ and $S_X$ are independent, model II with one shared spatial process ($S_Y = S_X$) and model III where $S_Y$ and $S_X$ are correlated components. Schematics of the directed acyclic graphs (DAG) describing the three models are reported in Figure 1, while the full description of the framework is given in Supplementary Materials.
We are aware that estimating the real probability of occurrence using presence-only data is not possible given the inherent sampling bias of these types of data (e.g. Gelfand & Shirota, 2019; Guillera-Arroita et al., 2014). Here, we refer to environmental suitability as the spatial variation across space that determines where a species lives, settles or occupies a given area. This definition disregards the scale of the given value for a particular area. In other situations, we use the term probability of occurrence to account for the spatial variation of the ecological process (i.e. environmental suitability) in a probabilistic context, that is, where the spatial variation ranges in values from 0 to 1. To exemplify this, compare the range in values of the latent variable \( S_Y \) (spatial effect) to those of the ecological process \( P_Y \). Values in \( P_Y \) range only within the \([0, 1]\) interval.

2.1.2 | Selection of explanatory variables

Our framework is based on the Grinnellian definition of ecological niche, that is, a niche defined by non-interactive and non-consumable (scenopoetic) variables with environmental conditions changing smoothly and coarsely in space (Soberón, 2007). The selection of these explanatory variables (covariates) is crucial for the interpretability of the model and, although, the general specifications for \( P_Y \) and \( P_X \) are mathematically similar (Equations 1 and 2), they describe very different processes. \( P_X \) models the environmental suitability for a ToI to occupy the area under study. Therefore, its associated explanatory variables (\( d_Y \)) should be of ecological interest. Examples of these variables are as follows: temperature, precipitation, evapotranspiration, elevation, slope and vegetation cover. On the other hand, \( P_Y \) models the probability of a ToI to be sampled, given that it has been observed. This process is assumed to be independent from the environmental suitability and it is fully determined by anthropic variables such as distance to closest road, population density, infrastructure, political borders and land use type. The selection of covariates depends on the nature and specificities of each problem and research question. Therefore, the classification between ecological and anthropological variables is not necessarily mutually exclusive. A detailed tutorial (i.e. users guide) for preparing the data and fitting all the three models is available in Supplementary Materials and as an interactive notebook in the Github repository: https://github.com/molgor/CAR-1SDM/blob/master/CAR-1SDM/notebooks/Users-guide.ipynb

![Directed acyclic graphs for the three model specifications. Variables in squares account for observations: \( Y \): presence of a taxon of interest (e.g. species) and \( X \): presence of sample. Circles in blue correspond to latent variables while circles in grey correspond to parameters. Variables \( P_X \) and \( P_Y \) correspond to the latent processes of the sampling effort and environmental suitability, variables \( R_X \) and \( R_Y \) correspond to the random effect for the sampling effort and the environmental suitability processes, respectively. Variables \( \beta_X \) and \( \beta_Y \) represent the parameters of the fixed effects (linear components) of the latent processes \( P_X \) and \( P_Y \), respectively. Squares in salmon colour indicate environmental (\( d_X \)) and anthropic (\( d_Y \)) explanatory variables. The variables inside the dark grey block define the random effects component; different in the three models. Variables \( S_X \), \( Z_Y \), and \( Z_Y \) describe the spatial component defined as Gaussian Markov Random Fields, while variables \( Z_X \) and \( Z_Y \) represent unstructured variability within an area.](image-url)
2.2 | A choosing principle for obtaining presences, relative absences and missing observations

Estimating the probability of occurrence using solely presence-only observations necessarily requires additional assumptions about non-existent absences (Ward et al. 2009). Thus, any non-recorded presence of the taxon of interest (ToI) can potentially be a real absence (i.e. the area is not inhabited by the ToI) or an unobserved presence (i.e. the ToI inhabits the area but there is no record about it). The fundamental concept of this work is to use occurrence records of other taxa that are considered to share a similar sampling pattern as the ToI. These occurrences are used to model a sample effort process that informs about the presence and absence of the taxon of interest.

Models I, II and III specify a joint bivariate process that uses two vectors of observations as inputs; one (Y) for fitting the ecological process ($P_Y$) and another (X) for fitting the associated sampling effort process ($P_X$). These input vectors (hereafter called response vectors) are composed of k entrys, one for each area element of the spatial lattice. Each entry has assigned one of three possible states: presence (1), relative absence (0) or missing data (N.A.). As such, for a given site ($Y_i$), a state of presence indicates that the taxa of interest (ToI) has been observed. A state of relative absence (0) indicates that the surrogate taxon is present (i.e. $X_i = 1$) but the ToI is absent (i.e. $Y_i = 0$). A state of missing data (also called missing observations) indicates that neither the ToI nor the surrogate taxa are present in the site k (i.e. $X_i = 0 = Y_i$).

As we are using exclusively occurrence data, we need a function for deriving response vectors Y and X from presence-only records. We call this mechanism the choosing principle and it receives two lists as inputs: target ($t$) and background ($b$). The lists are obtained by checking the existence of an occurrence on each area element of the spatial lattice. That is, if on a given area, there exists at least one record inside, assign a 1, otherwise assign a 0. This procedure is repeated on all the k areas of the spatial lattice. Contrary to the response vectors Y and X, where each entry can be 1, 0 or N.A., the entries of $t$ (target) and $b$ (background) are binary (i.e. 0 or 1). Obtaining the missing values (N.A.) is performed by transforming t and b into response vectors Y and X using the choosing principle. There are many possibilities to define a choosing principle. Here, we used one that, for a given site $i$, assigns: missing data (N.A.) where neither the background nor target observations are present (i.e. $t_i = 0 = b_i$), 0 where there is no presence of a target observation but an observation of the background (i.e. $t_i = 0$ and $b_i = 1$), and 1 to locations where there is presence of the target taxa (i.e. $t_i = 1$). Algorithm 1 describes this choosing principle for arbitrary target and background observations and response vector.

As such, the choosing principle defines the missing data for Y and X, given the binary lists of observations of the target (i.e. species of interest) and background observations. Each response vector needs both lists of observations (target ($t$) and background ($b$)). Specifically, to obtain the response vector of the ToI (i.e. Y), the target and background list correspond to the occurrences list of the taxon of interest and a surrogate taxa (or taxon) used as informative sample. In the case of the sample observations ($X$), the target list ($t$) corresponds to the aforementioned surrogate taxa and another background list that, upon consideration of the researcher, informs about the absences of the sampling effort process. This extra background list could be, for example, all available records that likely inform the sampling effort process about sites with known observations. This approach may be useful for fitting a general purpose sampling effort process.

Choosing principle: Obtaining a response vector $R$ using background $b$ and target observations $t$ over a spatial lattice composed of $K$ area elements. Binary values are as follows: 1 if there is at least one registered occurrence, and 0 otherwise. The symbol N.A. (Not a number) is assigned to missing values.

Require:band $t$.
for $i = 1$ to $i = K; i + +$ do.
if $b[i] = 1$ then.
    $R[i] ← 1$
else
    $R[i] ← 0$
end if.
end for.

The selected choosing principle is reasonable from an ecological view. If, on average, the existence of $X$ informs the occurrence of Y, we can argue that: if a site $j$ has no background information, the probability of $X$ and $Y$ is unknown and it is informed only by nearby sites. If on the other hand, the background information exists, but there is no known occurrence (i.e. a relative absence) of Y at area $i$, the probability of occurrence for Y will depend on the presence of X as well as its nearby areas. In this sense, the probability of occurrence of a taxon (e.g. species) depends on the presence, its relative absence, its sampling effort and the nearby areas where the taxon is present. The next section provides two practical examples.

3 | APPLICATIONS

To show the capabilities of the framework, we chose two examples for predicting presences. The first involves predicting the presence of pines, that is, occurrences of the class Pinopsida as the process $P_Y$ (Pines) using the available botanical records and occurrences of the kingdom Plantae as the sampling process $P_X$ (Plants). The second example predicts the presence of a relatively abundant family of flycatchers (family: Tyrannidae) as the process $P_Y$ (Tyranids), using the available records of birds (class Aves as the sampling process $P_X$ (Birds). In both cases, we chose Elevation and Precipitation as the scenopoetic variables for process $P_X$ and Distance to roads and Population density as the anthropological variables for process $P_X$. Following the model specification in Equations 1 and 2 (Supplementary Materials),
the model for the examples of Pines and flycatchers is defined as the joint Bernoulli process:

\[
\begin{align*}
\text{logit}(Y_i)_{\text{ToI}} &= \beta_{y_1} + \beta_{y_2} (\text{Elevation})_i + \beta_{y_3} (\text{Precipitation})_i + S_Y + Z_y \\
\text{logit}(Y_i)_{\text{Sample}} &= \beta_{y_1} + \beta_{y_2} (\text{Population density})_i + \beta_{y_3} (\text{Distance to roads})_i + S_Y + Z_y \\
\end{align*}
\]  

(3)

Where the word ToI indicates that the equation is used for the taxon of interest (i.e. pines or flycatchers) and Sample indicates that the equation is valid for the sampling effort (i.e. plants or birds). Note that, in the case of model II, \( S_X = S_Y \).

### 3.1 Study region

Both models were fitted to data from the same study region. The region comprises the inland area of a circular polygon centred in central-eastern Mexico at 19 N – 97 E with radius of 2° (ca. ~ 200 km). The area covers approximately 112,000 km² and intersects several Mexican states including Veracruz, Puebla, Tlaxcala, Hidalgo, Mexico City, Morelos and Oaxaca (see Figure 2 [i]). It includes heterogeneous landscapes with variability in biodiversity, geomorphological and climatic features. The region also includes distinct biomes such as coastal dunes, chaparrals, mesophyl forests, evergreen rainforest, grasslands, mangroves, broadleaf forests and coniferous forests (INEGI, 2015; Rzedowski, 2006). The circular polygon was intersected on a grid of 4 km spatial resolution to obtain a lattice \( \mathcal{W} \) composed of 4061 areal units. This lattice was used to define the spatial structure in models I, II and III.

![Figure 2](image)

**Figure 2** A map showing the study area (overlaid semi-circular polygon) over central Mexico. Cities and other urban areas are shown in orange. Green areas represent natural protected sites with high vegetation cover

### 3.2 Occurrence data

For the presence-only data, we used the available GBIF occurrence data (GBIF Secretariat, 2015) registered before January 2015, constrained to the region \( \mathcal{W} \). The raw data were downloaded from the GBIF portal with the catalogue id: DOI:10.15468/dl.oflvla Upon downloading, we performed minimal data cleansing to remove records with missing information in any of the seven taxonomic ranks (i.e. kingdom, phylum, class, order, family, genus and species), acquisition date and collection code. We kept occurrences with identical coordinates as, historically, these occurrences might represent distinct, different records collected in a common study area. Further information about this dataset, including all data attributions, can be found in GBIF.org (2016).

We aggregated the occurrence data following the *choosing principle* described in subsection 2.2 to obtain response variables \( y, x \) according to each example. The aggregation was by the class *Pinopsida* and kingdom *Plantae*, in the Pines example and, by the family *Tyrannidae* and class *Birds* for the *Tyrannidae* case. Both examples used all known living records (*Life*) as background signal \( b \). The taxonomic classification structure used was the GBIF Taxonomic Backbone (GBIF Secretariat, 2017).

### 3.3 Treatments for missing data

To assess the impact of using missing information in the prediction accuracy of the framework, we established two different treatments for fitting each model on each example. Recalling that both response vectors \( Y \) and \( X \) have entries of presence, relative absence and missing data, we defined the following treatments:

- **treatment i**: response vectors for the ToI \( Y \) and the sample \( X \) have missing data (i.e. \( \bar{X} \neq \emptyset \neq \bar{Y} \)).
- **treatment ii**: only the sample response vector \( X \) has missing data. That is, \( \bar{X} \) is the only source of missing information.

The motivation of using treatments is that they can serve as a middle hypothesis to assess the performance of the framework under scenarios with different proportions of missing data. The recommended scenario for use in practical applications is to use treatment i. We used the ROC-AUC estimate to measure the model’s performance within treatments. Using this estimate as an absolute measure between models may lead to the wrong conclusions. For example, treatment ii implies that all the absences of \( Y \) are real and the sample \( X \) provides no information in the data augmentation methodology and therefore results in lower variance. This may lead to the conclusion that treatment ii has greater predictive accuracy than treatment i. This conclusion would be true only under the assumption that the absences of the sampling effort are in fact true absences, which, in the case of presence-only data is false. Therefore, the comparison of presence-only models using the AUC-ROC estimate is only valid as a relative measure within models that used the same data, as it penalises models that estimate potential distributions (e.g. treating absences as missing information) while
favouring those that model realised distributions where absences are informative) (Jiménez-Valverde, 2012). Comparing the AUC makes sense only when they are conditioned to a specific treatment and not between treatments.

3.4 | Explanatory variables

The elevation data used were obtained from the Global Relief ModelETOPO1 at 1 arc-minute resolution (Amante & Eakins, 2009). The precipitation data were obtained from the World Climatic Data WorldClim version 2 (Fick & Hijmans, 2017). The original data are composed in a raster model with c.a 1 km spatial resolution averaged from the years 1970 to 2000. The raster data were aggregated (by mean) to a scalar value for each areal unit in the spatial lattice equivalent to a spatial resolution of 4 km. This approach was used for the raster data. The distance to road dataset was generated from the years 1970 to 2000. The raster data were aggregated (by mean) to a scalar value for each areal unit in the spatial lattice equivalent to a spatial resolution of 4 km. This approach was used for the raster data. The distance to road dataset was generated from the years 1970 to 2000. The raster data were aggregated (by mean) to a scalar value for each areal unit in the spatial lattice equivalent to a spatial resolution of 4 km. This approach was used for the raster data. The distance to road dataset was generated from the years 1970 to 2000. The raster data were aggregated (by mean) to a scalar value for each areal unit in the spatial lattice equivalent to a spatial resolution of 4 km. This approach was used for the raster data. The distance to road dataset was generated from the years 1970 to 2000.

3.5 | Data pre-processing

The occurrences, scenopoetic and anthropological data were spatially overlaid and aggregated on each areal unit of W. The aggregation method differed according to the data type. Mean and standard deviation were used for continuous variables, mode for categorical variables and the logical AND for binary data (y, x and b). The pipeline for processing the data was undertaken with Biospytial (Escamilla Molgora et al., 2020), a geospatial knowledge engine for processing environmental data.

3.6 | Inference and prediction

We used a customised version of the R package CarBayes (Lee, 2013) and adapted it to fit models I, II and III. It includes a wrapper for easily fitting SDMs using one of the three models proposed using any type of fixed effects. The code is available from: https://github.com/molgora/CARBayeSDM. The package fits the model with a Markov Chain Monte Carlo (MCMC) method using a combination of Gibbs sampling and the Metropolis-adjusted Langevin Method (MALA) (Roberts & Tweedie, 2006). The posterior distributions were sampled by running 10,000 iterations (using 5000 for burn-in) and a thinning interval of 5. Prediction for sites with missing information was done by sampling the posterior distributions of Y and X. This same configuration was used in models I, II and III.

3.7 | Comparison between models

Models I, II and III were compared with the Deviance Information Criterion (DIC) Spiegelhalter et al. (2002). The DIC accounts for the number of parameters used and the likelihood of the observed data, given the statistical model assumed to be generating the data. The DIC is a generalisation of the Akaike information criterion (AIC) for hierarchial models. Both measure the quality of the models in terms of their accuracy and parsimony. The DIC also serves as a Bayesian-based model selection tool. Model A is preferred to model B if its DIC value is lower than the one for B (i.e. DIC_A < DIC_B).

3.8 | Comparison against MaxEnt

As mentioned in the introduction, we used the maximum entropy (MaxEnt) algorithm (Phillips et al., 2006) as a benchmark to compare the prediction accuracy of the proposed models. Contrary to models I, II and III, MaxEnt does not have a hierarchical specification and, therefore, calculating a DIC for model comparison is not possible. To address this limitation, we used a k-fold (k = 7) cross-validation methodology to measure the quality of the predictions of all models. That is, on each fold, 1/7th of the data was excluded from the fitting process and used as testing data to be compared against the corresponding predictions. This procedure was performed seven times, until every observation had a corresponding predicted value. We then used the receiver operator characteristic (ROC) curve and its area under the curve (AUC) (Fielding & Bell, 1997) as a measure of prediction accuracy. The same sevenfold cross-validation was performed for models I, II and III with the difference that the excluded data were treated as missing data. The ROC/AUC values, as well as their corresponding 95% confidence intervals, were calculated with the R package pROC (Turck et al., 2011).

Recalling that the proposed models are based on a spatial lattice structure (i.e. a CAR-based model), the spatial variation is modelled on a finite set of areal units. In the following case studies, these units were defined as square cells on a regular grid with approximately 4 km spatial resolution. To make a fair comparison, we used the same spatial resolution and environmental values for fitting the MaxEnt models. Additionally, the background data (i.e. pseudo-absences in the MaxEnt jargon) used for fitting MaxEnt were obtained from locations with sampling observations but with no record of the taxon of interest, similar to the sample selection bias for background data proposed by Phillips et al. (2009). In other words, the choosing principle was also applied to the MaxEnt models resulting in the same input for all models (only valid for component Y [presence] of models I, II and III).
3.8.1 | MaxEnt optimisation

MaxEnt allows different configurations for model fitting. The most important are: the regularisation factor (reg) and the composition of mathematical transformations of the covariates, the so-called features (see: Merow et al., 2013). These features are equivalent to functions of the trend (i.e. they modify the fixed effect). To optimise the predictions of MaxEnt, we ran the sevenfold cross-validation using different combinations of regularisation factors (reg ∈ (0.1, 1.50)) and feature functions. In the case of the features, we used single and paired combinations of each of the following types: linear (l), quadratic (q), product (p), threshold (t) and hinge (h). The total number of different combinations (i.e. models) for MaxEnt was 2250. The model was fitted with the R package maxnet (Phillips et al., 2017).

4 | RESULTS

4.1 | Presence of pines

We performed the methods described in section 2.2 to obtain response variables for Pines (Pines) and the botanical sample (Plants) using a geographical lattice composed of 4060 cells (or unit areas). For the presence observations, 341 (8.4%) cells have known occurrences (class Pinopsida), 2559 (63%) have relative absences and 1160 (28.6%) are unknown (locations with missing observations). For the sample observations (botanical records), 2900 (71.4%) cells have known occurrence, 430 (8.4%) have relative absence and 730 (18%) unknown information (missing data).

The optimal MaxEnt, in terms of its higher predictive accuracy, measured by the AUC-ROC was the one with combined quadratic and product feature types (pq, Figure 4a) and regularisation factor of 5.1. This combination, however, achieved the lowest prediction AUC of 0.8 ± (0.78,0.82) 95% confidence interval (CI), when compared with models I, II and III (see Figure 4a). Results from the best MaxEnt model and Models I, II and III are described in Table 2.

For the treatment i (i.e. with both sources of missing information, see section 3.3), Model III (the one with correlated spatial structures) was the best ranked, that is, it achieved the lowest DIC of 3440.2 (see Table 2). The predictive accuracy of this model, measured as the area under the ROC curve (i.e. AUC-ROC), was the highest of all three models: 0.838 ± (0.814, 0.862) 95% CI (see Figure 4a). The AUC of the three models fell within a common 95% credible interval of [0.81,0.86], that is, the predictive accuracy of models I, II and III was not significantly different.

Treatment ii (i.e. the one with no missing data in the sample effort component) produced slightly different results. In this case, Model I (independent spatial effects) was the best ranked by achieving the lowest DIC value (3421.2). The AUC in all models was higher than those on treatment i. However, in a similar way, all of these values fell within a common 95% credible interval of [0.85, 0.89] (see ROC figure in Supplementary Materials). Possible reasons for this effect are explained in the next section. Additionally, the ROC curves in all models show similar variance described as the envelope of the ROC curve. Figures of this matter are shown in the Supplementary Materials. The framework allows testing the significance of the model’s parameters, in the same form as a Bayesian linear regression. In this sense, the variable Distance to road was found to be the only significant covariate common to models I, II and III. That is, the zero is out of the 95% credible intervals (CI) of its posterior distribution. The scenopoeitic variables (elevation and precipitation) were only significant in Model II. The selection of these specific covariates was done solely to demonstrate the capabilities of the model. As such, other covariates with stronger significance may be preferred in case-based applications.

4.1.1 | Spatial results

Figure 3 shows the mean predicted latent surfaces for the presence of Pines PX and sampling effort PX in all three models (left and right columns, respectively). PX shows higher probability of occurrence than PX across all the region. This is consistent in the three models. In contrast, the presence PX revealed clustered patterns of high probability (Figure 3). Of particular interest is the central zone that shows a high probability of occurrence. This area corresponds to the contact between the Eastern Sierra Madre and the Volcanic Axis and is of high elevation and high precipitation. In contrast, the MaxEnt model (fig: 3, bottom left panel) produced a smoother surface. The orographic features are more defined and the clustered patterns for presence are lost. Visual comparison between the models is difficult because of their similarity. However,

|        | DIC  | ROC-AUC  | 95% CI       | DIC* | ROC-AUC* | 95% CI*     |
|--------|------|----------|---------------|------|----------|-------------|
| Model I | 3517.6 | 0.835    | [0.81, 0.86]  | 3421.2 | 0.874    | [0.85, 0.89] |
| Model II| 3665.9 | 0.807    | [0.77, 0.84]  | 3647.9 | 0.842    | [0.83, 0.86] |
| Model III| 3440.2 | 0.838    | [0.81, 0.86]  | 3505.9 | 0.876    | [0.86, 0.89] |
| MaxEnt  | –     | –        | –             | 0.8   | –        | [0.78, 0.82] |
in treatment II (only one source of missing observations), Model II shows the compromise of estimating the sample $P_X$ while satisfying a common spatial component with $P_Y$ (for a clearer demonstration, refer to Supplementary Materials). In Model III, the median correlation obtained from the cross variance ($\Sigma$), between the presence of pines ($P_Y$) and the sampling effort ($P_X$), was 0.97 with (0.9,0.99) 95% credible interval. This result is consistent with the fact that the taxon of interest (i.e. pines) is totally contained in the sampling effort (i.e. plants). The complete estimates summary is given in Supplementary Materials.

### 4.2 | Results for the presence of flycatchers (family Tyrannidae)

The flycatchers case study was performed in the same study region (i.e. across the lattice $\mathcal{W}$). However, the data availability was significantly different and, therefore, the results are also different. In this example, we obtained 596 (14.6%) cells with known occurrences of flycatchers, 368 (9.1%) with relative absences and 3096 (76.2%) of unknown or missing information. The occurrences for the sample (birds in general) were composed of 990 (24.4%)
known occurrences, 2340 (57.6%) relative absences and 730 (18%) missing data.

The optimal MaxEnt, in terms of its higher predictive accuracy measured by the AUC-ROC was the one with a combination of linear (l), quadratic (q), product (p), threshold (t) and hinge (h) with pairwise combinations). The values in the y-axis correspond to the resulting AUC-ROC value according to that specific pair of parameters. The AUC-ROC values of models I (red), II (green) and III (blue) are shown as horizontal lines. The black curves are smooth interpolations (GAM) of a specific feature class of MaxEnt as a function of a regularisation factor, shaded areas correspond to the confidence interval of the GAM smoothing.

**TABLE 3** Comparison of the presence-only models: Independent Spatial Components (Model 1), Common Spatial Component (Model 2), Correlated Spatial Components (Model 3) and Maximum Entropy (MaxEnt) for the presence of the family **Tyrannidae** using birds as sample (class: Aves). A sevenfold cross-validation was performed to calculate the area under the receiver-operating characteristic curve (ROC-AUC) as a measure of quality for each model. Models with the ⋆ symbol were fitted using only missing data from X (sample) (i.e. treatment **ii**).

| Model   | DIC     | ROC-AUC | 95% CI        | DIC*    | ROC-AUC* | 95% CI* |
|---------|---------|---------|---------------|---------|----------|---------|
| Model I | 4445.8  | 0.556   | [0.47, 0.64]  | 5607.3  | 0.89     | [0.88, 0.91] |
| Model II| 4251.1  | 0.77    | [0.71, 0.84]  | N.A.    | N.A.     | N.A.    |
| Model III| 3905.0  | 0.54    | [0.45, 0.62]  | 3331.1  | 0.95     | [0.94, 0.96] |
| MaxEnt  | –       | –       | –             | –       | 0.72     | [0.69, 0.74] |

known occurrences, 2340 (57.6%) relative absences and 730 (18%) missing data.

The optimal MaxEnt, in terms of its higher predictive accuracy measured by the AUC-ROC was the one with a combination of linear and hinge feature types (nknots = 50), and a regularisation factor of 5.5 with a ROC-AUC of 0.72 ± (0.69,0.74) 95% confidence interval (CI). The optimal parameter combination was equivalent to models I and III in terms of its predictive accuracy. That is, all the MaxEnt models are covered by the 95% confidence intervals of the ROC-AUC estimation for models I, II and III. Nevertheless, Model II (the one with a common spatial random effect) was significantly more accurate than the rest of the models. **Figure 4b** shows a comprehensive view of the aforementioned results. Additionally, a quantitative summary of these results is described in **Table 3**.

In treatment **i** (i.e. missing data in both response vectors, for presence and for sample), Model III (correlated spatial components between the ecological process and the sampling effort) was the best ranked, achieving the lowest DIC value (3905), similar to the Pines example. However, its accuracy in terms of ROC-AUC was close to random classification, reaching an AUC of 0.54 with ± (0.45,0.62) at 95% CI. Model I (independent spatial effect for the ecological and the sampling components) obtained similar values of ROC-AUC (0.57 ± (0.48,0.65) at 95% CI). In contrast, Model II obtained the highest predictive accuracy (0.77 ± (0.71,0.84)) with a DIC of 3905, second in rank (see **Figure 4b**). In addition, models I and III achieved a low predictive power compared to the benchmark model (MaxEnt).

Treatment **ii** (i.e. only one response vector (X) with missing information) produced contrasting results. Although model III (correlated components) ranked best, in terms of a lowest DIC (3331.1), its AUC was 0.95 ± (0.94,0.96). Model I (independent spatial components) followed with an AUC of 0.89 ± (0.88,0.91). Model II could not obtain valid posterior distributions, as its log-likelihood diverged to −∞. We discuss possible reasons and circumventing strategies in the next section.
All results are shown in Table 3. Based solely on the DIC, Model III was ranked first in both treatments. However, in cases with large proportions of missing data (as in treatment i with 76.2% cells), the prediction accuracy (ROC-AUC) was low. This effect highlights the importance of selecting informative missing data as well as the type of model to use. These issues are explored further in the discussion section.

The covariate Distance to roads was found to be significant in models I and III. The rest (elevation, precipitation and population count) were not significant in all three models. The selection of these specific covariates was done solely to demonstrate the capabilities of the model. As such, other covariates with stronger significance may be found.

4.2.1 | Spatial results

Figure 5 shows the mean predicted latent surfaces for the presence of flycatchers $P_Y$ (Tyranids) and relative sample $P_X$ (Birds) in all three models.
(left and right columns, respectively). Model I presents a clear difference between $P_r$ and $P_x$ (Figure 5, first row). In this case, $P_r$ appears more smooth with patches of lower probability, although always with probability higher than 0.2. The surface $P_x$ in model I (Figure 5, top right panel) has clear shaped patterns with contrasting probabilities between interior regions (pocket shapes). This feature is present in both surfaces of model II (Figure 5, second row) and model III (Figure 5, third row). The fixed effects (covariates) for $P_r$ and $P_x$ are close to zero. Therefore, the spatial variation is driven only by the common structure $S$. In the case of model III, the sample surface $P_x$ presents greater connectivity and higher probabilities in places with known observations. Both surfaces, however, present a similar structure in shapes and patterns.

In contrast, the MaxEnt prediction lacks the random spatial effect component. The resulting probability surface is determined exclusively by the features derived from the covariates. Although it is possible to distinguish spatial patterns within the region, the predicted probability is in general close to a uniform random classification (i.e. 0.5). This effect is supported by the obtained AUC-ROC value of the cross-validation analysis (0.6) (fig: 4b [a]). In Model III, the median correlation, obtained from the cross variation (2) between the presence of flycatchers ($P_{nv}$) and the sampling effort ($P_x$), was 0.996 with (0.993,0.998) 95% credible interval. As in the latter example, this result is consistent with the fact that the taxon of interest (i.e. flycatchers) is totally contained in the sampling effort (i.e. birds). The complete estimates summary is shown in Supplementary Materials.

5 | DISCUSSION

The bivariate CAR modelling framework uses an additional source of information, apart from the presences of the target species. This extra information comes from sampling observations related to other species and other taxa that, according to the modeller, give complementary information relative to the occurrence of the taxon of interest (ToI). The framework relies on three fundamental concepts: (i) the sampling effort as complementary information for inferring the probability of presence, (ii) the spatial autocorrelation structure for determining the variability and occurrences likelihood across the landscape and (iii) the choosing principle, a mechanism for determining presences, relative absences and missing data from presence-only records. Both examples showed that, at least one of the three proposed models outperformed MaxEnt. The results in Tables 2 and 3 show that the models’ goodness-of-fit statistic (i.e. DIC) and predictive accuracy increased in treatment ii, that is, when the absence of records were treated as real absences. This is expected because assuming missing data as real absences reduces uncertainty.

These results show that the proportion of missing data plays a fundamental role in the predictive capability of the model. This effect is recognised in the flycatchers example, where the proportion of missing observations is much higher (76% of the total number of regions) compared to presences and relative absences. In this case, models I and III produced low predictive accuracy similar to random classification, while MaxEnt obtained an AUC-ROC of near 0.72. Although model II ranked second in terms of DIC, it achieved the highest predictive accuracy (AUC-ROC), outperforming MaxEnt. This result is also supported by the high number of missing data (increased uncertainty) and reduced number of spatial parameters to fit. In terms of model parsimony, one shared spatial latent effect (model II) has less parameters to fit compared with two spatial effects in the case of models I and II.

The three proposed models impose different restrictions on how the spatial autocorrelation structure affects the probability of a species to occur. The more complex the spatial structure is, the more presence-only observations (and less missing data) are needed. This can be modulated by the amount of missing data with respect to the relative absences determined by the sampling effort observations and the choosing principle. Consequently, using an appropriate informative sample becomes crucial for obtaining accurate inferences and predictions. This finding highlights interesting paths for future research: one related to the selection of informative observations for the sampling effort process, and the other for different choosing principles.

Model II may be a better alternative for taxa with sparse spatial distributions and large proportions of missing data. Nevertheless, model II presented problems with identifiability in treatment ii (i.e. missing data only in the ToI observations and assumed real absences in the sampling process). A possible reason is that the inference method could not find a suitable compromise in accounting for a common spatial effect that had two constraints. One, the accountability of residuals of both processes ($P_{rv}$ and $P_{px}$) and two, the restrictions imposed by the intrinsic CAR model specification. That is, the sum of the random effect on all the lattice areas should sum to one. A possibility to circumvent this last restriction is to specify, instead, a proper CAR model (e.g. Leroux et al., 2000). The package CARBayes (Lee, 2013) allows this specification. We recommend the practitioner to compare the three models accordingly to fit specific needs.

5.1 | The role of the choosing principle

When presence-only data are used, any choosing principle is inevitably a source of potential bias. Thus, the research question and the selection of the sampling effort observations play a fundamental role in determining the accuracy of predictions. The way relative absences and missing data are derived implies ecological assumptions that should be kept in mind when one tries to model species (taxon) distributions. For example, following the biotic, abiotic, movements (BAM) diagram proposed by Soberon and Nakamura (2009), if the objective is to model the realised distribution (i.e. places where the species lives in reality), absences become informative. If on the other hand, the objective is to model the species’ potential distribution (i.e. places where it can survive and thrive due to suitable environmental conditions), absences may constitute missing data. See equivalent concepts from a SDM approach (Jiménez-Valverde et al., 2008).

In our framework, we used the sample observations $X$ together with the choosing principle to discriminate between informative absences and missing data. If the sampling effort is chosen to be informative, it can increase significantly the accuracy of predictions (see Table 2).
The current choosing principle assumes that for every location \( k \), if the ToI (e.g., species) is not present, but the sample observation exists \( (X_k = 1) \), then the ToI is assumed to be absent \( (Y_k = 0) \). In some applications, this assertion may be incorrect and, if the sample observations \( X \) consist as well of presence-only data, the bias in false absences can propagate in both processes. This problem is present in all presence-only methods that tries to account for the sampling bias using pseudo-absences (e.g., target-background approach of Phillips et al. (2009)), given the intrinsic bias of the collected data. Ideally, the best way to rank distinct choosing principles, given a ToI, is using presence-absence data. The proposed choosing principle is not intended to be a general rule for all species and problems. It is encouraged that the modeller considers other definitions of a choosing principle. For example, another type of choosing principle can incorporate information on other species features, such as movement, since the accessibility of an area can be indicative of poor sampling and its use has been shown to reduce bias in occurrence data (Monsarrat et al., 2018).

5.2 An alternative to spatial point processes in real-world applications

As mentioned in the introduction, there is a generalised consensus in the SDM community that spatial point processes (SPP) are a “natural way” to model presence-only species distributions (Renner et al., 2015; Warton & Shepherd, 2010). We agree that this framework provides an elegant (and easy to interpret) approach in ideal circumstances, that is, when the presence-only data provide enough information about their context and sampling design to satisfy the assumptions of the SPP. The assumptions vary from one specification to another. However, necessary assumptions of the standard model include: (i) point locations are measured exactly, (ii) no two points can share the same location and (iii) the survey is exhaustive within the study region, that is, “there are no errors in detecting the presence of the random process” (Baddeley, 2010). The first assumption is difficult to achieve for several kinds of observation records. For example, if the target species are catalogued as endangered or threatened, data providers (e.g. GBIF or the IUCN Red List of Threatened Species) will not disclose their exact location to protect the organisms from poaching or other illegal forms of exploitation. Assumption ii is not applicable to all types of records either. Several ecological studies assign a common location to a given set of observations, due to the difficulties in field or other logistic obstacles. This was also a common practice in records prior to the widespread availability of GPS receivers. Still nowadays, depending on the case study, measuring each specific location for a given individual may be infeasible. Within the SPP framework, the event of each record is not the species presence but the recording of the species, similar to our joint modelling proposal. However, assumption iii implies that all records are exhaustive within the study region, that is, there are no detection errors. Again, this is not applicable for all species or records. Take as an example the records obtained by citizen observations or those of camera-traps. In both cases, the occurrences are registered only within the range of the observer (or camera) and not exhaustively across the region. Additionally, each occurrence is classified by a human (or an algorithm) with varying degrees of error. Lastly, SPPs treat locations as random point events. As such, they disregard any information related to the physical size of the taxa. Neglecting this information may lead to artefacts when accounting for interactions not caused by ecological phenomena (arguably the focus of interest) but simply, by modelling physical objects with zero size. See Isaac and Pocock (2015) for an extensive review of the biases, potential information and methodological caveats regarding the use multiple collections of biological records.

As such, using SPPs for modelling presence-only distributions using a collection of heterogeneous surveys (arguably taken opportunistically) will possibly lead to contradictions of the SPP model assumptions. In such circumstances, our approach offers a viable alternative, that of modelling presences as occupied areas over a fixed spatial lattice (i.e. grid or tessellation), rather than adapting each SPP model to a particular use. As such, the presented framework proposes a generic alternative for modelling species distributions when using large and diverse datasets.

5.3 Advantages of the framework

The model is defined on a fixed spatial lattice. Observations that occur on a given area element can be aggregated to reflect presences or abundances. The data aggregation eases the computational complexity of accounting for each individual data point (as opposed to Gaussian processes or SPP). Instead, all records contained within the boundaries of a given area unit (cell) are aggregated into binary or (in later implementations) count or continuous data. This approach, in conjunction with the sparse representation of the covariance matrix, makes it an ideal modelling framework for species distributions over wide areas using large and heterogeneous datasets. In addition, the probabilities for presence in areas that have not been sampled can be inferred from the neighbouring areas. The method is able to infer probabilities for places where data availability is limited. The model specifies a Bayesian hierarchical model and, thus, accounting for uncertainties in the parameters is possible. This brings the possibility to perform hypotheses testing on the posterior sample. As it is a hierarchical model, it is also possible to perform model selection using the DIC statistic. The structural components of the models (i.e. the ecological process and the sampling effort) can be modelled explicitly using different covariates and even feature classes, as the ones used by MaxEnt. Lastly, the choosing principle provides a flexible approach with which to assign absences and missing data.
5.4 Limitations

Manipulating the spatial random component of the model implies greater computational complexity on the order of $O(n^3)$ (in the worst scenario). Although the matrix is sparse and the inference uses optimised numerical methods that can reduce the computational complexity, the numerical methods involved are more intensive than MaxEnt or other models that are not based on hierarchical Bayesian inference. This is a limitation for studies that require extended regions involving hundreds of thousands of area elements.

Another limitation is that the specification of the spatial effect is based on discrete spatial distributions. This implies that, once the model is fitted, it is not possible to make predictions on unobserved locations or data (as opposed to geostatistical models). Depending on the specification, a modeller may need the spatial random effect to be continuous in space, instead of over a discrete lattice. If this is the case, we recommend the use of alternative SPP-based models (Renner et al., 2015; Renner et al., 2019; Velázquez et al., 2016).

ACKNOWLEDGEMENTS

This project was jointly sponsored by the Mexican Council of Science and Technology (CONACYT) under the doctoral program: Becas al Extranjero and the Faculty of Science and Technology from Lancaster University. 

CONFLICT OF INTEREST

The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

A snapshot of the code and data are stored in the FigShare https://doi.org/10.6084/m9.figshare.19093775.v1 for reproducibility. Updated versions of the package and user’s guide are hosted in: https://github.com/molgor/CARBayeSDM. The software and the data used are released under the GNU General Public Licence (GPL V.3) and, therefore, does not need any permission for use, copy, modify or distribute.

ORCID

Juan M. Escamilla Molgora https://orcid.org/0000-0002-3682-9828

Luigi Sedda https://orcid.org/0000-0002-9271-6596

Peter Diggle https://orcid.org/0000-0003-3521-5020

Peter M. Atkinson https://orcid.org/0000-0002-5489-6880

REFERENCES

Amante, C., & Eakins, B. (2009). ETOPO1 1 arc-minute global relief model: Procedures, data sources and analysis. Technical report March.

Baddeley, A. J. (2010). Modelling strategies. In A. E. Gelfand, P. J. Diggle, M. Fuentes & P. Guttorp (Eds.), Handbook of spatial statistics (Chapter 20, pp. 340–369), CRC Press.

Baddeley, A. J., & van Lieshout, M. N. (1995). Area-interaction point processes. Annals of the Institute of Statistical Mathematics, 47(4), 601–619.

Banerjee, S., Carlin, B. P., & Gelfand, A. E. (2014). Hierarchical modeling and analysis for spatial data (2nd ed.), Chapman and Hall/CRC.

Banerjee, S., Gelfand, A. E., Finley, A. O., & Sang, H. (2008). Gaussian predictive process models for large spatial data sets. Royal Statistical Society (Series B), 70(4), 825–848.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. Journal of the Royal Statistical Society. Series B (Methodological), 36(2), 192–236.

Chakraborty, A., Gelfand, A. E., Wilson, A. M., Latimer, A. M., & Silander, J. A. (2011). Point pattern modelling for degraded presence-only data over large regions. Journal of the Royal Statistical Society. Series C: Applied Statistics, 60(5), 757–776.

Cheung, W., Rondonini, C., Avtar, R., van den Belt, M., Hickler, T., Metzger, J., Scharlemann, J., Velez-Liendo, X., & Yue, T. (2016). IPBES Scenarios and Models of Biodiversity and Ecosystem Services. Secretariat of the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services.

CONAFOR. (2018). Inventario Nacional Forestal y de Suelos 2009-2014. Inventario Nacional de Suelos (p. 432).

Cressie, N., Calder, C. A., Clark, J. S., Ver Hoef, J. M., & Wikle, C. K. (2009). Accounting for uncertainty in ecological analysis: The strengths and limitations of hierarchical statistical modeling. Ecological Applications, 19(3), 553–570.

Diggle, P. J. (2013). Statistical analysis of spatial and spatio-temporal point patterns. Chapman and Hall/CRC.

Elith, J., Graham, C. H., Anderson, R. P., Dudik, M., Ferrier, S., Guisan, A., Hijmans, R. J., Huettmann, F., Leathwick, J. R., Lehmann, A., Li, J., Lohmann, L. G., Loiselle, B. A., Manion, G., Moritz, C., Nakamura, M., Nakazawa, Y., Overton, J. M., Peterson, A. T., ... Zimmermann, N. E. (2006). Novel methods improve prediction of species’ distributions from occurrence data. Ecography, 29(2), 129–151.

Elith, J., & Leathwick, J. R. (2009). Species distribution models: Ecological explanation and prediction across space and time. Annual Review of Ecology, Evolution, and Systematics, 40(1), 677–697.

Escamilla Molgora, J. M., Sedda, L., & Atkinson, P. M. (2020). Biospytal: Spatial graph-based computing for ecological Big Data. GigaScience, 9(5), 1–25.

Fick, S., & Hijmans, R. (2017). Worldclim 2: New 1-km spatial resolution climate surfaces for global land areas. International Journal of Climatology, 37, 4302–4315.

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

Fithian, W., Elith, J., Hastie, T., & Keith, D. A. (2015). Bias correction in species distribution models: Pooling survey and collection data for multiple species. Methods in Ecology and Evolution, 6(4), 424–438.

Foden, W. B., & Young, B. E. (2016). IUCN SSC guidelines for assessing species’ vulnerability to climate change. Technical report, Cambridge, UK.

GBIF Secretariat. (2015). Global biodiversity infrastructure. GBIF Secretariat. (2017). GBIF backbone taxonomy. GBIF.org. (2016). GBIF occurrence download.

GDAL/OGR Contributors. (2018). GDAL/OGR - Geospatial Data Abstraction software Library.

Gelfand, A. E., Shirota, S. (2019). Preferential sampling for presence/absence data and for fusion of presence/absence data with presence-only data. Ecological Monographs, 89(3), e01372.

Gelfand, A. E., Silander, J. A., Wu, S., Latimer, A., Lewis, P. O., Rebelo, A. G., & Holder, M. (2006). Explaining species distribution patterns through hierarchical modeling. Bayesian Analysis, 1(1 A), 41–92.

Golding, N., & Purse, B. V. (2016). Fast and flexible Bayesian species distribution modelling using Gaussian processes. Methods in Ecology and Evolution, 7(5), 598–608.
BIOSKETCH
Juan Escamilla Mólgora is interested in developing computational and statistical methodologies for studying life's spatial patterns at different scales. This work was part of his PhD research at Lancaster University on the development of a computational and statistical framework for modelling species distributions using presence-only data. The co-authors collaborate in developing spatial statistical methods applied to epidemiological and environmental problems.

Author Contributions: All authors developed the general framework and provided critical feedback in all the stages of this work. More specifically, PD proposed the three model specifications. PA proposed the choosing principle. LS and JEM designed the modelling and simulations strategies. JEM prepared the data, implemented the models, performed the analysis and visualisations and wrote the manuscript with inputs and edits from all co-authors. PA, LS and PD supervised the project.

SUPPORTING INFORMATION
Additional supporting information may be found in the online version of the article at the publisher’s website.

How to cite this article: Escamilla Molgora, J. M., Sedda, L., Diggle, P. & Atkinson, P. M. (2022). A joint distribution framework to improve presence-only species distribution models by exploiting opportunistic surveys. Journal of Biogeography, 49, 1176–1192. https://doi.org/10.1111/jbi.14365