Integrating Presence-only and Abundance Data to Predict Baobab (Adansonia digitata L.) Distribution: A Bayesian Data Fusion Framework

Akoeugnigan Idelphonse Sode

sdidelphonse@gmail.com

Laboratoire de Biomathématiques et d'Estimations Forestières, Faculté des Sciences Agronomiques, Université d'Abomey-Calavi

Adande Belarmain Fandohan

Unité de Recherche en Foresterie et Conservation des Bioressources, Ecole de Foresterie Tropicale, Université Nationale d'Agriculture

Elias Teixeira Krainski

Program, Computer, Electrical and Mathematical Sciences and Engineering Division, King Abdullah University of Science and Technology

Achille Ephrem Assogbadjo

Laboratoire d'Ecologie Appliquée, Faculté des Sciences Agronomiques, Université d'Abomey-Calavi

Romain Glèlè Kakaï

Laboratoire de Biomathématiques et d'Estimations Forestières, Faculté des Sciences Agronomiques, Université d'Abomey-Calavi

---

Research Article

Keywords: Log-Gaussian Cox process, count data, ISDM, INLA, joint spatial modelling, shared latent components

Posted Date: October 30th, 2025

DOI: https://doi.org/10.21203/rs.3.rs-7871875/v1

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at Environmental and Ecological Statistics on June 15th, 2026. See the published version at https://doi.org/10.1007/s10651-026-00737-2.

Integrating Presence-only and Abundance Data to Predict Baobab (*Adansonia digitata L.*) Distribution: A Bayesian Data Fusion Framework

Akoeugnigan Idelphonse Sode[1*], Adande Belarmain Fandohan[2], Elias Teixeira Krainski[3], Achille Ephrem Assogbadjo[4], Romain Glèlè Kakaï[1]

[1*]Laboratoire de Biomathématiques et d'Estimations Forestières, Faculté des Sciences Agronomiques, Université d'Abomey-Calavi, Abomey-Calavi, 04 BP 1525, Cotonou, Benin.
[2]Unité de Recherche en Foresterie et Conservation des Bioressources, Ecole de Foresterie Tropicale, Université Nationale d'Agriculture, Kétou, BP 43, Kétou, Benin.
[3]Program, Computer, Electrical and Mathematical Sciences and Engineering Division, King Abdullah University of Science and Technology, Thuwal, 23955-6900, Saudi Arabia.
[4]Laboratoire d'Ecologie Appliquée, Faculté des Sciences Agronomiques, Université d'Abomey-Calavi, Abomey-Calavi, 01 BP 526, Cotonou, Benin.

*Corresponding author(s). E-mail(s): sdidelphonse@gmail.com;
Contributing authors: bfandohan@gmail.com; eliaskrainski@gmail.com; assogbadjo@yahoo.fr; glele.romain@gmail.com;

Abstract

Species distribution models (SDMs) are vital tools in ecology and conservation. The integration of increasingly available citizen science data with planned survey data offers a significant opportunity to enhance species distribution estimates. While integrated SDMs often combine presence-only and abundance data, the interdependence between the conditional distributions of these outcomes remains

to be elucidated. This study proposes a Bayesian spatial fusion modelling framework to jointly analyse presence-only and abundance data for the African baobab in Benin. The aim was to understand and map the spatial variation in the species' distribution. We briefly reviewed process-based models for count and point process data and explored various data fusion strategies using Integrated Nested Laplace Approximations (INLA) and Stochastic Partial Differential Equations (SPDE) for inference. The results revealed a heterogeneous baobab distribution across Benin, characterised by a spatial autocorrelation range of 34.4 km (95% Bayesian credible interval, BCI = 27.59-42.52). Key drivers of this distribution include environmental factors such as annual temperature, rainfall of the driest month, soil texture (silt/clay fractions), and slope. A spatial fusion model incorporating shared latent components and common covariates' effects demonstrated the highest performance level, surpassing alternative fusion approaches. The model achieved the highest mean composite scores for the Area Under the ROC Curve (AUC) (0.85±0.02), accuracy (0.77±0.02), and True Skill Statistics (TSS) (0.66±0.05). The shared component model has the capacity to explain datasets' interdependence, estimate covariate effects missed by separate models, and enhance prediction precision. Despite relying on the assumption of an identically shared spatial signal across target responses, this research underscores the potential of spatial fusion modelling for integrating disparate data sources. The findings contribute to advancing SDM inference, particularly in data-limited contexts, and have wider applicability to spatial regression problems involving multisource outcomes.

Keywords: Log-Gaussian Cox process, count data, ISDM, INLA, joint spatial modelling, shared latent components

1 Introduction

Species distribution models (SDMs) are essential to ecology, evolution, biogeography, and conservation biology. They are particularly useful for understanding environmental relationships and predicting species distribution, projecting climate change effects (Elith and Leathwick 2009; Guillera-Arroita et al. 2015), assessing the effectiveness of protected areas in conserving species communities (Adjahossou et al. 2016), or identifying areas with a high risk of invasion by exotic species (Fandohan et al. 2015). Ecologists usually build SDMs from various data sources, particularly presence-absence data (i.e. binary response at a set of sampled locations), presence-only data (i.e. incompletely observed point process), or count data (i.e. number of individuals

at a set of sampled locations). In addition, for presence-only (PO) data, *location* is a point, whereas, for the other two types of data, *location* is presumably a finite area (e.g., a circle of pre-specified radius) and can then be called *site*. While presence-only data are often collected through opportunistic sampling and are widely available from citizen science programs (eBird, iNaturalist, biodiversity databases, etc.), presence-absence (PA) and abundance data are usually obtained through planned survey designs throughout the study region (Giraud et al. 2016; Dorazio 2014).

Multiple data sources have become increasingly available due to recent technological and digital developments (Guillera-Arroita et al. 2015; Giraud et al. 2016). However, this comes with additional challenges for species distribution modelling (SDM). Usually, presence-only data are subject to sample selection bias, while planned survey data, such as abundance outcomes, are more costly and consequently suffer from the limited amount of data and the restricted geographic extent of the sampling design. Some data sources may have biases that can be reduced or eliminated by combining an additional data source that does not have similar limitations. Thus, recent advances in SDM motivate the use of integrated species distribution models (ISDMs) to address these challenges (Fletcher Jr. et al. 2019; Foster et al. 2024; Mostert and O'Hara 2023). ISDM, also known as model-based data integration (or fusion), is an innovative statistical approach that combines spatial data from multiple sources to understand and predict species distribution. Its strength lies in its ability to share parameters between submodels, allowing better estimation of unknown parameters under investigation than those of independent models (Isaac et al. 2020; Fletcher Jr. et al. 2019).

In recent years, several studies have proposed ISDM to demonstrate how ecological data from various sources can be combined to leverage the strengths of each dataset.

Dorazio (2014) developed a hierarchical statistical model to analyse presence-only data in conjunction with count data observed in repeated planned surveys using the maximum likelihood approach and found an improvement in parameter estimation over traditional approaches based on analysing both data sets separately. Fithian et al. (2015) proposed a statistical model that is inspired by Dorazio (2014) to allow a joint analysis of presence-only and presence-absence data from multiple species using the maximum likelihood approach for parameter estimation and found a substantial improvement in the predictive performance of their data fusion approach. However, these approaches do not allow the inclusion of a stochastic spatial component in the linear predictors of the target outcomes to represent the spatial variation explained by the unmeasured covariates as suggested in Renner et al. (2015). The consideration of spatial autocorrelation (SAC) in ISDM is relevant for ecological data such as presence-only and abundance outcomes, which are often spatially correlated due to anthropogenic factors or ecological processes such as biotic interactions between species (Wisz et al. 2013), species dynamics (Zurell et al. 2016) or dispersal constraints (Sydenham et al. 2017). The failure to adjust for SAC in model residuals leads to biased estimates of the effects of environmental covariates (Dormann 2007). The SAC can strongly influence the quality and precision of abundance predictions if it co-occurs with an imperfect detection of the species (Guélat and Kéry 2018). In addition, model estimates for coarse-scale covariates, such as climate covariates, may be subject to bias if the species distribution is also influenced by an unmeasured covariate operating on a finer spatial scale (Mäkinen et al. 2022). Consequently, in instances where the species response is spatially autocorrelated and the model does not adequately account for this, the conclusion drawn from the analysis may be erroneous.

However, some recent studies have attempted to build spatially explicit models to analyse spatial data from multiple sources. For example, Adde et al. (2021) fitted ISDMs to presence-absence and presence-only datasets to analyse the distribution of waterfowl in the western boreal forest of Canada by considering two independent spatial processes. They found that the distributions from the ISDM and the model based on PA data alone were more consistent with the observed species distribution than the model based solely on PO data. Dambly et al. (2023) demonstrated how mesh parameterisation can affect spatial predictions from ISDM built from PO and PA data using the Bayesian method based on integrated nested Laplace approximations (INLA) and stochastic partial differential equations (SPDE) techniques. The INLA approach (Rue et al. 2009) is a computationally efficient alternative to the Markov Chain Monte Carlo (MCMC) methods for Bayesian inference. Other recent advances have developed new implementations of ISDM for the combined analysis of ecological data from multiple sources (Dovers et al. 2023; Jung 2023; Foster et al. 2024). Although these tools can handle spatial autocorrelation among observed outcomes, some are restricted to integrating presence-absence and presence-only data. Conversely, alternative approaches use the traditional *gridding* method to approximate the Gaussian Random Field (GRF). The presence-absence data can be useful in a wide range of ecological applications. However, collecting true absence data is often more difficult than recording abundance data. A species can be absent at a given spatial location due to detection issues or competition with other species, especially in wildlife surveys. Indeed, as demonstrated by Mugumaarhahama et al. (2022), the presence of sampling bias and/or low species detection can result in a low level of precision in the baseline estimate of species intensity. These challenges provide an opportunity to rely on the joint analysis of presence-only and count data to understand the spatial distribution of the species of interest, while searching for robust models that can take advantage of both datasets.

Furthermore, in ISDM, accommodating the dependence structure among the latent components of the observation models of integrated outcomes is another crucial aspect that requires a special class of modelling approaches, such as the shared latent component modelling approach. Despite the recent focus on this family of models in the spatial modelling context (Mahaki et al. 2018; Amoah et al. 2020), its uptake among ecologists for the joint analysis of presence-only and abundance data remains limited. ISDM can be regarded as a valuable opportunity for ecologists and conservationists to model the distribution of agroforestry tree species that provide several environmental services to rural and global communities (Mbow et al. 2014; Luedeling et al. 2014). This is because the model enables the integration of presence-only and abundance datasets into a unified modelling framework to map the species' hotspots in a particular area of interest, while facilitating the information sharing across submodels. The resulting insights thus help support informed decision-making processes in that particular area.

This study proposes a Bayesian spatial fusion modelling framework based on INLA-SPDE techniques to integrate presence-only and abundance data, accounting for dependence within and between data sources using a case study of the African baobab, *Adansonia digitata* in Benin (West Africa). The African baobab is a widely distributed multipurpose agroforestry tree species in Africa. The species provides rural inhabitants with various nutritional and medicinal products (Assogbadjo et al. 2005). It has been ranked by Bioversity International as one of the ten most significant agroforestry species for domestication in West Africa and identified as the second most common wild tree providing leafy vegetables in Benin (Dansi et al. 2008). Subsequently, understanding the habitat suitability of the species is a pivotal step in the process of its successful domestication in the region. The present study proposes a Bayesian integrated modelling approach to assess the spatial variation in

the baobab distribution. Specifically, our aim was (i) to assess the spatial dependence among observations of species occurrence and abundance, (ii) to explain the relationships between the species distribution and some potential environmental factors, and (iii) to predict the species distribution from the joint analysis of presence-only and abundance data using the optimal fusion model. Unlike previous ISDMs that often treated the latent components of integrated datasets as independent when utilising the joint likelihood approach, our framework explicitly models the dependence structure among the joint model's latent components, ensuring that spatial dependence within and between outcomes is properly accounted for. This results in enhanced inference, more coherent spatial predictions, and more robust integration of heterogeneous data sources. The INLA approach, when used in conjunction with the SPDE, enables the implementation of fast computational methods to make inferences from the observed data by approximating the continuously indexed GRF as a discretely indexed Gaussian Markov Random Field (GMRF) (Lindgren et al. 2011; Simpson et al. 2016; Rue et al. 2017). The development of an R package, entitled *isdmtools*, was motivated by the necessity to facilitate the evaluation of integrated model performance through the utilisation of block cross-validation techniques and to calculate some standard performance indices well-known in the SDM and machine learning literature. The package is publicly available for installation on a GitHub repository (https://github.com/sodeidelphonse/isdmtools).

The remainder of the paper is organised as follows. In Section 2, we present the two datasets of the model species and the common modelling approaches to analyse each dataset. Then, we propose various spatial fusion models for the joint analysis of both datasets, along with the methodology used for inferring models and conducting

statistical analysis. Then, the main results are presented in Section 3. The potential implications and challenges of the work, as well as the opportunities for further research, are discussed in Section 4. The concluding remarks are presented in Section 5.

2 Methods

2.1 Description of the data

In this study, we modelled two datasets on the African baobab species in Benin. The first set is presence-only (or point pattern) data, which are the geographical coordinates of the species locations, while the second set is abundance (or point referenced) data, which represents the density of the species in terms of the number of individuals per km^2. The abundance data were collected using four transects of 3 kilometres in length and 500 meters in width. Each transect was oriented from the centre of each village to the north, south, west and east. This results in an area of 6 km^2, surveyed around each village centroid. The transects were established in each village, randomly selected along megatransects established in the three climatic regions of the country (Assogbadjo et al. 2005). The use of transects is an effective strategy for achieving a standardised and randomised approach to sampling species density in the region. The presence-only data were recorded from the Global Biodiversity Information Facility (GBIF) database (https://www.gbif.org/) (Figure 1).

The work of Assogbadjo et al. (2005) found that environmental factors such as temperature, rainfall, moisture, and soil characteristics influence the relative abundance and productivity of African baobab in Benin. Thus, for this study, we collected climatic variables from the AFRICLIM website (https://www.york.ac.uk/environment/research/kite/resources/) (Platts et al. 2015). These climatic data have a resolution of 30 arcseconds (approximately 1 km at the equator) and derive from the *WorldClim* baseline station data, which has been

interpolated from the 1950-2000 period. We extracted demographic and topographic covariates from WorldPop (https://hub.worldpop.org/) and soil variables from the ISRIC - World Soil Information (http://www.isric.org/data/AfSoilGrids250m) databases, respectively. The WorldPop and soil variables, which have a spatial resolution of 100 m and 250 m at the equator, respectively, were resampled to the same resolution as the climatic data using the package *terra* in R. Further details of potential covariates included in the study can be found in Appendix A of the Supplementary Information.

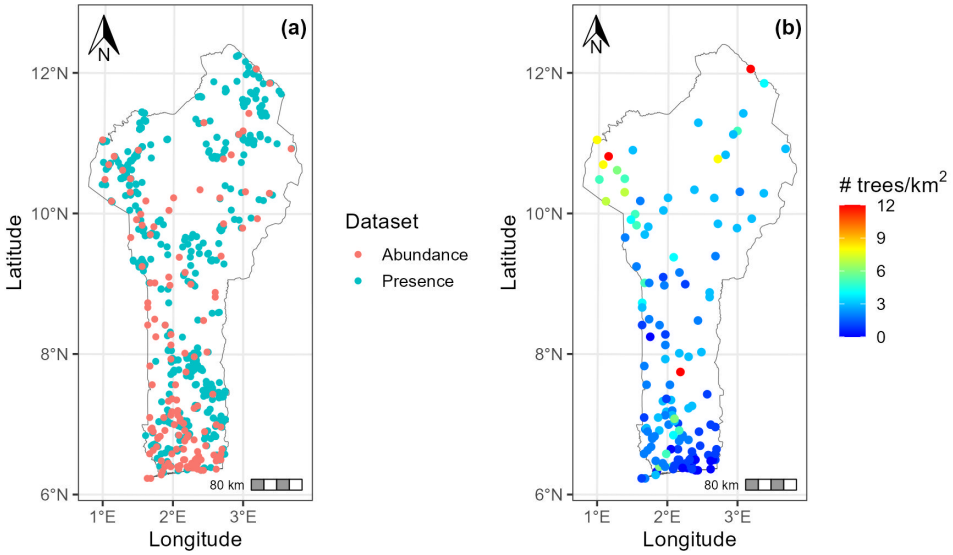

Fig. 1 Map of the presence-only and abundance observations of the African baobab in the study region - (a) the locations of 723 presence-only records (green) and 142 sites for abundance records (light red), and (b) the density of baobab in terms of number of individuals per km^2

2.2 Modeling approach

We assume that the sampling design underlying the observed relative abundance is nonpreferential. That means that the process that generates the locations of observed abundance is independent of the underlying spatial latent process (Diggle et al. 2010), since the sampling protocol used for collecting abundance data is structured through transects established in villages selected randomly along megatransects, as described earlier. Although a systematic sampling scheme would have been better here, we argue that the species shows a distribution that a systematic sampling scheme would most likely miss. Thus, under the assumption of nonpreferential sampling, we present common process-based models to analyse abundance and presence-only data and propose different candidate models for the joint analysis of both outcomes.

2.2.1 Geostatistical model for abundance data

Model-based geostatistics is often utilised to understand and provide quantitative descriptions of natural phenomena distributed in space (or both space and time). The aim is to model a random variable Y as the realisation of a spatial process $S = \{S(x) : x \in \mathcal{D} \subseteq \mathbb{R}^2\}$ observed at a set of discrete locations $\mathcal{X} = \{x_i\}_{i=1}^n$, where x is a vector of the location's coordinates u and v, and predict its values at unobserved locations. Thus, count responses can be analysed using a Generalised Linear Mixed Model (GLMM) (McCulloch and Searle 2001). This is achieved by adding spatially structured random effects $S(x)$ to the linear predictor. These random effects are typically modelled as a zero-mean Gaussian process with a specific correlation function and marginal variance σ_s^2 (Diggle and Ribeiro 2007). In this study, we model the relative abundance of baobab populations as a Poisson random variable of mean μ_i, that is, $Y_i|\mu_i \sim \text{Poisson}(a_i\mu_i)$. The linear predictor of the model is defined as follows:

$$\eta_i = \log(\mu_i) = \beta_0 + \mathbf{Z}'(x_i)\boldsymbol{\beta} + S(x_i), \quad i = 1, \ldots, n, \tag{1}$$

where $\mathbf{Z}_i'$ is the ith row vector of fixed covariates values recorded at site x_i, $\boldsymbol{\beta}$ the vector of covariates' weights, β_0 the intercept, a_i is the area surveyed around each site x_i which is constant throughout the survey domain, and η_i is the expected density of species on the log scale. The linear predictor η_i is related to each observation Y_i through a log link function as follows:

$$\log(\mathbb{E}(Y_i|\mathbf{Z}_i)) = log(a_i) + \eta_i, \tag{2}$$

where Y_i is the response observed at location x_i and $log(a_i)$ is the offset term that allows us to model the counts as a relative abundance. Note that Eq. 1 can also accommodate nonlinear relationships (e.g. piecewise linear functions, B-splines, or one-dimensional SPDE) (Gómez-Rubio 2021), depending on the scientific question and the amount of data available to capture such effects.

The underlying latent process $S(x)$ allows one to capture any spatial dependence or unmeasured covariates associated with the species distribution. We assume that $S(x)$ is a second-order stationary and isotropic Gaussian process with *Matérn* covariance, which is well known for its flexibility and its direct connection to GMRFs through the SPDE approach. That means that the expectation and variance of $S(x)$ are constant in all locations and that its correlation function $\varrho(x_i - x_j) = \text{corr}\{S(x_i), S(x_j)\}$ depends only on the Euclidean distance $h = \|x_i - x_j\|$ between two generic locations x_i and x_j, i.e. $\varrho(x_i - x_j) = \varrho(h)$. This dependence structure is typically chosen to make closer observations more strongly correlated in space than those further apart. The *Matérn* covariance family takes the form:

$$C(h; \kappa, \nu) = \frac{\sigma^2}{2^{\nu-1}\Gamma(\nu)}(\kappa h)^\nu K_\nu(\kappa h), \qquad h > 0, \tag{3}$$

where $K_\nu(.)$ is the modified Bessel function of the second kind, of order $\nu > 0$; $\kappa > 0$

is the scaling parameter related to the practical range, which is the distance at which the correlation between two locations is close to 0. Due to the difficulty in identifying all parameters of the Matérn model, the smoothness parameter ν, which measures the mean-square differentiability of the underlying spatial process, is usually fixed for practical applications. Diggle and Ribeiro (2007) recommended values between 0.5, 1.5, and 2.5. The Matérn model includes special cases of exponential covariance ($\nu = 0.5$) and Whittle covariance ($\nu = 1$) (Whittle 1954). The log-likelihood of the set of observations $\{Y_i\}_{i=1}^n$ is given as:

$$l(\mu|Y) = \sum_{i=1}^{n} (Y_i \log(\mu_i) - a_i\mu_i)\,. \tag{4}$$

2.2.2 Log-Gaussian Cox process model for presence-only data

A point process provides a convenient way to study the locations of an event that arises in a continuously indexed domain. It has applications in various fields, including ecology, forestry, geography, and epidemiology. Let $X = \{x_i : i = 1, \ldots, m\} \subseteq \mathcal{D}$ be the random set of point locations arising from an inhomogeneous Poisson process (IPP). We can model the occurrence of the African baobab as arising from an IPP. The intensity function $\lambda(.)$ defined at each location x can take a log-linear form (Renner and Warton 2013):

$$\lambda(x) = \exp\{\beta_0 + \mathbf{Z}'(x)\boldsymbol{\beta}\}. \tag{5}$$

The intensity $\lambda(x)$ quantifies the expected limit of the number of points per unit of area, i.e., the density of the process. The model allows for adjustment for the effects of fixed covariates defined through the matrix $\mathbf{Z}$, with $\boldsymbol{\beta}$ the vector of associated weights and β_0 the intercept. For any subregion $A \subseteq \mathcal{D}$, the total number of points $x_i \in X$ falling into A, denoted by $N(A)$ follows a Poisson distribution with mean intensity $\Lambda(A)$ defined by:

$$\Lambda(A) = \int_A \lambda(x)\, dx. \tag{6}$$

The expected intensity $\Lambda(A)$ helps connect the point patterns to the count data (Fletcher Jr. et al. 2019). The integral $\Lambda(\mathcal{D})$ across the study region $\mathcal{D}$ is the expectation of population size $N(\mathcal{D})$, that is, $\mathbb{E}(N(\mathcal{D})) = \Lambda(\mathcal{D})$.

In the SDM setting, the intensity $\lambda(x)$ is normalized to obtain a probability distribution $p_\lambda(x) = \frac{\lambda(x)}{\Lambda(\mathcal{D})}$ that integrates to 1. An IPP is then defined as an independent random sample from $p_\lambda(x)$ whose size $N(\mathcal{D})$ itself has a Poisson distribution with mean $\Lambda(\mathcal{D})$ (Fithian et al. 2015). The quantity $p_\lambda(x)$ represents the species distribution or the relative abundance. A transformation such as the inverse complementary log-log (*cloglog*) function helps convert this distribution into a probability of presence (Phillips et al. 2017; Fletcher Jr. et al. 2019). The most common approach to estimate the parameters of this model is to maximise the log-likelihood function, which is expressed by:

$$l(\lambda|X) = \sum_{i=1}^{m} \log \lambda(x_i) - \int_{\mathcal{D}} \lambda(x)\, dx. \tag{7}$$

Note that a term $\log(m!)$ is discarded since it does not depend on the model parameters. This likelihood is intractable due to the second term $\int_{\mathcal{D}} \lambda(x)\, dx$, which has no closed analytical form. However, it can be approximated using numerical integration methods, including weighted quadrature (Renner et al. 2015).

Now, assuming that the intensity function is the realisation of a stationary Gaussian process $S(x)$ with mean zero and marginal variance σ^2 and covariance function $C(h;\kappa)$ as described earlier, Eq. 5 can be rewritten as (Renner et al. 2015):

$$\lambda(x) = \exp\{\beta_0 + \mathbf{Z}'(x)\boldsymbol{\beta} + S(x)\}. \tag{8}$$

Conditional on S and $\mathbf{Z}$, the set of events $X = \{x_i : i = 1, \ldots, m\}$ is the realization of the inhomogeneous Poisson process $\mathcal{S}$ with intensity $\lambda(x)$. This doubly stochastic process is known as *Log-Gaussian Cox Process* (LGCP) (Møller et al. 1998). Simpson et al. (2016) proposed a deterministic integration rule to approximate *stochastic integral* in Eq. 7 and constructed pseudo-observations to obtain a likelihood approximation similar to that of *Poisson* random variables.

However, since a fraction of species is usually observed with the presence-only data due to limited sampling effort, one can model the thinned version of the intensity function as $\lambda^*(x) = p(x)\lambda(x)$, where $p(x)$ is the detection probability depending on some environmental covariates (e.g., distance to main roads or populated cities) (Dorazio 2014; Fithian et al. 2015). In this paper, we eschewed the explicit modelling of a thinned Poisson process in favour of using an additional spatial random component to account for the eventual sampling bias associated with the presence-only data (Mostert and O'Hara 2023) (see Section 2.2.3 for more details).

2.2.3 Spatial fusion models for integrating abundance and presence-only data

We now explore how a model-based data fusion can be specified to account for the dependence structure between both datasets. Let X and Y be the random variables that generate the observed presence-only and abundance outcomes, respectively, $\mathbf{Z}$ an $n \times p$ matrix of potential covariates associated with the observed target responses, $\boldsymbol{\beta}$ a $p \times 1$ vector of covariates' weights, β_0^X and β_0^Y the intercepts of the presence-only and abundance outcomes, respectively. The model has now been expanded to incorporate the $\{f_k(W)\}$ components, which represent a collection of smooth functions

of nonlinear covariates $W_k, k = 1, \ldots, K$. We then propose the following candidate models for the joint analysis of both outcomes.

(i) *Spatial fusion model with common and specific spatial components*

In this model, we assume that the point process observation model has a specific spatial component $S(.)$, and a common spatial component $S_0(.)$. The rationale under this formulation is that the spatial signal $S(.)$ added to the linear predictor of the point process model at the log scale can capture any unknown bias specific to the presence-only data (e.g. uneven sampling effort) as mentioned in Section 2.2.2. The model is defined as follows:

$$\begin{aligned} Y_i|\mathbf{Z}_i, W_{k,i}, S_0(.) &\sim \text{Pois}\left(a_i\mu_i\right), \quad i = 1, \ldots, n, \\ \log(\mu_i) &= \beta_0^Y + \mathbf{Z}'(x_i)\boldsymbol{\beta} + \sum_k f_k(W_{k,i}) + \delta S_0(x_i) \\ X|\mathbf{Z}, W_k, S_0(.), S(.) &\sim \text{IPP}(\lambda(x)), \\ \log(\lambda(x)) &= \beta_0^X + \mathbf{Z}'(x)\boldsymbol{\beta} + \sum_k f_k(W_k(x)) + S_0(x) + S(x), \end{aligned} \tag{9}$$

where $S_0(.)$ and $S(.)$ are independent zero-mean Gaussian processes, and the weight δ, operates as a scaling parameter. The parameter δ can be fixed at 1 if it is assumed that a similar magnitude of spatial signal drives both outcomes. When δ is set to 1, the model becomes:

$$\begin{aligned} Y_i|\mathbf{Z}_i, W_{k,i}, S_0(.) &\sim \text{Pois}\left(a_i\mu_i\right), \quad i = 1, \ldots, n, \\ \log(\mu_i) &= \beta_0^Y + \mathbf{Z}'(x_i)\boldsymbol{\beta} + \sum_k f_k(W_{k,i}) + S_0(x_i) \\ X|\mathbf{Z}, W_k, S_0(.), S(.) &\sim \text{IPP}(\lambda(x)), \\ \log(\lambda(x)) &= \beta_0^X + \mathbf{Z}'(x)\boldsymbol{\beta} + \sum_k f_k(W_k(x)) + S_0(x) + S(x). \end{aligned} \tag{10}$$

Allowing the models in Eqs. 9 and 10 to estimate different regression coefficients and smooth functions of covariates for each dataset, i.e., $\boldsymbol{\beta}_{\mathrm{X}}$ and $\{f_k^{\mathrm{X}}(W_k)\}$ for presence-only response, and $\boldsymbol{\beta}_{\mathrm{Y}}$ and $\{f_k^{\mathrm{Y}}(W_k)\}$ for abundance response, we obtain two other candidate models.

(ii) *Spatial fusion model with a shared spatial component*

Assuming that there is a spatially shared latent field $S_0(.)$ between the two datasets and that the sampling bias in the PO data is negligible, the fusion model can be formulated as follows:

$$\begin{aligned} Y_i|\mathbf{Z}_i, W_{k,i}, S_0(.) &\sim \text{Pois}\left(a_i\mu_i\right), \quad i = 1, \ldots, n, \\ \log(\mu_i) &= \beta_0^Y + \mathbf{Z}'(x_i)\boldsymbol{\beta} + \sum_k f_k(W_{k,i}) + \delta S_0(x_i) \\ X|\mathbf{Z}, W_k, S_0(.) &\sim \text{IPP}(\lambda(x)), \\ \log(\lambda(x)) &= \beta_0^X + \mathbf{Z}'(x)\boldsymbol{\beta} + \sum_k f_k(W_k(x)) + S_0(x), \end{aligned} \tag{11}$$

where $S_0(.)$ is a zero-mean Gaussian process and δ is the weight of the shared spatial signal. The weight δ can be fixed to 1 (i.e. both outcomes are assumed to have a spatial signal with similar magnitude). This results in the following model:

$$\begin{aligned} Y_i|\mathbf{Z}_i, W_{k,i}, S_0(.) &\sim \text{Pois}\left(a_i\mu_i\right), \quad i = 1, \ldots, n, \\ \log(\mu_i) &= \beta_0^Y + \mathbf{Z}'(x_i)\boldsymbol{\beta} + \sum_k f_k(W_{k,i}) + S_0(x_i) \\ X|\mathbf{Z}, W_k, S_0(.) &\sim \text{IPP}(\lambda(x)), \\ \log(\lambda(x)) &= \beta_0^X + \mathbf{Z}'(x)\boldsymbol{\beta} + \sum_k f_k(W_k(x)) + S_0(x). \end{aligned} \tag{12}$$

Allowing the models in Eqs. 11 and 12 to estimate dataset-specific regression coefficients and non-linear smooth functions of covariates as defined above, we obtain two other candidate models.

(iii) *Spatial fusion model with two independent spatial processes*

The underlying rationale for proposing this model is that two distinct spatial processes may drive the two responses. It is assumed that the latent component S_2 captures spatial variation (and uneven sampling intensity if there is any) in PO data, whilst S_1 captures spatial variation in the abundance data. The model can be formulated as follows:

$$\begin{aligned} Y_i|\mathbf{Z}_i, W_{k,i}, S_1(.) &\sim \text{Pois}\left(a_i\mu_i\right), \quad i = 1, \ldots, n, \\ \log(\mu_i) &= \beta_0^Y + \mathbf{Z}'(x_i)\boldsymbol{\beta} + \sum_k f_k(W_{k,i}) + S_1(x_i) \\ X|\mathbf{Z}, W_k, S_2(.) &\sim \text{IPP}(\lambda(x)), \\ \log(\lambda(x)) &= \beta_0^X + \mathbf{Z}'(x)\boldsymbol{\beta} + \sum_k f_k(W_k(x)) + S_2(x), \end{aligned} \tag{13}$$

where $S_1(.)$ and $S_2(.)$ are two independent zero-mean Gaussian processes. Allowing the models in Eq. 13 to estimate dataset-specific regression coefficients and non-linear functions of covariates, we obtain the following candidate model:

$$\begin{aligned} Y_i|\mathbf{Z}_i, W_{k,i}, S_1(.) &\sim \text{Pois}\left(a_i\mu_i\right), \quad i = 1, \ldots, n, \\ \log(\mu_i) &= \beta_0^Y + \mathbf{Z}'(x_i)\boldsymbol{\beta}_{\text{Y}} + \sum_k f_k^{\text{Y}}(W_{k,i}) + S_1(x_i) \\ X|\mathbf{Z}, W_k, S_2(.) &\sim \text{IPP}(\lambda(x)), \\ \log(\lambda(x)) &= \beta_0^X + \mathbf{Z}'(x)\boldsymbol{\beta}_{\text{X}} + \sum_k f_k^{\text{X}}(W_k(x)) + S_2(x). \end{aligned} \tag{14}$$

2.3 Bayesian inference

2.3.1 The stochastic partial differential equations approach

As mentioned earlier, the main focus in spatial modelling is to approximate *entire process(es)* defined in a continuous domain $\mathcal{D}$ to derive some quantitative information

and make predictions based on observations at a set of discrete locations $\{x_i\}_{i=1}^{n}$. The stochastic Partial Differential Equations (SPDE) proposed by Lindgren et al. (2011) and discussed by Simpson et al. (2012) is a computationally efficient approximation of the traditional covariance-based approach to approximate a continuously indexed GRF as a discretely indexed Gaussian Markov Random Field (GMRF) using a finite element method (FEM). The GRF is represented as a GMRF with additional conditional independence properties using weighted sums of deterministic basis function expansions defined over a triangulation of the domain. The rationale for using the GMRF is to avoid factorising dense covariance matrices $n \times n$ with complexity $\mathcal{O}(n^3)$ (Banerjee et al. 2014). The simplest model to represent the latent field $S(x)$ is the stationary *Matérn family* (Eq.3) which is the solution to the following fractional SPDE in $\mathbb{R}^d$ (Whittle 1954, 1963):

$$(\kappa^2 - \Delta)^{\alpha/2}(\tau S(x)) = \mathcal{W}(x), \quad x \in \mathcal{D}, \tag{15}$$

where $\Delta = \sum_{i=1}^{d} \partial^2/\partial x_i^2$ is the Laplacian operator, κ is the scaling parameter, α controls the smoothness of the realizations, τ the variance, and $\mathcal{W}(.)$ a Gaussian spatial white noise. The parameters of the Matérn model (Eq. 3) and the SPDE above (Eq. 15) are coupled in such a way that $\nu = \alpha - \frac{d}{2}$ and the marginal variance of the process is given by (Lindgren and Rue 2015)

$$\sigma^2 = \frac{\Gamma(\nu)}{\Gamma(\nu + \frac{d}{2})(4\pi)^{\frac{d}{2}}\kappa^{2\nu}\tau^2}. \tag{16}$$

For a two-dimensional space problem like in this study, $d = 2$ and then, $\nu = 1$ as α is fixed to 2 (i.e., R-INLA default, but non-integer values $0 \leq \alpha < 2$ are also available). This results in $\sigma^2 = 1/(4\pi\kappa^2\tau^2)$. The practical range is $\rho = \sqrt{8\nu}/\kappa$, the distance at which the spatial autocorrelation is close to 0.10. With this re-parameterization, the Matérn covariance in Eq. 3 is rewritten as:

$$C(h;\rho,\nu) = \frac{\sigma^2}{2^{\nu-1}\Gamma(\nu)}(\frac{\sqrt{8\nu}}{\rho}h)^{\nu}K_{\nu}(\frac{\sqrt{8\nu}}{\rho}h), \qquad h > 0, \tag{17}$$

The basis representation of the GRF discussed in Lindgren et al. (2011) and implemented in *R-INLA* is given by:

$$S(x) = \sum_{g=1}^{G} w_g \phi_g(x), \tag{18}$$

where $\{\phi_g(.)\}$ are piecewise linear basis functions defined by triangulation of the domain, G is the number of vertices of the triangulation and $\{w_g\}$ are the associated weights, with zero-mean Gaussian multivariate distribution, i.e. $\{w_g\} \sim \mathcal{N}(0, \mathbf{Q}_S^{-1})$, where $\mathbf{Q}_S$ is a precision matrix. This specification allows the distribution of $S(x)$ to approximate the distribution of stochastic weak solutions to the SPDE. For count data, the GRMF is constructed by projecting the SPDE onto the basis representation at the mesh vertices, where $\phi_g(.)$ equals 1 at the gth vertex of the triangle (or takes values based on barycentric weights from the three neighbouring vertices of the triangle containing the location) and 0 at other vertices. This procedure results in *projector matrix* $\boldsymbol{A}_Y$ of dimension $n \times G$. For PO data, the projector matrix $\boldsymbol{A}_X$ is composed of an identity matrix with dimension $G \times G$, row-binded to an $m \times G$ matrix obtained similarly to $\boldsymbol{A}_Y$ (see Krainski et al. (2018) for details). This approximation allows the precision matrix to be sparse with elements that have explicit expressions with hyperparameters ρ and τ (Lindgren and Rue 2015), and the computational complexity is reduced to $\mathcal{O}(n^{3/2})$. The mesh defined on the Benin map with the *Delaunay triangulation* to represent the GRF has 916 vertices and can be found in Appendix A (Figure A.1).

2.3.2 Prior and posterior distributions of parameters and inference with INLA

The integrated nested Laplace approximations (INLA) approach is designed for latent Gaussian models (LGMs), a vast and flexible class of models ranging from generalised linear (or additive) mixed to spatiotemporal models. The general hierarchical form of LGMs is presented in Rue et al. (2009, 2017). The fusion models proposed in the present study are all part of latent Gaussian models. Consequently, R-INLA, or any tool built on it, is a viable option for their implementation. In the joint modelling framework, each observation model has its own specific linear predictor that can share some or all of its components. The complete likelihood for the two datasets $\boldsymbol{Y} = \{Y, X\}$ is then expressed as the product of the likelihood of each dataset as follows:

$$\pi(\boldsymbol{Y}|\boldsymbol{u}, \boldsymbol{\theta}) = p(Y|\boldsymbol{\eta}(\boldsymbol{u}_Y), \boldsymbol{\theta}_Y)\ p(X|\boldsymbol{\eta}(\boldsymbol{u}_X), \boldsymbol{\theta}_X). \tag{19}$$

where $\boldsymbol{u}$ is the vector of the Gaussian latent field (i.e. $\boldsymbol{\beta}$, $f_k(W_k(x))$ and $S(x)$) and $\boldsymbol{\theta}$ is the vector of hyperparameters (i.e. δ, ρ and σ_s). The prior distribution of the latent field $\pi(\boldsymbol{u}|\boldsymbol{\theta})$ and the likelihood function $\pi(\boldsymbol{Y}|\boldsymbol{u}, \boldsymbol{\theta})$ are governed by hyperparameters $\boldsymbol{\theta}$ with prior density $\pi(\boldsymbol{\theta})$. In practice, we make an inference about the joint posterior distribution of the unknown parameters , which is defined as follows:

$$\pi(\boldsymbol{u}, \boldsymbol{\theta}|\boldsymbol{Y}) \propto \pi(\boldsymbol{Y}|\boldsymbol{u}, \boldsymbol{\theta})\pi(\boldsymbol{u}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}), \tag{20}$$

We refer the reader to Rue et al. (2009, 2017) for more details on the INLA procedure to estimate the parameters in Eq. 20.

Prior assumptions were formulated on the unknown parameters to complete the proposed models. For the hyperparameters of the Matérn covariance function (i.e., ρ

and σ), we assign the penalised complexity (PC) prior introduced by Simpson et al. (2017) to reduce model overfitting. The PC prior for two-dimensional space is defined as follows:

$$\pi(\rho, \sigma) = \lambda_\rho \lambda_\sigma \rho^{-2} \exp\left(-\lambda_\rho \rho^{-1} - \lambda_\sigma \rho\right), \quad \sigma \geq 0, \rho \geq 0, \tag{21}$$

where, $p(\rho \leq \rho_0) = \alpha_1$ and $p(\sigma \geq \sigma_0) = \alpha_2$ are obtained by $\lambda_\rho = -log(\alpha_1)\rho_0$, and $\lambda_\sigma = -log(\alpha_2)/\sigma_0$. For all spatial components, we set $\rho_0 = 10$, $\sigma_0 = 0.1$, and $\alpha_1 = \alpha_2 = 0.01$. Default priors were defined as Gaussian distributions with mean 0 and precision 0.001 for regression coefficients, and mean 1 and precision 0.1 for the shared latent weight.

2.4 Data analysis

First, we performed an exploratory data analysis on the abundance data by evaluating the spatial dependence between observations using an experimental variogram. Under the assumption of intrinsic stationarity, the variogram is expressed as $2\gamma(h) = Var\{Y(x+h) - Y(x)\}$, where $\gamma(h)$ is the semivariogram whose empirical estimate is given as follows:

$$\hat{\gamma}(h) = \frac{1}{2|N(h)|} \sum_{(i,j) \in N(h)} \left(Y(x_i) - Y(x_j)\right)^2, \tag{22}$$

where $|N(h)|$ is the number of distinct pairs in $N(h) = \{(x_i, x_j) : \|x_i - x_j\| = h,\, i \neq j\}$. A Box-Cox transformation was applied to the response variable Y before estimating $\hat{\gamma}(h)$, which was compared to its Monte Carlo envelope simulated from 99 random permutations (Diggle and Ribeiro 2007). Furthermore, we evaluated spatial dependence within the baobab point pattern using the inhomogeneous *K-function* and its Monte Carlo envelope simulated from 99 random permutations of the data. The K-function, which is defined as the expected number of events within a radius r of

a typical point, normalised by the intensity, has an explicit link with the LGCP covariance function defined in Eq. 17. Using the second-order properties of the LGCP (Diggle 2013), the K-function of an LGCP model is given as follows:

$$K(r) = 2\pi \int_0^r \mathrm{g}(s) s ds, \tag{23}$$

where, $\mathrm{g}(r) := \exp(C(r))$ is the pairwise correlation function and $C(r)$ is the covariance function of the form $C(r) = \sigma^2 c(r/\phi)$. The parameters σ^2 and ϕ control the strength and scale of the SAC, respectively. When $C(r) \to 0$ quickly, $K(r) \to \pi r^2$, corresponding to a homogeneous Poisson process. In this study, a comparison was made between $\hat{\mathrm{g}}(r)$ estimated empirically from the point pattern and $\exp(\hat{C}(r))$ estimated using both the classic and Bayesian LGCP model.

With regard to the selection of covariates, we conducted a variance inflation factor (VIF) analysis on the initial covariates to eliminate the most correlated ones employing a VIF threshold of 5. We then used our expert knowledge of the species ecology and the statistical significance of covariates to guide the selection of the final covariates for the Bayesian models. Potential covariates considered in the final model include annual temperature (Bio1), rainfall of the driest month (Bio14), slope, silt and clay fractions (soil texture). The effect of the rainfall of the driest month was estimated using a one-dimensional SPDE with PC-priors, say, $p(\rho \leq 1) = 0.01$ and $p(\sigma \geq 1) = 0.01$ (see the mesh depicted in Figure A.2 in the Appendix A). This was due to its nonlinear relationship with the species intensity. All environmental covariates were standardised to have a mean of 0 and a standard deviation of 1. The map of the selected covariates is presented in Figure A.3 in the Appendix A.

Furthermore, we implemented the generalised linear mixed-effects model (GLMM) and the log-Gaussian Cox process (LGCP) model for analysing abundance and PO

data sets, respectively, within a Bayesian framework using Integrated Nested Laplace Approximations (INLA) (Rue et al. 2009) combined with Stochastic Partial Differential Equations (SPDE) techniques (Lindgren et al. 2011; Simpson et al. 2016). The implementation of different fusion models with Bayesian framework was facilitated by the use of *PointedSDMs* (Mostert and O'Hara 2023), an R package which is built on *inlabru* (Bachl et al. 2019) to accommodate diverse configurations of spatial latent components. The selection of the best fusion model was achieved by implementing the block cross-validation technique. The procedure uses the K-means algorithm to obtain spatially separated test data, thus ensuring that models are robust to spatial autocorrelation (Figure 2). Performance metrics used in this study include various ROC-based indices, such as the area under the receiver operating characteristic (ROC) curve (AUC), the true skill statistic (TSS), the accuracy, and the F1 score. The *Youden* optimality criterion was used for ROC-based scores. This criterion involves selecting a threshold that maximises both sensitivity and specificity. In the event of multiple thresholds, the one which maximises the F1 score (i.e. both precision and recall) was selected. The calculation of each metric was derived as a *weighted composite score*, taking into account the respective sample size of the presence-only and abundance responses. Let M_j be an evaluation metric for a dataset j $(j = 1, \cdots, J)$ and w_j the corresponding sample size in the model testing set. The weighted composite score is defined as follows:

$$M_{\text{composite}} = \frac{\sum_{j=1}^{J} w_j M_j}{\sum_{j=1}^{J} w_j}. \tag{24}$$

The calculation of error-based metrics, including the root mean square error (RMSE) and the mean absolute error (MAE), was performed exclusively for the standalone abundance model. The deviance information criterion (DIC) was also calculated to assess the quality of the integrated models. A comparison of the evaluation scores of the models was conducted using either the analysis of variance if the metric residuals were normal with constant variance or the Kruskal-Wallis test if the normality

and homoscedasticity assumptions were not met. The Pearson correlation coefficient was utilised to evaluate the relationship between the predictions of the selected fusion model and those of the standalone model, which were derived from PO and abundance data. We estimated the spatial latent field, species intensity and habitat suitability from the optimal model as the posterior mean, the posterior standard deviation, and the 95% Bayesian credible interval using 1000 Monte Carlo samples. Data partitioning for cross-validation, suitability analysis, calculation of performance indices, and mapping were performed using the R package *isdmtools*, which we developed and made accessible on the GitHub repository mentioned in Section 1. All data analyses were performed in R version 4.4.1 (R Core Team 2024).

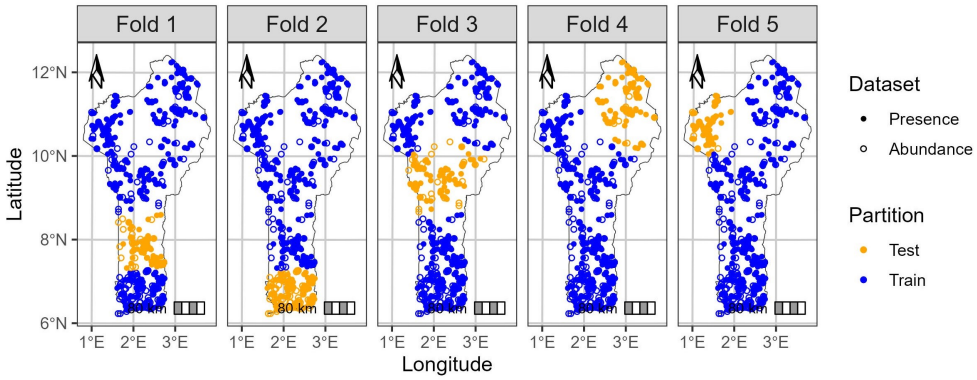

Fig. 2 Visualisation of data partitions for the block cross-validation procedure. The representation of PO data is illustrated by closed circles, whilst the representation of abundance data is illustrated by open circles. For each fold, training and testing sets are each comprised of a list of datasets which have been used to train and validate the integrated model, respectively

3 Results

3.1 Spatial heterogeneity in the distribution of baobab populations

The exploratory data analysis revealed that the baobab distribution pattern in the study region is not homogeneous. A low density was observed in the southern part of Benin, while a relatively high density was observed in the northern part of the country (Figure 1b). In the south (i.e. the Guinean zone), the majority of observed sites exhibited a lower density, ranging from 0 to 3 per km^2. This is in contrast to the north (i.e. the Sudanian zone), where some sites demonstrate higher densities, exceeding 5 individuals per km^2. Only two sites in the Guinean zone exhibit a maximum density of 6 individuals per km^2. In the Soudano-Guinean zone, few sites contain more than 5 individuals per km^2 compared to the Soudanian zone. The maximum observed density in the study region was recorded in the Sudanian (Dassari and Karimama) and Sudano-Guinean (Dassa-zoumé) zones with an average of approximately 12 individuals per km^2. Furthermore, the baobab tree occurs in all three climatic regions of the country, but with a relatively clustered pattern in certain spatial locations (Figure 1a).

The empirical variogram demonstrates evidence of spatial autocorrelation (SAC) between the observations of baobab density, as it lies partially outside the Monte Carlo envelope (Figure 3a). It depicts a typical shape of a stationary, correlated spatial process. The spatial range estimated from the Matérn model fitted to the empirical variogram of the abundance response is $\rho = 33.54\,\text{km}$, with a practical range of $\rho_p = 38.12\,\text{km}$ (Table 1). Moreover, the simulated point pattern obtained using the empirical K-function, $K(r)$, shows a clustering of species occurrence at distances below $r \approx 45\,\text{km}$ and repulsion beyond this distance (Figure 3b). The practical range

estimated from the conventional LGCP model is $\rho_p = 29.84\,\text{km}$, while the variance of the latent field is $\sigma^2 = 3.82$; a value that is relatively higher than its Bayesian counterpart (Table 1). Visualising the two data sets as a single map provides a more comprehensive and accurate representation of the baobab distribution in the study region (Figure 1a).

Furthermore, we showed the explicit link between the covariance function of the log-Gaussian Cox process and the empirical K-function (Figure 4). The model-based pairwise correlation function (i.e., the exponential of the covariance function of the conventional (i.e. frequentist) LGCP fitted to the point pattern data) tends to overlap the empirical pairwise correlation function $g(r)$ derived from the inhomogeneous K-function $K(r)$ (Figure 4). This finding reveals that the explicit relationship between the LGCP covariance function and $K(r)$ is valid for the observed point pattern. However, the present study demonstrates that the pairwise correlation function estimated from the conventional LGCP rapidly approaches zero compared to its analogue function estimated from the Bayesian model (Figure 4). Moreover, it has been observed that the distance from the origin to the crossing point of $K(r)$ with its theoretical value πr^2 is marginally greater than the practical ranges estimated from both LGCP models (Table 1). To assess spatial variation in the spatial distribution of baobab, it is necessary to construct a spatial model that incorporates latent component(s) that drive the realisation of both responses.

3.2 Selection of the optimal fusion model

ROC-based metrics calculated for each candidate model show that the shared component model with common covariate effects (Eq. 12) is the best fusion model for the joint analysis of the relative abundance and presence of baobab populations in the study region. This model shows the highest average composite scores of 0.85 ± 0.02

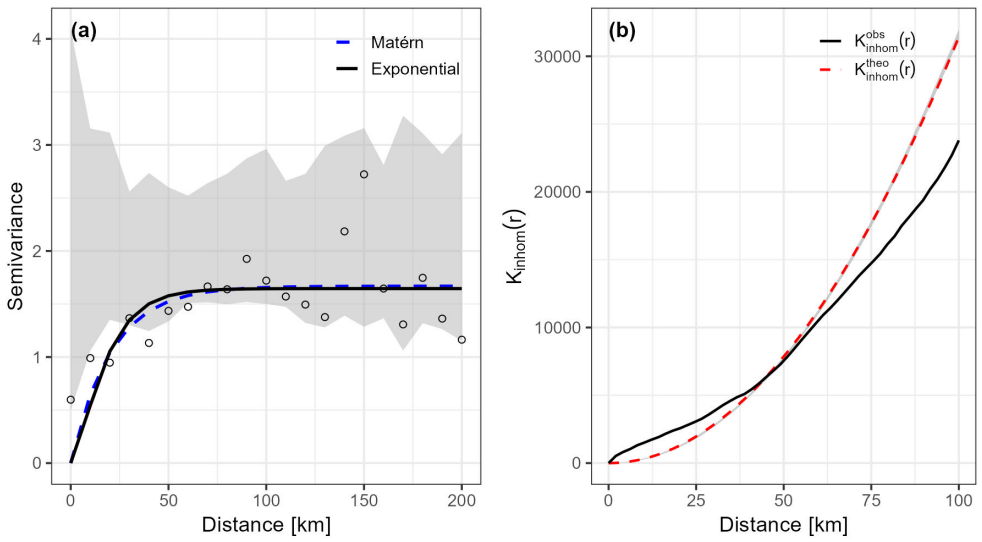

Fig. 3 (a) The empirical variogram with the Monte Carlo envelope (grey band) and (b) the inhomogeneous K-function with the simulated envelope (shaded band) showing the spatial dependence among observations of abundance and presence of baobab, respectively. As illustrated in panel (a), the empirical variogram (open circles) lies partially outside the Monte Carlo envelope for distances below $h \approx 38$ km, indicating evidence of SAC in the data. The black and blue lines represent the fitted variograms using Matérn and exponential covariance functions, respectively, as implemented in the *geoR* library. In panel (b), the empirical K-function (black line) is situated above the simulated envelope (grey band around the theoretical K-function, i.e πr^2) for distances below $r \approx 45$ km, indicating evidence of spatial clustering among observed locations

for area under the ROC curve (AUC), 0.77 ± 0.02 for accuracy and 0.66 ± 0.05 for true skill statistics (TSS) (Table 2). The selection of the shared component model with common covariates' effects is further supported by the DIC, which achieves the lowest cross-validation score of all fusion models. However, when the DIC criterion is used, no significant differences in cross-validation scores are evident between models. Moreover, the fusion model with independent spatial latent components was ranked second among the fusion models that were evaluated, indicating that the hypothesis concerning two independent spatial latent signals driving the observed responses may not be entirely erroneous. In contrast, the fusion model that accounts for additional unknown spatial variation or sampling effort in presence-only data demonstrates a relatively lower performance, compared to fusion models without bias component

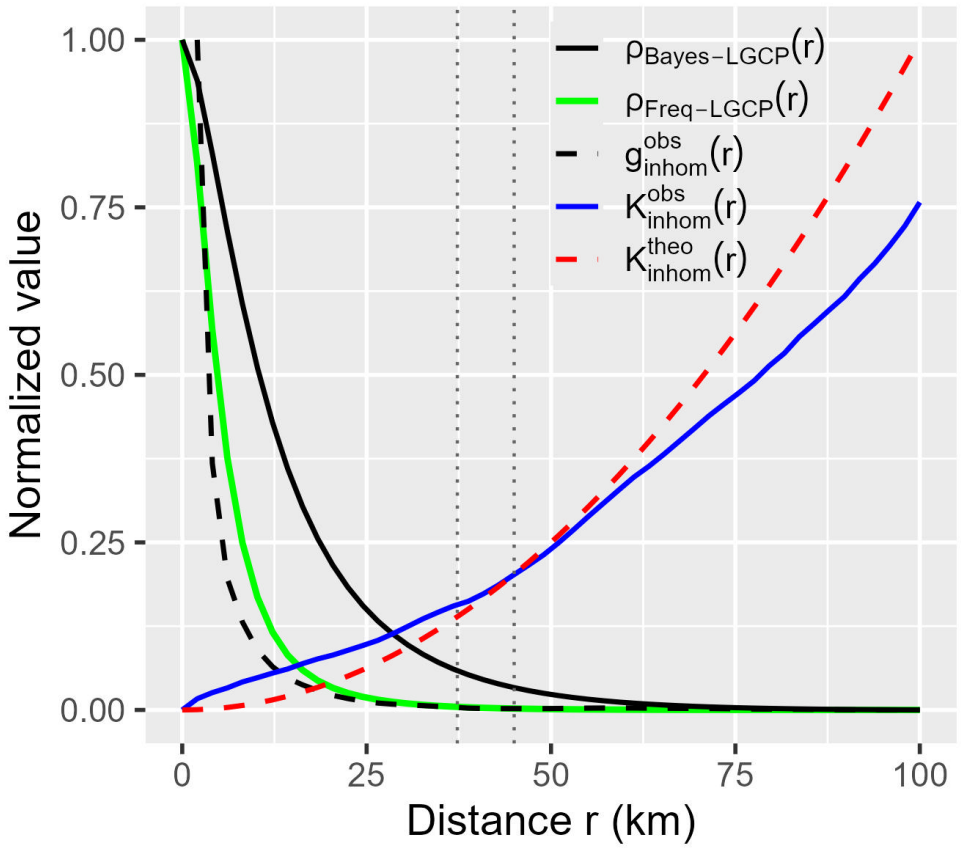

Fig. 4 Explicit link between the covariance function of the log-Gaussian Cox process model and the empirical K-function. Each function value is normalised using *min-max* normalisation, thereby facilitating the comparison of its shape across functions. On the legend, $\rho_{\text{Freq}-\text{LGCP}}(r)$ and $\rho_{\text{Bayes}-\text{LGCP}}(r)$ represent the pairwise correlation functions derived from the LGCP model (without covariates) fitted to the observed point pattern with the *Matérn* covariance using the minimum contrast algorithm in *spatstat* and the INLA-SPDE techniques in *inlabru*, respectively; $\text{g}^{\text{obs}}_{\text{inhom}}(r)$ is the empirical pairwise correlation estimated from the inhomogeneous K-function $\text{K}^{\text{obs}}_{\text{inhom}}(r)$ and $\text{K}^{\text{theo}}_{\text{inhom}}(r)$ the theoretical K-function

and/or specific covariates' effects (Table 2). Among the ten fusion model configurations, the fusion model incorporating a scaling factor, a bias in the presence-only observation model, and the effects of data-specific covariates demonstrates the poorest performance in terms of AUC (0.615 ± 0.026), accuracy (0.626 ± 0.059), and TSS (0.305 ± 0.047) scores. This finding suggests that sampling bias or the effects of data-specific covariates do not need to be considered in the proposed fusion modelling

Table 1 Variance and range parameters calculated from the analysis of spatial dependence among the observed sites for the abundance data and observed locations for the point pattern data

Method	Decay parameter	σ^2 (variance)	κ (kappa)	ρ (range, SPDE-style)	Practical range (10%)
Classic LGCP for PO data (*kppm*, $\nu = 1$)	$\alpha = 13.106$ km (scale)	$\sigma^2 = 3.8205$	0.1079	26.212 km	29.79 km
Bayesian LGCP for PO data (*INLA-SPDE*, $\nu = 1$)	$\rho = 37.30$ km (95% BCI: 31.42-44.06)	$\sigma = 1.31$ (95% BCI: 1.18-1.46); $\sigma^2 = 1.7161$	0.0877	37.30 km	37.30 km (95% BCI: 31.42-44.06)
Empirical K-function for PO data (graphically)	-	-	-	-	Crossing point: $\mathrm{K}(r) \cap \pi \mathrm{r}^2 \approx 45$ km
Variogram model for density (*variofit*, $\nu = 1$)	$\phi = 11.86$ km (scale)	$\sigma^2 = 1.6455$	0.119	33.54 km	38.12 km

Note: The estimated parameters were harmonised in accordance with the INLA-SPDE parametrisation, for the purpose of facilitating comparison across software and deriving a unified covariance function for visualisation. Typically, the scale parameter is defined as $\kappa = \sqrt{2\nu}/\alpha$ in *spatstat*, $1/\phi$ in *geoR* and $\sqrt{8\nu}/\rho$ in R-INLA. From the three expressions, we derived the range parameter in the classic setting as $\rho = 2\alpha$ for *spatstat* and $\rho = \phi\sqrt{8\nu}$ for *geoR*. The 10% practical range is the distance where $C(r) = 0.1 \cdot \sigma^2$

framework for analysing the spatial datasets under consideration. When fitted to abundance data alone, the Bayesian Poisson model demonstrates satisfactory mean absolute error (MAE) and root mean square error (RMSE) values of 1.63 ± 0.35 and 2.26 ± 0.35, respectively. These results suggest that the Bayesian GLMM is capable of accurately predicting species abundance at the observed sites.

3.3 Relationships between environmental factors and the distribution of baobab populations

The Bayesian fusion model, based on a shared latent component fitted to the two datasets, demonstrates a relatively strong relationship between potential environmental factors and the distribution of the baobab population in Benin. Annual temperature, rainfall in the driest month, slope, and soil texture (silt fraction) have been shown to significantly affect the abundance and intensity of the species, as evidenced by the non-zero credibility interval of the associated parameters (Table 3). These environmental factors have been found to have varying effects on the distribution of baobab populations. The study's findings demonstrate that the annual temperature and slope have a positive multiplicative influence on the intensity of the baobab populations. In contrast, soil textures (silt and clay fractions) have a multiplicative decreasing effect on species intensity (Table 3). In addition, we found that the integrated model can detect the effects of some covariates that were not significant in models based on a single dataset. For example, the separate model based on abundance data did not detect the significant effect of soil texture (silt fraction) and slope (Table 4). We also observed that the intensity of the baobab population has an increasing linear relationship with the slope covariate: $\beta_{\text{slope}} = 0.135$ (95% BCI = 0.073-0.197) (Table 4 and Figure 5c). In contrast, the rainfall in the driest month depicts a non-linear relationship with the predicted species intensity (Figure 5b). The remaining predictors depict an almost linear relationship with the estimated intensity (Figure 5). The predicted habitat suitability and relative abundance for presence-only and abundance data, respectively, can be found in Appendix B (Figures B.1 and B.2).

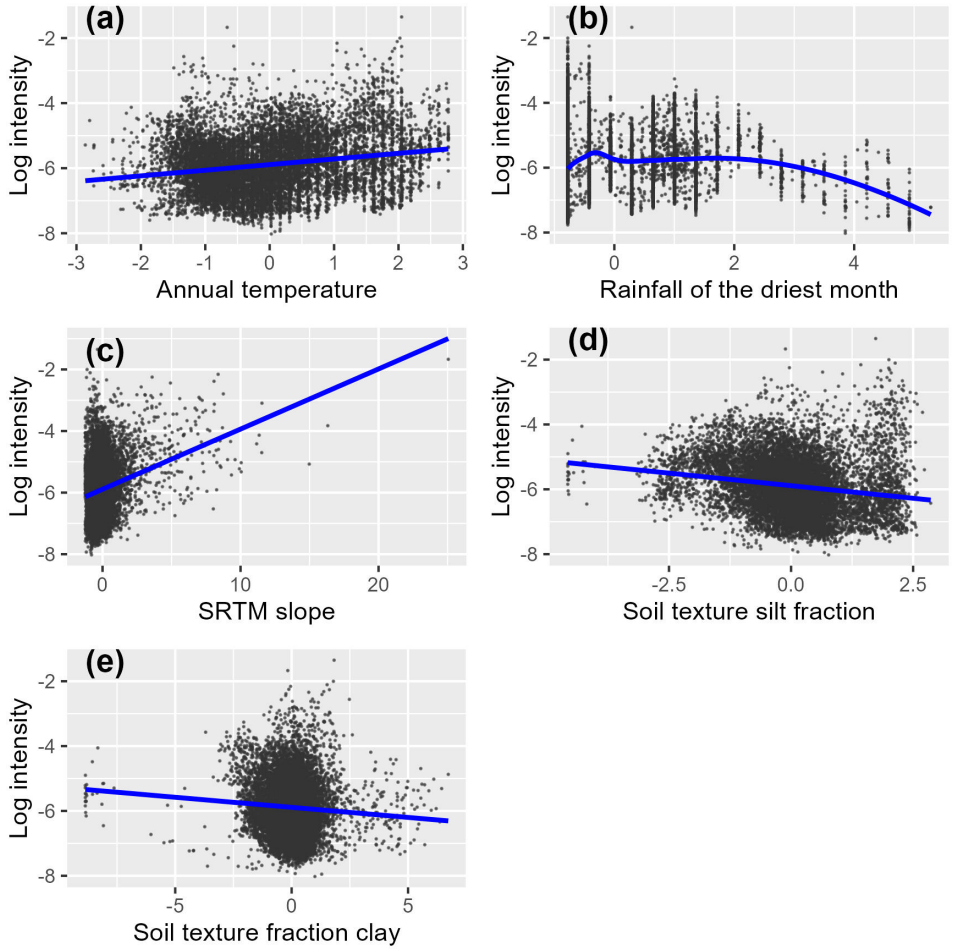

Fig. 5 Relationship between log intensity estimated from the shared component model and environmental covariates - The blue line or curve on each panel is the function that has been fitted graphically to the corresponding standardised covariate to assess the overall shape of the relationship

3.4 Prediction of the baobab spatial distribution from the integrated model

The results of the joint modelling of the abundance and occurrence data of baobab showed some heterogeneity in the spatial pattern of baobab in Benin. We found that the distribution map predicted from the joint analysis of both outcomes provides a smooth representation of species intensity in the study region. We observe

some patches in the species distribution across the three climatic zones of Benin, with a relatively higher intensity in the northern part of the country than in other regions. Some hotspots of the species can be observed in the Guinean and Soudano-Guinean zones (Figures 6 and 7). However, we noticed that the integrated model with the shared spatial component and common covariates' effects slightly reduces the marginal variance of the latent field and increases the magnitude of the spatial range compared to the standalone models. Indeed, the spatial range estimated from the shared component model revealed a mean nominal range of 34.47 km (95% Bayesian credible interval, BCI = 27.60 km - 42.61 km) (Table 3), which is slightly higher than the values of 21.4 km (95% BCI = 12.78 km - 33.34 km) and 32.24 km (95% BCI = 26.84 km - 38.4 km) estimated from separate models for abundance and presence-only outcomes, respectively. The mean estimate of the practical range from the Bayesian LGCP aligns with the corresponding value ($\rho_p = 29.8$ km) derived from the conventional LGCP, which falls within the 95% Bayesian credible interval. However, the Bayesian model is more precise (i.e. has a smaller variance) than the classic model (Tables 4 and 1).

Moreover, the marginal standard deviation of the spatial latent component predicted by the selected fusion model spans the range of 0.2 and 0.8, indicating that the spatial signal does not vary significantly around the mean. However, its high variability is observed in areas where few locations have been recorded for the species (Figure 8). Furthermore, the spatial range for the smooth function of the rainfall of the driest month, i.e. f(Bio14), is 0.96 (95% BCI = 0.31 - 2.29). This index corresponds to a posterior mean of 0.44 (95% BCI = [-0.28, 1.17]) for f(Bio14) on the standardized scale (see Figure B.3 in Appendix B). The marginal distributions for all parameters and hyperparameters of the shared latent component model with common covariates' effects fitted to baobab datasets are presented in Figure B.4 in the Appendix B.

Furthermore, the relative abundance predicted from the fusion model, using the count model components, revealed a near-perfect agreement with the relative abundance estimated from the count data alone at the observed sites ($r(133) = 0.93, P < 0.001$) (Figure 9b). In contrast, the habitat suitability estimated from the fusion model demonstrates a relatively strong correlation with the suitability index derived from the presence-only data ($r(692) = 0.66, P < 0.001$) (Figure 9a). The two scatter plots tend to follow the 1:1 relationship at the observed locations for the point pattern and sites for the abundance response. However, the intensity values predicted from the PO standalone model are relatively scattered above the perfect agreement line for medium to high values. Conversely, the integrated model tends to predict low to intermediate intensity values, thereby mitigating the overfitting inherent to PO data.

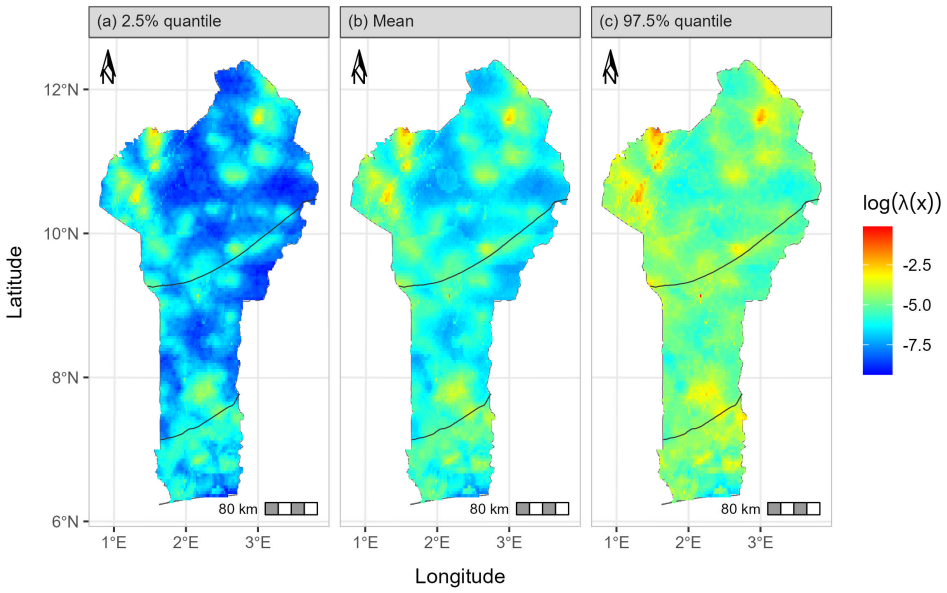

Fig. 6 Map of the predicted log intensity from the shared spatial component model with common covariates' effects fitted to PO and abundance data - (a) posterior mean, (b) 2.5% posterior quantile, and (c) 97.5% posterior quantile

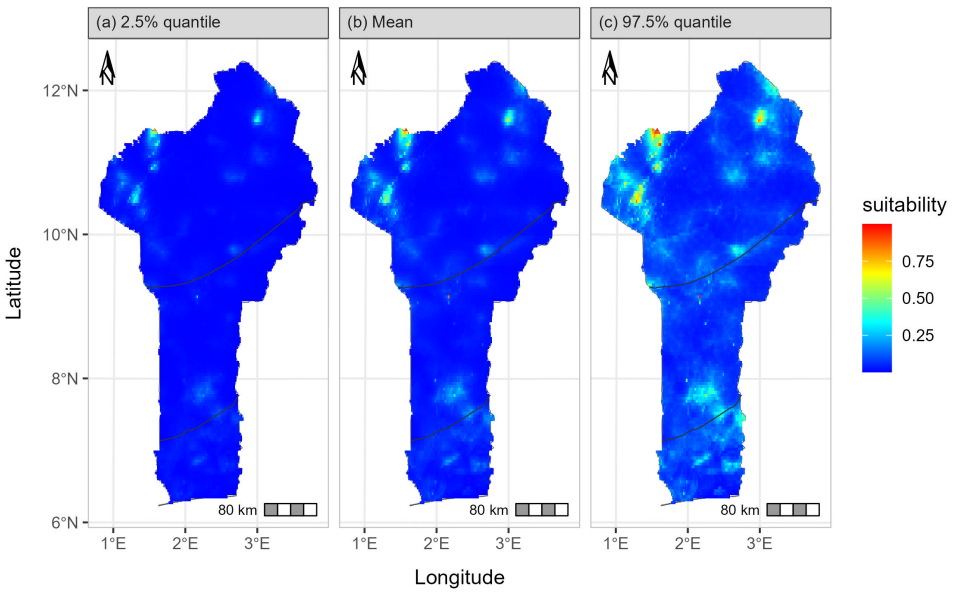

Fig. 7 Map of the predicted habitat suitability from the shared spatial component model with common covariates' effects fitted to PO and abundance data - (a) posterior mean, (b) 2.5% posterior quantile, and (c) 97.5% posterior quantile

4 Discussion

4.1 Spatial heterogeneity in the baobab distribution and its relationship with environmental factors

This study assessed spatial variation in African baobab populations using a Bayesian model-based data fusion framework based on integrated nested Laplace approximations and stochastic partial differential equation techniques. The INLA-SPDE approach allows us to account for spatial autocorrelation driven by spatially structured covariates that were not included in the models (Bachl et al. 2019). The significance of spatial dependence highlighted in our findings indicates the manner in which ecological datasets can be spatially structured due to ecological processes such as interspecific interactions (e.g. competition and symbiosis) (Wisz et al. 2013), population dynamics (Zurell et al. 2016), and dispersal barriers (Sydenham et al. 2017).

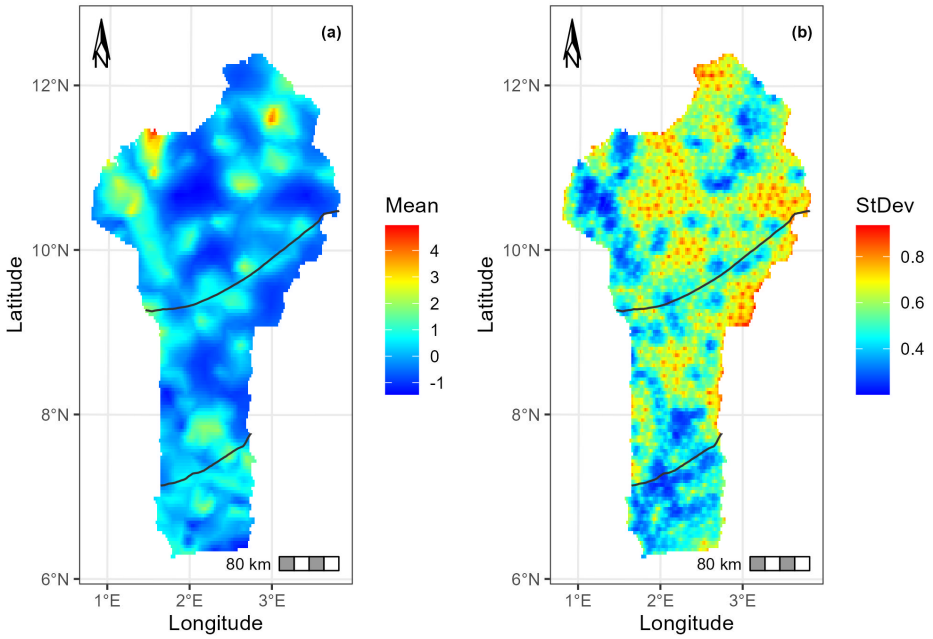

Fig. 8 Map of the predicted Gaussian field from the shared latent component model with common covariates' effects fitted to PO and abundance data - (a) posterior mean, and (b) posterior standard deviation

In addition, the significant effects of climatic and edaphic factors on the species intensity confirm how climate change influences the spatial pattern of many plant species in the region. Thus, the relationship between the distribution of the baobab population and environmental variables such as annual temperature, rainfall in the driest month, slope, and soil characteristics indicates that these factors shape the distribution of species in Benin. This finding is consistent with that of Assogbadjo et al. (2005), who found a decrease in baobab abundance due to certain edaphic factors, including climatic factors and soil properties. As posited by these authors, the baobab population has been observed to occur on sandy soils in the Sudanian and Guinean zones as well as on sandy–clayey soils in the Sudano-Guinean zone. A recent study evaluating the distribution of baobab using presence-only data and the maximum entropy (Maxent) approach has identified soil characteristics, rainfall

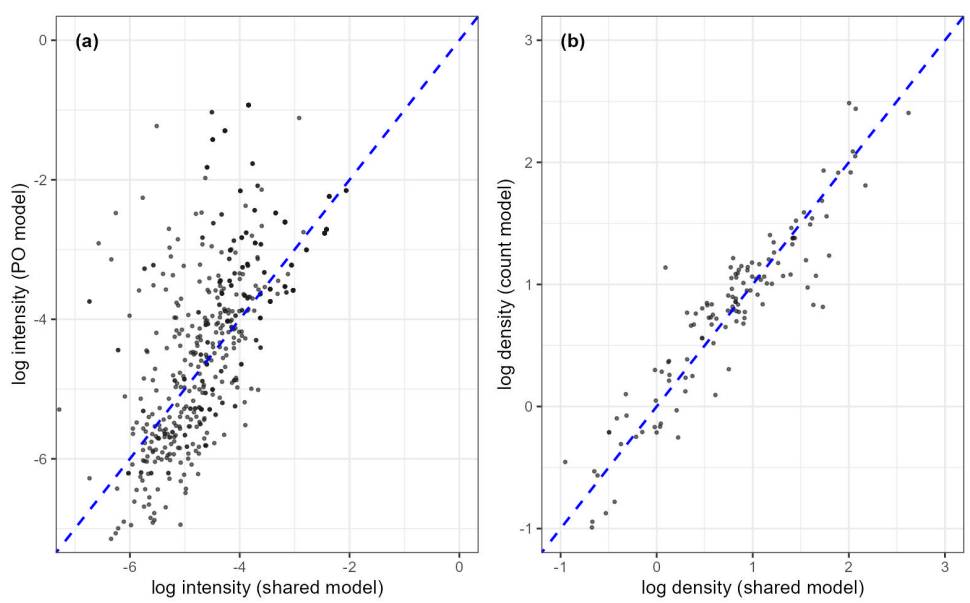

Fig. 9 Visualisation of the relationship between the predictions from the shared spatial component model and those from the standalone model based on (a) presence-only data at the observed locations and (b) abundance data at the observed sites. The dashed blue line represents the benchmark for the expected level of perfect congruence among predictions

in the driest month, and annual rainfall as the three main environmental variables contributing to the model (Assogba et al. 2022). Although some key environmental variables similar to those found here have been identified, the resulting interpretation differs. The exponentiation of the covariates' effects estimated from the ISDM has a direct interpretation as the multiplicative factor of the expected species density for a unit increase in the covariates, rather than the contribution of the covariates to the model. In addition, the approach used by these authors did not incorporate spatially explicit latent components into the model, unlike in the present study. This limitation would also have a substantial impact on the identification of the primary environmental drivers in their model and the subsequent predictive insights. Moreover, the increasing relationship between the intensity of the baobab trees and the slope may be attributable to the fact that the species finds its suitable climatic

conditions when moving from the southern to the northern part of the country.

Despite the harmonisation of estimated covariance parameters, minor discrepancies observed in the practical range (and/or variance) of the latent Gaussian field when using different estimation approaches may be due to the internal settings inherent to each approach influencing the resulting estimates. Therefore, the longest practical range observed graphically from the inhomogeneous K-function can be explained by the fact that it integrates over all distances, resulting in a longer estimate of the practical range. Additionally, it is worth noting that the frequentist LGCP model uses the minimum contrast algorithm to adjust the second-order summary statistic in order to match the empirical pairwise correlation function. This justifies the overlap of the empirical pairwise correlation function $\hat{\mathrm{g}}(r)$ with its theoretical LGCP counterpart, $\exp(\hat{C}(r))$. This finding confirms the explicit link between the empirical K-function and the conventional LGCP covariance function. In contrast, the Bayesian LGCP model employs INLA-SPDE techniques and makes some prior assumptions about the covariance structure of the Gaussian field by penalising model complexity. However, there is no guarantee that the resulting pairwise correlation function will match its empirical counterpart derived from the observed data. This disparate modelling philosophy may explain the discrepancy observed between the pairwise correlation functions estimated from the two modelling approaches. Nonetheless, the fact that the practical range of the frequentist LGCP model lies within the Bayesian credible interval, coupled with the lower marginal variance from the Bayesian LGCP model, proves that the Bayesian estimation of parameters for the latent field covariance structure is valid overall.

4.2 Modelling approach and relevance

The findings of the present study demonstrate that the shared latent component model without a scaling factor fits the observed datasets better than the shared latent component model that incorporates a scaling factor. This result may be attributable to the fact that the relative abundance is modelled using an offset term, rather than the original counts. This makes it easier to handle the underlying shared spatial signal at almost the same scale. In addition, the results indicate a slight increase in the magnitude of spatial range estimates and a slight decrease in the marginal standard deviation of the latent effects compared to the model based on PO data alone. This suggests that the precision gained in the estimates of these model components is due to the integration of heterogeneous data sources. Similar findings were obtained by Seaton et al. (2024), who proposed a spatiotemporal model-based data integration approach to understand the spatiotemporal change in the Gatekeeper butterfly (*Pyronia tithonus*) and caddisfly (*Trichoptera*) populations across Great Britain and the River Thames catchment, respectively. Furthermore, the integrated model implemented in this study has been shown to significantly enhance habitat suitability predicted for PO data compared to modelling PO data alone. This is primarily due to the substantial congruence between the fusion model predictions and the standalone model predictions. This finding corroborates the hypothesis that the fusion model, by borrowing information from the abundance data, has the capacity to correct for bias in PO data (Dorazio 2014), even if this bias appears to be negligible in the present study. The slightly lower estimates produced by the shared component model at very high intensities compared to the standalone PO model may be due to the shared spatial latent field. This component of the integrated model, which may function as a bias filter for PO response and stabiliser for abundance response, tends to impose a more globally consistent structure, thereby preventing the standalone model from predicting excessively high localised peaks. For instance, an analysis of the observed

species densities at the sampled sites reveals three sites with the maximum density of 12 individuals per square kilometre, whereas the second-highest density in the region is 8 individuals per square kilometre. In the integrated model, the predictions of relative abundance at locations with extreme densities would have been slightly reduced by the shared spatial signal, which acts as a smoother or regulariser of the density estimates. Furthermore, the near-perfect agreement in local density predictions ($r = 0.93$) combined with the high correlation on the log-intensity scale ($r = 0.66$) confirms the capacity of the integrated model to simultaneously preserve the local fidelity of the abundance information while maintaining the structural consistency of the underlying spatial process under investigation.

Despite the superior performance of the proposed integrated model compared to the standalone models concerning unseen data and its more accurate prediction of observed abundance, it is possible that the entire range of species abundance may not be adequately captured by a fusion model, particularly in unsampled areas. This insight may be attributable to the disparity in the size of the presence-only and abundance datasets (723 locations for PO and 142 for counts), with the presence-only data seemingly predominating the results. This fact has been highlighted in recent studies by Fletcher Jr. et al. (2019) and Simmonds et al. (2020). One potential avenue for further exploration could involve testing extensions to fusion model structures, such as weighting the likelihoods of integrated datasets to account for variations in either data size or its quality (Fletcher Jr. et al. 2019). Furthermore, potentially biased data, such as presence-only data, are easier to collect than abundance data, which are often unbiased by design but more costly to obtain. Therefore, it is crucial to be able to investigate evidence of bias within the datasets being analysed. This is the function of the proposed spatial fusion modelling framework that tends to borrow from the strengths of each dataset. Furthermore, previous studies have shown that integrated

models can improve estimates of environmental covariates and model predictions even with a small proportion of structured data sets (i.e. presence-absence or count data) (Dorazio 2014; Fithian et al. 2015). The present findings are consistent with those of the aforementioned studies, due to the shared model's capacity to estimate common regression coefficients or smoothed functions, identify the species hotspots (areas with species presence and high local abundance) and enhance the precision of the resulting predictions. However, it is possible that the full range of species abundance may not be adequately captured by a fusion model, particularly when certain datasets are of poor quality and appropriate weights are not used (Fletcher Jr. et al. 2019). These findings demonstrate that when high-quality structured data are available and the primary objective is to predict species abundance, integrating them with lower-quality data may not result in additional benefits. Nonetheless, it can be argued that critical observations derived from this empirical study should not be regarded as a shortcoming of the integrated model in comparison to the standalone models. This is because the true species distribution against which we can compare the predictions from different models under investigation remains unknown. A simulation study based on a virtual species in which the experimenter sets the model's parameters (and subsequently knows the true distribution of the species) may facilitate further investigations into the benefits of a fusion modelling framework under different model configurations and characteristics of disparate datasets. Furthermore, it is argued that the standard resolution of one kilometre utilised in many SDM approaches is adequate for capturing the species' response to the observed environmental conditions at the scale of analysis. Nonetheless, we recommend exploring alternative modelling strategies for the integrated SDMs. One such strategy is the use of deep learning-based SDMs (Botella et al. 2018) to predict the extent of species distribution and its relationship with environmental drivers in the context of the data fusion framework. Convolutional neural networks represent a class of deep learning techniques that

have been demonstrated to facilitate the efficient learning of complex relationships between species distribution and environmental drivers. This is achieved by identifying spatial patterns in environmental variables (Botella et al. 2018) or the influence of local environmental landscapes on the species distribution (Deneu et al. 2021).

Notwithstanding the challenges previously mentioned, the Bayesian ISDM implemented in this study facilitates the estimation of the average habitat suitability of the species and other relevant quantities of interest. Furthermore, it enables the propagation of uncertainty surrounding the mean by providing a Bayesian credible interval. These intervals are narrower for the integrated model than for models based on a single dataset, indicating the benefit of combining datasets from multiple sources. Another benefit of the proposed fusion modelling framework is its ability to account for the dependence between the conditional distributions of both target outcomes by allowing some parameters to be shared across the individual observation models, thereby enhancing the estimation and interpretability of the resulting models. The shared latent component modelling approach, or model-based data fusion in general, has been demonstrated to apply to a wide range of spatial modelling settings involving multiple outcomes. These settings include disease mapping (Mahaki et al. 2018), remote sensing (Shi and Kang 2017) and crime prevention (Law et al. 2020). Moreover, the cross-validation procedure implemented in this study has enabled the identification of a fusion model that is robust to spatial autocorrelation and its subsequent generalisation to unobserved locations. In this regard, it is essential to note that a similar block cross-validation technique can be readily applied to other studies dealing with disparate sources of spatial data using the proposed *isdmtools* R package. The validation of models by means of independent or withheld data is of great importance when assessing the performance of SDMs or machine learning models, particularly in the context of limited data.

4.3 Limitations of the study and future perspectives

The model-based data integration approach demonstrated in this study, which aims to understand the spatial variation of the baobab population and its relationships with environmental factors, presents certain benefits and challenges. Indeed, the spatial models discussed here include spatial latent components to account for other spatially correlated covariates that were not included in the analysis, the uneven sampling efforts inherent in the presence-only data, and unmeasured environmental drivers. The latter category includes dispersal barriers, which can influence the distribution of certain plant and animal species. Although spatial latent components can capture dispersal constraints, their correlation structure was modelled under a stationarity assumption and could not incorporate dispersal barriers, as proposed by Bakka et al. (2019). Nonetheless, baobab is a long-dispersal plant species and is widely distributed in sub-Saharan Africa. It is possible that it does not have physical dispersal barriers on the scale of this analysis. Moreover, it is conceivable that the assumption of linear relationships between the species' intensity and some environmental covariates may be relaxed following the visualisation of the relationships between the selected covariates and the predicted species intensity. In this instance, suspicion was raised about a nonlinear relationship between the species intensity and the rainfall in the driest month, which was then modelled using a one-dimensional SPDE approach. Although this approach and other well-known smoothing spline techniques (Gómez-Rubio 2021) may be more appealing in a data fusion application with few covariates, they have the potential to increase model complexity and computational cost and to result in a loss of direct interpretability in situations involving many covariates with complex relationships to species intensity. Generally, more structured data on the target species can help to better estimate the complex environmental relationships with the species' habitat. Deep learning approaches, such as convolutional neural network (CNN) architectures, can also help to capture the multidimensional spatial

patterns in raster images used as environmental features (Deneu et al. 2021; Botella et al. 2018). However, it may be challenging to interpret the resulting relationships between environmental features and species intensity directly, unless the primary objective of the study is essentially prediction.

Moreover, the integrated model results are conditional on the critical assumption that the underlying spatial signal is shared similarly across data types. However, it should be noted that this assumption may be invalidated if the two sampling processes exhibit structurally different biases that have not been accounted for by the fusion model. The sampling protocol used to record the observed abundance of baobab trees was assumed to be nonpreferential, on the basis of its random nature. However, this assumption may need to be reconsidered if the sites that have generated the abundance are not stochastically independent of the underlying spatial latent signal. This scenario typically arises when the sampling process is not randomised, resulting in a biased sample. In this case, the appropriate joint modelling strategy would be a data fusion framework under a preferential sampling (Gelfand and Shirota 2019). Finally, the integrated model discussed in the present study employed a Poisson distribution to model the baobab abundance response, which includes some extra zeros. However, the estimated proportion of these observations through a ZIP model is small ($p(Y_i = 0) \approx 0.09$) and has a negligible effect on the estimated abundance compared to the traditional Poisson model. Nonetheless, future studies could consider using a zero-inflated Poisson (ZIP) distribution to model additional zeros in the count data. This would allow their impact on model estimates and predictions to be accounted for, particularly in cases involving a rare species where the proportion of additional zeros is significant. The data fusion framework addressed in this study for baobab datasets could undoubtedly be applied to spatio-temporal modelling contexts if structured data on species abundance are available at different time points (Seaton et al. 2024). This

would assist in assessing the spatiotemporal dynamics of the baobab distribution in the country.

5 Conclusion

This study aimed to elucidate the most suitable spatial fusion models for the joint analysis of abundance and presence-only target responses, to assess the spatial distribution of African baobab in Benin. The fusion models discussed are constructed based on geostatistical and log-Gaussian Cox process observation models for abundance and presence-only data, respectively, using approximate Bayesian inference. The results demonstrate the existence of spatial heterogeneity in the distribution of species across Benin, with a notable concentration of individuals in the northern region of the country. Key environmental factors, including annual temperature, rainfall in the driest month, slope, and soil texture, have a significant influence on the baobab distribution in the region. This suggests that environmental factors play an important role in shaping the distribution of plant species in the region. The present study illustrates the utility of integrated species distribution modelling in elucidating the spatial variation of a species intensity using disparate sources of spatial data. The approach enables the sharing of particular environmental parameters among submodels, thereby enhancing the estimation and interpretability of the resulting models. However, further investigation is required to determine the circumstances under which a joint model will be preferred to separate models when the primary objective is the prediction of local abundance. The methodological principles demonstrated here in the specific context of SDM are more widely applicable to any spatial regression model aiming to analyse multiple outcomes within a Bayesian framework. The insights derived from the joint analysis of the baobab datasets have the potential to provide valuable guidelines for statisticians, biometricians, and ecologists, enabling the development of robust spatial models to support planning and management strategies for sustainable agroforestry

practices in sub-Saharan Africa. It is recommended that future research explore the potential of integrating temporal climate variability with additional structured data to enhance the predictive performance of ISDMs.

Supplementary Information. This paper is accompanied by supplementary information, which can be found in the web appendix.

Acknowledgements. The authors express their gratitude to the African Institute of Mathematical Sciences Research and Innovation Centre (AIMS-RIC) for its general support. They are also grateful to Peter J. Diggle for his valuable comments and suggestions on the initial manuscript draft.

Author contributions. SAI, FAB, and GKR conceived the study. AAE and SAI collected the data. SAI developed the methodology and wrote the code and the original draft. SAI and KET analysed the data. FAB and GKR supervised the work. All authors have read, reviewed and approved the manuscript for publication.

Code availability. The R package *isdmtools* that has been developed for facilitating the cross-validation procedure, model evaluation and habitat suitability mapping is publicly available for installation on the following GitHub repository: https://github.com/sodeidelphonse/isdmtools.

Data availability. Data is available upon request.

Funding. Not applicable.

Declarations

Conflict of interest. The authors declare no conflict of interest.

References

Adde, A., Amat, C., Mazerolle, M.J., Darveau, M., Cumming, S.G., O'Hara, R.B.: Integrated modeling of waterfowl distribution in western Canada using aerial survey and citizen science (eBird) data. Ecosphere **12**(10), 03790 (2021) https://doi.org/10.1002/ecs2.3790 .

Amoah, B., Diggle, P.J., Giorgi, E.: A geostatistical framework for combining spatially referenced disease prevalence data from multiple diagnostics. Biometrics **76**(1), 158–170 (2020) https://doi.org/10.1111/biom.13142 .

Adjahossou, S.G.C., Gouwakinnou, G.N., Houehanou, D.T., Sode, A.I., Yaoitcha, A.S., Houinato, M.R.B., Sinsin, B.: Efficacité des aires protégées dans la conservation d'habitats favorables prioritaires de ligneux de valeur au Bénin. Bois & Forêts des Tropiques **328**, 67–76 (2016) https://doi.org/10.19182/bft2016.328.a31303 .

Assogba, D., Idohou, R., Chirwa, P., Assogbadjo, A.E.: On opportunities and challenges to conserve the African baobab under present and future climates in Benin (West Africa). Journal of Arid Environments **198**, 104692 (2022) https://doi.org/10.1016/j.jaridenv.2021.104692 .

Assogbadjo, A.E., Sinsin, B., Codjia, J.T.C., Van Damme, P.: Ecological diversity and pulp, seed and kernel production of the baobab (Adansonia digitata) in Benin. Belgian Journal of Botany, 47–56 (2005).

Banerjee, S., Carlin, B.P., Gelfand, A.E.: Hierarchical Modeling and Analysis for Spatial Data, Second Edition, 2ed. edn. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. CRC Press, New York, NY (2014).

Botella, C., Joly, A., Bonnet, P., Monestiez, P., Munoz, F.: A Deep Learning Approach

to Species Distribution Modelling. In: Joly, A., Vrochidis, S., Karatzas, K., Karppinen, A., Bonnet, P. (eds.) Multimedia Tools and Applications for Environmental & Biodiversity Informatics, pp. 169–199. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-76445-0_10 .

Bachl, F.E., Lindgren, F., Borchers, D.L., Illian, J.B.: inlabru: an R package for Bayesian spatial modelling from ecological survey data. Methods in Ecology and Evolution **10**(6), 760–766 (2019) https://doi.org/10.1111/2041-210X.13168 .

Bakka, H., Vanhatalo, J., Illian, J.B., Simpson, D., Rue, H.: Non-stationary Gaussian models with physical barriers. Spatial Statistics **29**, 268–288 (2019) https://doi.org/10.1016/j.spasta.2019.01.002 .

Dansi, A., Adjatin, A., Adoukonou-Sagbadja, H., Faladé, V., Yedomonhan, H., Odou, D., Dossou, B.: Traditional leafy vegetables and their use in the Benin Republic. Genetic Resources and Crop Evolution **55**(8), 1239–1256 (2008) https://doi.org/10.1007/s10722-008-9324-z .

Diggle, P.J.: Statistical Analysis of Spatial and Spatio-Temporal Point Patterns, Third Edition, 3rd edn. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. CRC Press, Boca Raton, FL (2013).

Dambly, L.I., Isaac, N.J.B., Jones, K.E., Boughey, K.L., O'Hara, R.B.: Integrated species distribution models fitted in INLA are sensitive to mesh parameterisation. Ecography **2023**(7), 06391 (2023) https://doi.org/10.1111/ecog.06391 .

Diggle, P.J., Menezes, R., Su, T.-l.: Geostatistical inference under preferential sampling. Journal of the Royal Statistical Society Series C: Applied Statistics **59**(2), 191–232 (2010) https://doi.org/10.1111/j.1467-9876.2009.00701.x .

Dormann, C.F.: Effects of incorporating spatial autocorrelation into the analysis of

species distribution data. Global Ecology and Biogeography **16**(2), 129–138 (2007) https://doi.org/10.1111/j.1466-8238.2006.00279.x .

Dorazio, R.M.: Accounting for imperfect detection and survey bias in statistical analysis of presence-only data. Global Ecology and Biogeography **23**(12), 1472–1484 (2014) https://doi.org/10.1111/geb.12216 .

Dovers, E., Popovic, G.C., Warton, D.I.: A fast method for fitting integrated species distribution models. Methods in Ecology and Evolution (2023) https://doi.org/10.1111/2041-210X.14252 .

Diggle, P.J., Ribeiro, P.J.: Model-based Geostatistics, 1st edn. Springer Series in Statistics. Springer, New York, NY (2007). https://doi.org/10.1007/978-0-387-48536-2 .

Deneu, B., Servajean, M., Bonnet, P., Botella, C., Munoz, F., Joly, A.: Convolutional neural networks improve species distribution modelling by capturing the spatial structure of the environment. PLOS Computational Biology **17**(4), 1008856 (2021) https://doi.org/10.1371/journal.pcbi.1008856 .

Elith, J., Leathwick, J.R.: Species Distribution Models: Ecological Explanation and Prediction Across Space and Time. Annual Review of Ecology, Evolution, and Systematics **40**(Volume 40, 2009), 677–697 (2009) https://doi.org/10.1146/annurev.ecolsys.110308.120159 .

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

Fletcher Jr., R.J., Hefley, T.J., Robertson, E.P., Zuckerberg, B., McCleery, R.A., Dorazio, R.M.: A practical guide for combining data to model species distributions.

Ecology **100**(6), 02710 (2019) https://doi.org/10.1002/ecy.2710 .

Fandohan, A.B., Oduor, A.M.O., Sodé, A.I., Wu, L., Cuni-sanchez, A., Assédé, E., Gouwakinnou, G.N.: Modeling vulnerability of protected areas to invasion by chromolaena odorata under current and future climates. Ecosystem Health and Sustainability **1**(6), 1–12 (2015) https://doi.org/10.1890/EHS15-0003.1 .

Foster, S.D., Peel, D., Hosack, G.R., Hoskins, A., Mitchell, D.J., Proft, K., Yang, W.-H., Uribe-Rivera, D.E., Froese, J.G.: 'RISDM': species distribution modelling from multiple data sources in R. Ecography **2024**(6), 06964 (2024) https://doi.org/10.1111/ecog.06964 .

Guillera-Arroita, G., Lahoz-Monfort, J.J., Elith, J., Gordon, A., Kujala, H., Lentini, P.E., McCarthy, M.A., Tingley, R., Wintle, B.A.: Is my species distribution model fit for purpose? Matching data and models to applications. Global Ecology and Biogeography **24**(3), 276–292 (2015) https://doi.org/10.1111/geb.12268 .

Giraud, C., Calenge, C., Coron, C., Julliard, R.: Capitalizing on opportunistic data for monitoring relative abundances of species. Biometrics **72**(2), 649–658 (2016) https://doi.org/10.1111/biom.12431 .

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

Gómez-Rubio, V.: Bayesian Inference with INLA. Chapman & Hall/CRC, Boca Raton, FL (2021). https://becarioprecario.bitbucket.io/inla-gitbook/index.html .

Gelfand, A.E., Shirota, S.: Preferential sampling for presence/absence data and for fusion of presence/absence data with presence-only data. Ecological Monographs **89**(3), 01372 (2019) https://doi.org/10.1002/ecm.1372 .

Isaac, N.J.B., Jarzyna, M.A., Keil, P., Dambly, L.I., Boersch-Supan, P.H., Browning, E., Freeman, S.N., Golding, N., Guillera-Arroita, G., Henrys, P.A., Jarvis, S., Lahoz-Monfort, J., Pagel, J., Pescott, O.L., Schmucki, R., Simmonds, E.G., O'Hara, R.B.: Data Integration for Large-Scale Models of Species Distributions. Trends in Ecology & Evolution **35**(1), 56–67 (2020) https://doi.org/10.1016/j.tree.2019.08.006 .

Jung, M.: An integrated species distribution modelling framework for heterogeneous biodiversity data. Ecological Informatics **76**, 102127 (2023) https://doi.org/10.1016/j.ecoinf.2023.102127 .

Krainski, E., Gómez-Rubio, V., Bakka, H., Lenzi, A., Castro-Camilo, D., Simpson, D., Lindgren, F., Rue, H.: Advanced Spatial Modeling with Stochastic Partial Differential Equations Using R and INLA. Chapman and Hall/CRC, Boca Raton, FL (2018) https://becarioprecario.bitbucket.io/spde-gitbook/index.html .

Luedeling, E., Kindt, R., Huth, N.I., Koenig, K.: Agroforestry systems in a changing climate — challenges in projecting future performance. Current Opinion in Environmental Sustainability **6**, 1–7 (2014) https://doi.org/10.1016/j.cosust.2013.07.013 .

Law, J., Quick, M., Jadavji, A.: A Bayesian spatial shared component model for identifying crime-general and crime-specific hotspots. Annals of GIS **26**(1), 65–79 (2020) https://doi.org/10.1080/19475683.2020.1720290 .

Lindgren, F., Rue, H.: Bayesian spatial modelling with R-INLA. Journal of statistical software **63**, 1–25 (2015) https://doi.org/10.18637/jss.v063.i19

Lindgren, F., Rue, H., Lindström, J.: An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation

approach. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **73**(4), 423–498 (2011) https://doi.org/10.1111/j.1467-9868.2011.00777.x .

Mugumaarhahama, Y., Fandohan, A.B., Mushagalusa, A.C., Sode, A.I., Glèlè Kakaï, R.L.: Inhomogeneous Poisson point process for species distribution modelling: relative performance of methods accounting for sampling bias and imperfect detection. Modeling Earth Systems and Environment **8**(4), 5419–5432 (2022) https://doi.org/10.1007/s40808-022-01417-3 .

Mahaki, B., Mehrabi, Y., Kavousi, A., Schmid, V.J.: Joint spatio-temporal shared component model with an application in Iran Cancer Data. Asian Pacific journal of cancer prevention: APJCP **19**(6), 1553 (2018) https://doi.org/10.22034/APJCP.2018.19.6.1553 .

Mäkinen, J., Numminen, E., Niittynen, P., Luoto, M., Vanhatalo, J.: Spatial confounding in Bayesian species distribution modeling. Ecography **2022**(11), 06183 (2022) https://doi.org/10.1111/ecog.06183 .

Mostert, P.S., O'Hara, R.B.: PointedSDMs: An R package to help facilitate the construction of integrated species distribution models. Methods in Ecology and Evolution (2023) https://doi.org/10.1111/2041-210X.14091 .

McCulloch, C.E., Searle, S.R.: Generalized, Linear, and Mixed Models. John Wiley & Sons, New York, NY (2001) https://doi.org/10.1002/0471722073 .

Møller, J., Syversveen, A.R., Waagepetersen, R.P.: Log Gaussian Cox Processes. Scandinavian Journal of Statistics **25**(3), 451–482 (1998) https://doi.org/10.1111/1467-9469.00115 .

Mbow, C., Van Noordwijk, M., Luedeling, E., Neufeldt, H., Minang, P.A., Kowero, G.: Agroforestry solutions to address food security and climate change challenges

in Africa. Current Opinion in Environmental Sustainability **6**, 61–67 (2014) https://doi.org/10.1016/j.cosust.2013.10.014 .

Phillips, S.J., Anderson, R.P., Dudík, M., Schapire, R.E., Blair, M.E.: Opening the black box: an open-source release of Maxent. Ecography **40**(7), 887–893 (2017) https://doi.org/10.1111/ecog.03049 .

Platts, P.J., Omeny, P.A., Marchant, R.: AFRICLIM: high-resolution climate projections for ecological applications in Africa. African Journal of Ecology **53**(1), 103–108 (2015) https://doi.org/10.1111/aje.12180 .

R Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2024). https://www.R-project.org/

Renner, I.W., Elith, J., Baddeley, A., Fithian, W., Hastie, T., Phillips, S.J., Popovic, G., Warton, D.I.: Point process models for presence-only analysis. Methods in Ecology and Evolution **6**(4), 366–379 (2015) https://doi.org/10.1111/2041-210X.12352 .

Rue, H., Martino, S., Chopin, N.: Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. Journal of the Royal Statistical Society Series B: Statistical Methodology **71**(2), 319–392 (2009) https://doi.org/10.1111/j.1467-9868.2008.00700.x .

Rue, H., Riebler, A., Sørbye, S.H., Illian, J.B., Simpson, D.P., Lindgren, F.K.: Bayesian computing with INLA: a review. Annual Review of Statistics and Its Application **4**, 395–421 (2017) https://doi.org/10.1146/annurev-statistics-060116-054045 .

Renner, I.W., Warton, D.I.: Equivalence of MAXENT and Poisson Point Process Models for Species Distribution Modeling in Ecology. Biometrics **69**(1), 274–281 (2013) https://doi.org/10.1111/j.1541-0420.2012.01824.x .

Simpson, D., Illian, J.B., Lindgren, F., Sørbye, S.H., Rue, H.: Going off grid: Computationally efficient inference for log-Gaussian Cox processes. Biometrika **103**(1), 49–70 (2016) https://doi.org/10.1093/biomet/asv064 .

Simmonds, E.G., Jarvis, S.G., Henrys, P.A., Isaac, N.J.B., O'Hara, R.B.: Is more data always better? A simulation study of benefits and limitations of integrated distribution models. Ecography **43**(10), 1413–1422 (2020) https://doi.org/10.1111/ecog.05146 .

Seaton, F.M., Jarvis, S.G., Henrys, P.A.: Spatio-temporal data integration for species distribution modelling in R-INLA. Methods in Ecology and Evolution **n**/**a**(n/a) (2024) https://doi.org/10.1111/2041-210X.14356 .

Shi, H., Kang, E.L.: Spatial data fusion for large non-Gaussian remote sensing datasets. Stat (International Statistical Institute) **6**(1), 390–404 (2017) https://doi.org/10.1002/sta4.165 .

Simpson, D., Lindgren, F., Rue, H.: In order to make spatial statistics computationally feasible, we need to forget about the covariance function. Environmetrics **23**(1), 65–74 (2012) https://doi.org/10.1002/env.1137 .

Sydenham, M.A.K., Moe, S.R., Kuhlmann, M., Potts, S.G., Roberts, S.P.M., Totland, O., Eldegard, K.: Disentangling the contributions of dispersal limitation, ecological drift, and ecological filtering to wild bee community assembly. Ecosphere **8**(1), 01650 (2017) https://doi.org/10.1002/ecs2.1650 .

Simpson, D., Rue, H., Riebler, A., Martins, T.G., Sørbye, S.H.: Penalising Model Component Complexity: A Principled, Practical Approach to Constructing Priors. Statistical Science **32**(1), 1–28 (2017) https://doi.org/10.1214/16-STS576 .

Whittle, P.: On Stationary Processes in the Plane. Biometrika **41**(3/4), 434–449 (1954)

https://doi.org/10.2307/2332724 .

Whittle, P.: Stochastic-processes in several dimensions. Bulletin of the International Statistical Institute **40**(2), 974–994 (1963).

Wisz, M.S., Pottier, J., Kissling, W.D., Pellissier, L., Lenoir, J., Damgaard, C.F., Dormann, C.F., Forchhammer, M.C., Grytnes, J., Guisan, A., Heikkinen, R.K., Høye, T.T., Kühn, I., Luoto, M., Maiorano, L., Nilsson, M., Normand, S., Öckinger, E., Schmidt, N.M., Termansen, M., Timmermann, A., Wardle, D.A., Aastrup, P., Svenning, J.: The role of biotic interactions in shaping distributions and realised assemblages of species: implications for species distribution modelling. Biological Reviews **88**(1), 15–30 (2013) https://doi.org/10.1111/j.1469-185X.2012.00235.x .

Zurell, D., Thuiller, W., Pagel, J., Cabral, J.S., Münkemüller, T., Gravel, D., Dullinger, S., Normand, S., Schiffers, K.H., Moore, K.A., Zimmermann, N.E.: Benchmarking novel approaches for modelling species range dynamics. Global Change Biology **22**(8), 2651–2664 (2016) https://doi.org/10.1111/gcb.13251 .

Appendix A Supplementary information for the methods section

Appendix B Supplementary information for the results section

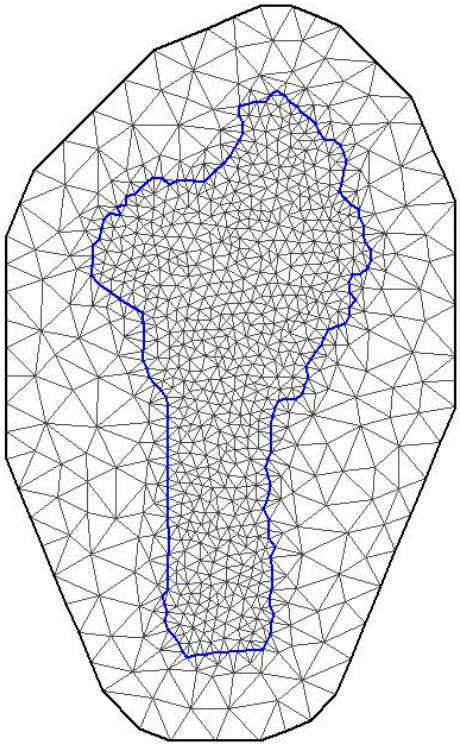

Fig. A.1 The two-dimensional mesh constructed from *Delaunay triangulation* of the study domain to approximate the Gaussian latent fields through the use of the finite element method. This mesh was also employed in the construction of the integration points that were used for approximating the log-Gaussian Cox process likelihood

Table 2 Comparison of integrated models using the ROC composite scores and the Deviance Information Criterion (DIC) from block cross-validation

Integrated models	AUC		TSS		Accuracy		Overall ROC Score		DIC	
	Mean	SE	Mean	SE	Mean	SE	Mean	SE	Mean	SE
Shared	0.855^{a}	0.021	0.662^{a}	0.048	0.773^{a}	0.022	0.763^{a}	0.025	-8897	324
Independent	0.833^{ab}	0.023	0.634^{ab}	0.047	0.760^{ab}	0.021	0.742^{ab}	0.025	-8701	339
Shared-scaled	0.827^{ab}	0.026	0.625^{ab}	0.040	0.743^{ab}	0.022	0.732^{abc}	0.027	-8657	333
Shared-scov	0.781^{ab}	0.021	0.513^{abc}	0.033	0.702^{abc}	0.032	0.666^{abcd}	0.010	-8707	322
Shared-bias	0.759^{ab}	0.030	0.504^{abc}	0.044	0.725^{abc}	0.028	0.663^{abcd}	0.025	-8668	337
Shared-scaled-scov	0.748^{abc}	0.016	0.474^{abc}	0.032	0.697^{abc}	0.036	0.64^{abcde}	0.015	-8393	295
Shared-scaled-bias	0.721^{bcd}	0.037	0.465^{bc}	0.056	0.676^{abc}	0.033	0.62^{bcde}	0.038	-8695	337
Independent-scov	0.721^{bcd}	0.011	0.447^{abc}	0.033	0.662^{bc}	0.038	0.61^{cde}	0.017	-8427	316
Shared-bias-scov	0.632^{cd}	0.024	0.328^{c}	0.042	0.691^{abc}	0.052	0.55^{de}	0.022	-8409	317
Shared-scaled-bias-scov	0.615^{d}	0.026	0.305^{c}	0.047	0.626^{c}	0.059	0.515^{c}	0.038	-8421	316

Note: Shared = Shared spatial effect, Independent = Independent spatial effects, Shared-scaled = Shared spatial effect + scaling factor, Shared-scov = Shared spatial effect + data-specific covariates effects, Shared-bias = Shared spatial effect + bias in presence-only data, Shared-scaled-scov = Shared spatial effect + scaling factor + data-specific covariates effects, Shared-scaled-bias = Shared spatial effect + scaling factor + bias in presence-only data, Independent-scov = Independent spatial effects + data-specific covariates effects, Shared-bias-scov = Shared spatial effect + bias in presence-only data + data-specific covariates effects, and Shared-scaled-bias-scov = Shared spatial effect + scaling factor + bias in presence-only data + data-specific covariates effects and SE = standard error. Note that models were ranked based on the overall ROC score. Models that have the same small letter(s) for a given metric are not significantly different according to the least significant difference (LSD) post-hoc test

Table 3 Effects of environmental factors affecting baobab distribution and hyperparameters of the spatial latent field estimated from the shared component model with common covariates' effects

Model parameters	Mean	SD	2.5% quantile	Median	97.5% quantile	Mode
Fixed effects of covariates						
Annual temperature	0.185	0.084	0.02	0.185	0.35	0.185
SRTM slope	0.135	0.031	0.073	0.135	0.197	0.135
Soil texture silt fraction	-0.267	0.072	-0.407	-0.267	-0.127	-0.267
Soil texture clay fraction	-0.046	0.038	-0.122	-0.046	0.029	-0.046
Presence intercept	-6.103	0.211	-6.516	-6.103	-5.689	-6.103
Abundance intercept	-0.199	0.211	-0.614	-0.199	0.215	-0.199
Hyperparameters						
Range for $S_0(x)$	34.475	3.816	27.6	34.257	42.61	33.807
SD for $S_0(x)$	1.022	0.058	0.914	1.02	1.14	1.017
Range for f(Bio14)	0.966	0.52	0.311	0.851	2.29	0.66
SD for f(Bio14)	0.59	0.214	0.275	0.555	1.11	0.491

Note: Bio14 = Rainfall of the driest month, $f(.)$ = a smooth function estimated with a one-dimensional SPDE, SD = Standard deviation of the posterior distribution and $S_0(.)$ = the shared spatial latent component. The limited spatial range of f(Bio14) is indicative of the manner in which the one-dimensional SPDE functions to capture the local variation of the smoothed effect of a covariate

Table 4 Effects of environmental factors affecting baobab distribution and hyperparameters of the spatial latent field estimated from the LGCP model for presence-only data and Poisson GLMM for abundance data

Model parameters	Mean	SD	2.5% quantile	Median	97.5% quantile	Mode
GLMM for abundance data						
Annual temperature	0.165	0.067	0.033	0.165	0.297	0.165
SRTM slope	0.021	0.04	-0.058	0.021	0.101	0.021
Soil texture silt fraction	0.083	0.071	-0.055	0.083	0.221	0.083
Soil texture clay fraction	-0.092	0.038	-0.166	-0.092	-0.018	-0.092
Abundance intercept	0.112	0.176	-0.233	0.112	0.457	0.112
Hyperparameters						
Range for S(x)	21.405	5.254	12.784	20.83	33.34	19.772
SD for S(x)	0.578	0.07	0.454	0.574	0.73	0.564
Range for f(Bio14)	3.395	2.441	1.093	2.715	9.95	1.837
SD for f(Bio14)	0.874	0.255	0.472	0.841	1.47	0.779
LGCP model for presence-only data						
Annual temperature	1.024	0.102	0.824	1.024	1.224	1.024
SRTM slope	0.322	0.05	0.223	0.322	0.42	0.322
Soil texture silt fraction	-0.91	0.109	-1.124	-0.91	-0.696	-0.91
Soil texture clay fraction	-0.079	0.079	-0.234	-0.079	0.076	-0.079
Presence-only intercept	-6.353	0.247	-6.837	-6.353	-5.868	-6.353
Hyperparameters						
Range for S(x)	32.241	2.94	26.838	32.11	38.4	31.851
SD for S(x)	1.043	0.054	0.94	1.041	1.15	1.039
Range for f(Bio14)	0.754	0.255	0.367	0.716	1.36	0.646
SD for f(Bio14)	1.081	0.251	0.669	1.054	1.65	1.002

Note: Bio14 = Rainfall of the driest month, $f(.)$ = a smooth function estimated with a one-dimensional SPDE, SD = Standard deviation of the posterior distribution and S(.) = the spatial latent component

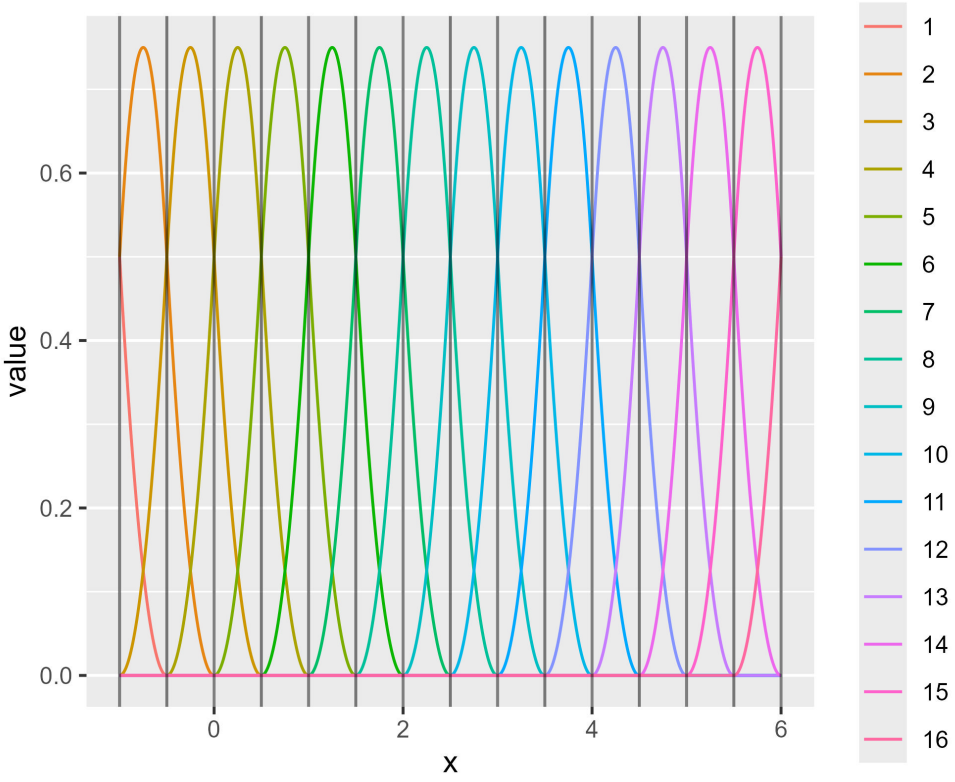

Fig. A.2 The one-dimensional mesh constructed to estimate the nonlinear effect of the rainfall of the driest month (Bio14) using the finite element method

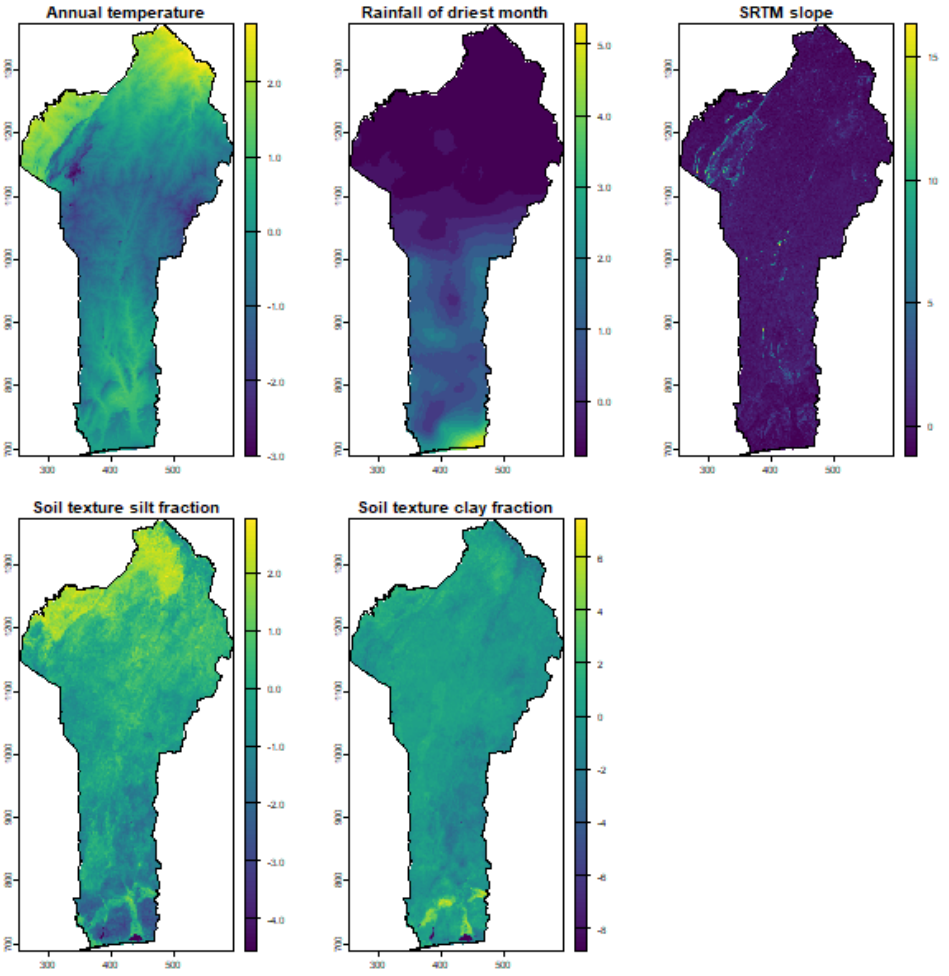

Fig. A.3 Map of the standardised covariates included in all spatial models that were fitted to the baobab datasets. The soil texture clay fraction is for 150 centimetres (cm) depth, while the silt fraction is for 10 centimetres depth

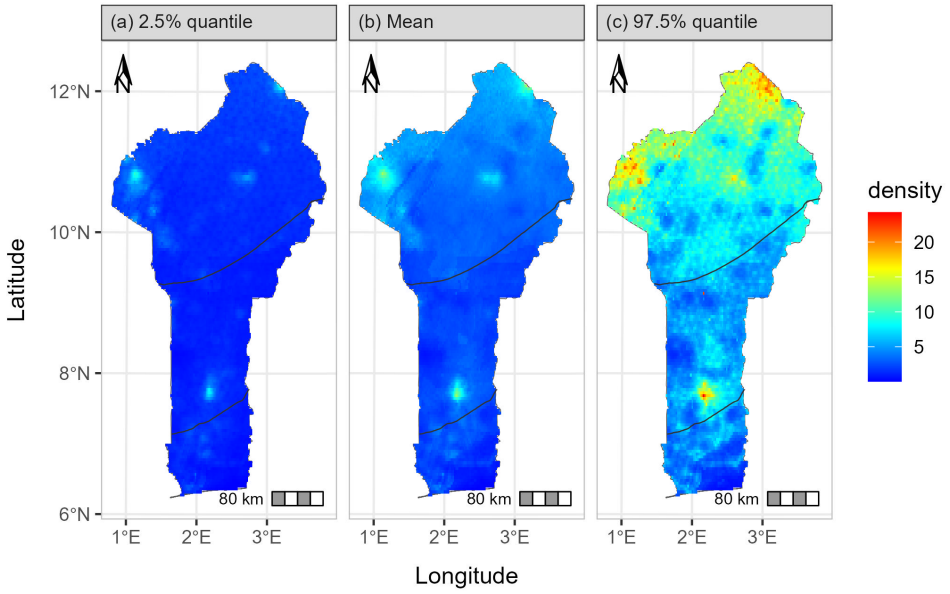

Fig. B.1 Map of the predicted density (relative abundance) of baobab from the Bayesian generalised linear mixed-effects model (GLMM) fitted to the count data - (a) posterior mean, (b) 2.5% posterior quantile, and (c) 97.5% posterior quantile

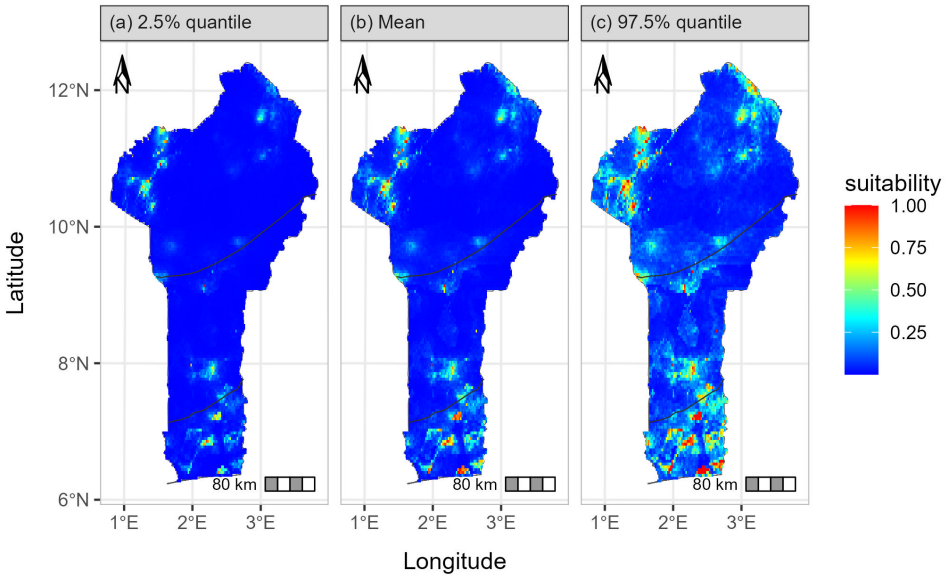

Fig. B.2 Map of the predicted habitat suitability of baobab from the Bayesian log-Gaussian Cox process (LGCP) model fitted to the presence-only data - (a) posterior mean, (b) 2.5% posterior quantile, and (c) 97.5% posterior quantile

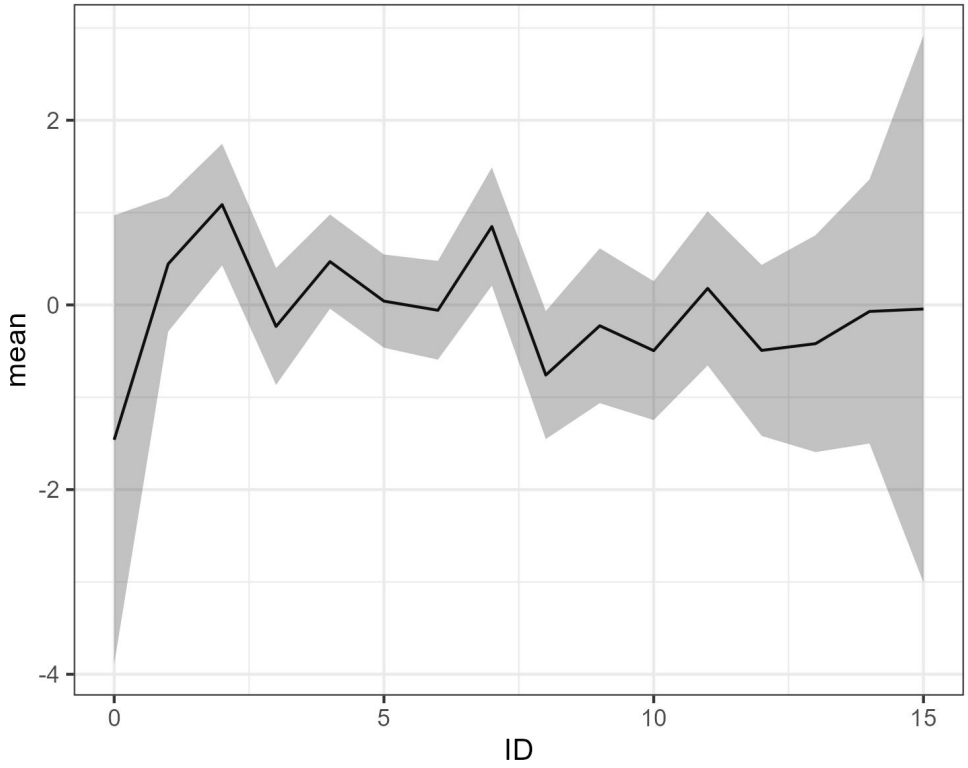

Fig. B.3 Visualisation of the non-linear effect of the rainfall in the driest month (solid line) estimated from the shared spatial component model with common covariates' effects. The grey band around the solid line represents the 95% credible interval around the estimated function

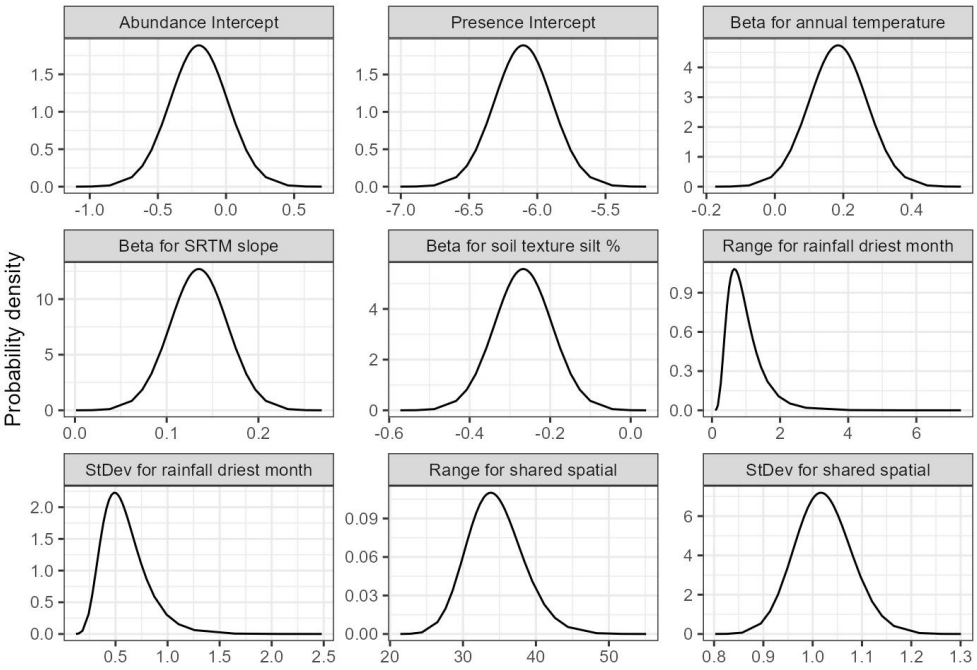

Fig. B.4 Marginal distribution of parameters and hyperparameters estimated from the shared spatial component model with common covariates' effects