



# Species distribution modelling through Bayesian hierarchical approach

Oscar Rodríguez de Rivera<sup>1,2</sup> · Marta Blangiardo<sup>2</sup> · Antonio López-Quílez<sup>1</sup> · Ignacio Martín-Sanz<sup>3</sup>

Received: 30 June 2017 / Accepted: 21 June 2018 / Published online: 5 July 2018  
© Springer Nature B.V. 2018

## Abstract

Usually in Ecology, the availability and quality of the data is not as good as we would like. For some species, the typical environmental study focuses on presence/absence data, and particularly with small animals as amphibians and reptiles, the number of presences can be rather small. The aim of this study is to develop a spatial model for studying animal data with a low level of presences; we specify a Gaussian Markov Random Field for modelling the spatial component and evaluate the inclusion of environmental covariates. To assess the model suitability, we use Watanabe-Akaike information criteria (WAIC) and the conditional predictive ordinate (CPO). We apply this framework to model each species of amphibian and reptiles present in the Las Tablas de Daimiel National Park (Spain).

**Keywords** Stochastic partial differential equation · Integrated nested Laplace approximation · Species distribution · Spatial model

## Introduction

Species distribution models (SDM) commonly used in ecology consist of numerical tools that combine observations of species occurrence or abundance with environmental covariates. They are used to gain ecological and evolutionary insight from SDM and to predict distributions across landscapes, sometimes requiring extrapolation in space and time (Elith and Leathwick 2009). In SDM, the following steps are usually taken: (1) locations of occurrence of a species (or other phenomenon) are compiled; (2) values of environmental predictor variables (such as climate) at these locations are extracted from spatial databases; (3) the environmental values are used to fit a model to estimate similarity to the sites of occurrence, or another measure such as abundance of the

species; (4) the model is used to predict the variable of interest across the study region (and perhaps for a future or past climate) (Hijmans and Elith 2015).

Currently, the statistical understanding of applied scientists is increasing and new techniques can cope with larger, more complex data sets, so applied statisticians are faced with the need to specify sophisticated models. Logically, as the complexity of these models increase, it becomes harder to perform inference. The Bayesian approach is particularly appropriate as it is flexible and can deal with complex models, for instance including hierarchical structure or including missing data. Undoubtedly, the most popular family of approximate inference methods in Bayesian statistics is the class of Markov Chain Monte Carlo (MCMC) methods. These methods, which exploded into popularity in the mid-1980s have remained at the forefront of Bayesian statistics ever since, with the basic framework being extended to cope with increasingly more complex problems (Simpson et al. 2011).

Modelling patterns of the presence/absence of the species using local environmental factors have been a growing problem in Ecology in the last few years (Chakraborty et al. 2010). This kind of modelling has been extensively used to address several issues, including the identification of essential fauna habitats in order to classify and manage conservation areas (Pressey et al. 2007), and predicting the response of species to environmental features (Midgley and Thuiller 2007; Loarie et al. 2008). Different approaches and methodologies have been proposed in this perspective (see for instance Guisan

✉ Oscar Rodríguez de Rivera  
osroderi@alumni.uv.es

<sup>1</sup> Departament d'Estadística i I.O., Universitat de Valencia, Dr. Moliner 50, 46100 Burjassot, Spain

<sup>2</sup> MRC Centre for Environment and Health, Department of Epidemiology and Biostatistics, Imperial College London, London, UK

<sup>3</sup> Departament Sistemas y Recursos Naturales, ETSI de Montes, Forestal y del Medio Natural, Universidad Politécnica de Madrid, Ciudad Universitaria s/n, Madrid, Spain

and Thuiller 2005; Hijman and Graham 2006; Wisz et al. 2008), with Maximum Entropy modelling (MaxEnt) (Elith and Burgman 2002), more flexible models as generalised linear and additive ones (GLM and GAM) (Guisan et al. 2002), species envelope models such as BIOCLIM (Busby 1991) and the multivariate adaptive regression splines (MARS) (Leathwick et al. 2005) being some of the most commonly used (Muñoz et al. 2013).

Several projects have focused on comparing these different methods (see for instance Rivera and López-Quílez 2017); also a summary this comparison has been developed recently by Lecours 2017. Most of these applications consist of explanatory models that seek to assess the relationship between environmental variables (e.g. precipitation, bathymetry, etc.) (Guisan et al. 2002). Moreover, the theory of these methods is based on the fact that the observations are independent, while spatial autocorrelation is common in georeferenced ecological data (Crase et al. 2012). Spatial autocorrelation should be taken into account in the species distribution models, even if the data were collected through a standardised sampling scheme, since the observations are often close and subject to similar environmental features (Underwood 1981; Hurlbert 1984). In addition observer error (Royle et al. 2007; Cressie et al. 2009), gaps in the sampling, missing data, and spatial mobility of the species (Gelfand et al. 2006) can also affect the models.

The value of both reptiles and amphibians has been recognised as an integral part of natural ecosystems and as heralds of environmental quality (Gibbons and Stangel 1999). In recent years, as overall environmental awareness among the public has increased, concerns have raised on the ecological state of reptile and amphibian species as well as of their habitats (Gibbons et al. 2000). Habitats of many amphibians populations are small, temporary ponds and the surrounding forested area, which are usually affected by many stressors such as UV-radiation (Cummins 2003; Hatch and Blaustein 2003), the use of pesticides (Gendron et al. 2006; Fellers et al. 2004), industrial chemicals (Bishop and Gendron 1998; Sower et al. 2000) and climate change (Corn 2005). Since amphibians are sensitive to the alterations of their environment, they could be used as bioindicator organisms to follow changes in their habitats and in ecotoxicological studies (Henry 2000). As their population usually contains high numbers of individuals and they are good representatives of freshwater environments, they are ideal model organisms for pollution studies (Burger and Snodgrass 1998).

Gibbons et al. (2000), consider the vulnerability of reptiles within the context of the factors known or suspected to be associated with amphibian declines, using six categories of concern established by Partners in Amphibian and Reptile Conservation (PARC; Gibbons and Stangel 1999): habitat loss and degradation, introduced invasive species, environmental pollution, disease and parasitism, unsustainable use and global climate change.

The aim of this paper is to build a spatial model to predict the spatial distribution of several species characterised by a low level of presences, which leads to data sparsity. We will use real data on five species of amphibians obtained from inventories developed in Las Tablas de Daimiel National Park (TDNP-Spain) in 2011–2012 supported with environmental variables. On these species, we have presence/absence at geographical coordinates and we will generate distribution models for each species as well as combine these into the corresponding the class (Amphibia). Our approach is to specify a Bayesian hierarchical geostatistical modelling framework accounting for spatial dependency.

Hierarchical models can simplify complex interactions by allowing parameters to vary at more than one level via the introduction of random effects. The expected value of the response is then expressed conditional on these random effects (Cosandey-Godin et al. 2015). The advantages of using hierarchical Bayesian models emerge more so as complexity increases, when, for example, spatio-temporal variability needs to be modelled explicitly (Cressie et al. 2009). The Bayesian framework also offers the advantage of providing full inference, such that model parameters and uncertainty can be quantified, which has great utility in applied conservation (Wade 2000; Wintle et al. 2003).

Several authors have used a Bayesian approach to analyse species distribution. For instance Golding and Purse 2016, compared Gaussian processes against more traditional techniques obtaining better performance, while Gelfand et al. 2006, tried to illustrate spatial patterns applying hierarchical logistic regression through a Bayesian framework. Also, Latimer et al. 2006, developed Bayesian regression models for species presence/absence and Royle 2004, estimated abundance of birds applying N-mixture model in a Bayesian perspective. Another interesting work developed by Mackenzie et al. 2002, focused on understanding site occupancy of some amphibians species when detection probabilities are below 1, again in a Bayesian approach.

Hierarchical Bayesian models have traditionally relied on MCMC simulation techniques, which are computationally expensive and technically challenging, consequently limiting their use. However, a new statistical approach is now readily available, namely integrated nested Laplace approximations (INLA) via the R-INLA package (<http://www.r-inla.org>) (Cosandey-Godin et al. 2015). INLA methodology and its powerful application to modelling complex datasets has recently been introduced to a wider non-technical audience (Illian et al. 2013). As opposed to MCMC simulations, INLA uses an approximation for inference and hence avoids the intense computational demands, convergence, and mixing problems sometimes encountered by MCMC algorithms (Rue and Martino 2007). It can only be used for Gaussian models but this includes the class of models which we consider here for species distribution. Moreover, R-INLA can be compiled

with the stochastic partial differential equations (SPDE) approach (Lindgren et al. 2011) which through a discretisation of a continuous Gaussian field can cope efficiently with variables characterised by a complex spatial structure. This is the case of this environmental inventory, since environmentalists or field workers start the inventory to target particular species, resulting in clustered spatial patterns and large regions without any values. Together, these new statistical methods and their implementation in R (R Core Team 2016) allow scientists to fit complex spatio-temporal models considerably faster and more reliably (Rue et al. 2009).

The structure of the paper is as follows. In Section 2, we introduce our motivating problem regarding spatial distribution for amphibian species in Las Tablas de Daimiel National Park (TDNP) (Ciudad Real, Spain). Then, after discussing the available data, we describe the geostatistical spatial model. In Section 3, we explain the model evaluation and the comparison. In Section 4, we present the results of the analysis of the spatial distributions and show how the environmental variables could affect the presence of the species. Finally, in Section 5, we resume the conclusions of this work.

## Motivating example

The data set come from an inventory developed in Las Tablas de Daimiel National Park (TDNP) during 2011 and 2012, comprising 234 sample points with coordinates. Each sample point has the presence or absence of each species, elevation in meters and information about the ambient (categorical variable with the following categories: Salt marsh, Reed bed, Islands, areas of *Typha latifolia*, *Cladium mariscus* and free of vegetation).

The following species are included: *Bufo bufo*, *Bufo calamita*, *Pelobates cultripipes*, *Pelodytes punctatus* and *Triturus pygmaeus* (see Table 1). As the aim of this analysis is not to study the distribution of each species, we are not going to explain the biology or characteristics of each species.

The Tablas de Daimiel National Park is a floodplain wetland located in the Upper Guadiana Basin, central Spain (see

Fig. 1). The landscape of Las Tablas de Daimiel is characterised by the horizontality of the terrain, with a range of altitude between the 599 m above sea level in the confluence of the rivers Guadiana and Cigüela, and the 623 m above sea level in the Pochela hill.

The Guadiana River is one of the three main drainage units of the Iberian Peninsula, having its source in central Spain before flowing into Portugal and then, in its lower reaches, acting as a natural border between the two countries. The TDNP is one of the core areas of the Mancha Húmeda, declared a Biosphere Reserve in 1980 by UNESCO. The wetland is the result of the mixture of inputs from Cigüela and Guadiana rivers, together with groundwater discharge from the West Mancha aquifer. The peripheral surface of the wetland is 1928 ha, but at present, the potentially flooded area is 1587 ha (Sánchez-Carrillo et al. 2010). The climate of the Upper Guadiana Basin is semiarid, with an average temperature of 14.1 °C (1955–2009) and an average precipitation of 448 mm (1945–2009) (Yustres et al. 2013). The spatio-temporal distribution of rain is irregular (Acreman et al. 2000), and the high temperature values in the summer cause the potential evapotranspiration to be notably high. Subject to this semiarid climate regime, the water balance is particularly fragile in the river basin, with water shortage considered a structural characteristic of the system (Cots et al. 2007).

Currently, the TDNP suffers a reduction in water inputs mainly because the groundwater discharge to the wetland is decreasing, since the aquifer has been suffering intensive groundwater overexploitation since the late 1970s (Navarro et al. 2011).

The water quality has been also affected, since one of its peculiarities is the variety of hydrochemical and hydrological processes, i.e. the two main sources of water and salts and at the same time TDNP is an area where saline sulfate-rich surface waters flowing from the Cigüela River mix with calcium carbonate freshwater coming from groundwater (Coronado et al. 1974; Sánchez-Ramos et al. 2015).

We can see in Fig. 2 the distribution of the ambients in our data set, summarised as follows: Salt marsh (2% of the samples), Reed bed (31%), Islands (17%), areas of *Typha latifolia* (34%), *Cladium mariscus* (4%) and free of vegetation (13%).

**Table 1** Species presence (in percentage) by ambient (C.m. = *Cladium mariscus*, F.v. = Free of vegetation, I. = Islands, R.b. = Reed bed, S.m. = Salt marsh and T.l. = *Typha latifolia*) and by number of presences

	C.m.	F.v.	I.	R.b.	S.m.	T.l.	Presences
<i>Bufo bufo</i>	0%	0%	0%	100%	0%	0%	3%
<i>Bufo calamita</i>	5%	5%	48%	19%	0%	24%	54%
<i>Pelobates cultripipes</i>	0%	0%	0%	100%	0%	0%	3%
<i>Pelodytes punctatus</i>	0%	21%	14%	50%	0%	14%	36%
<i>Triturus pygmaeus</i>	0%	0%	0%	50%	0%	50%	5%
Amphibia	3%	10%	31%	36%	0%	21%	100%

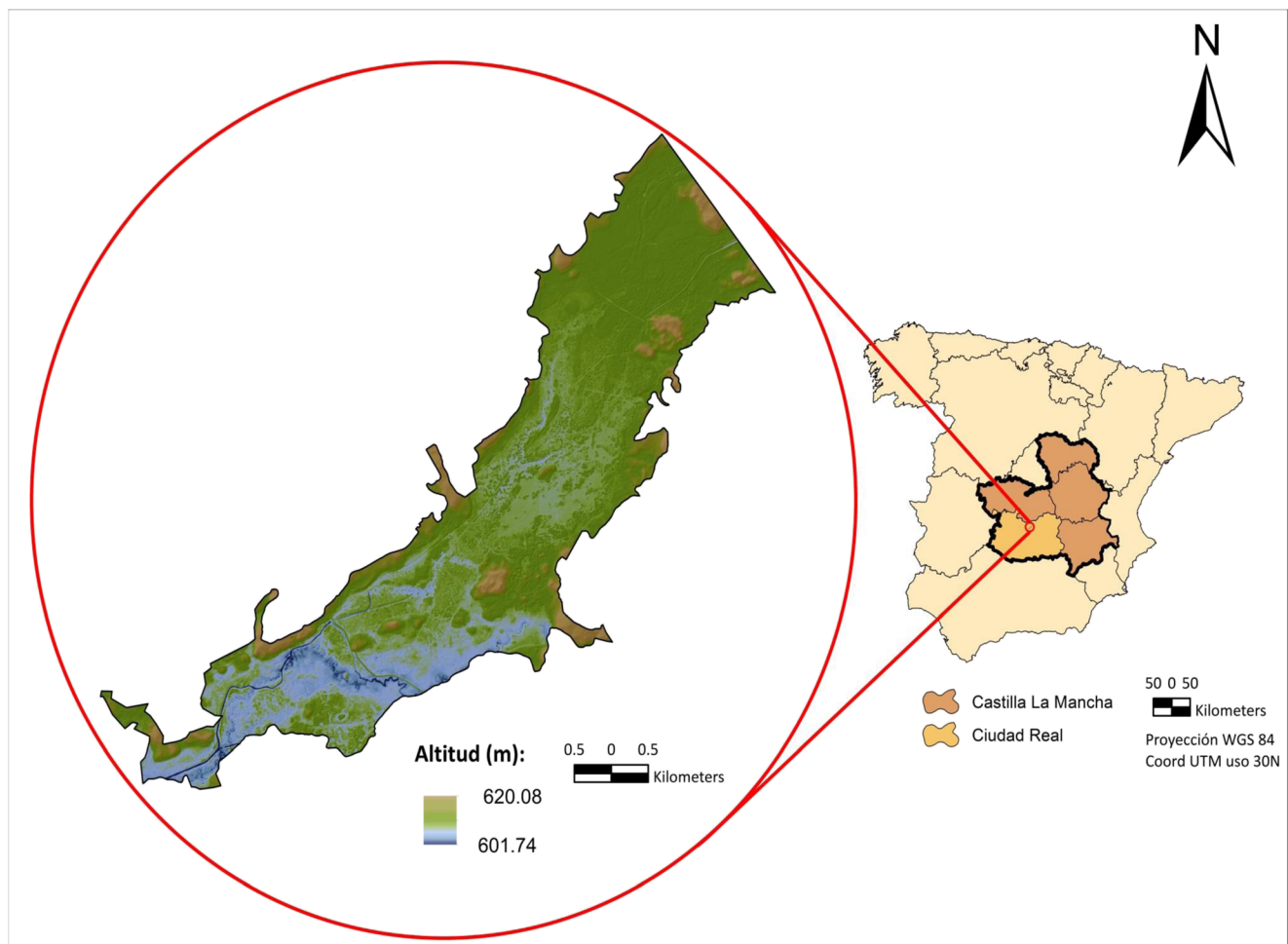
## Species distribution model

### Model

Spatial data are defined as realisations of a stochastic process indexed by space:

$$Y(s) \equiv \{y(s), s \in D\}$$

where  $D$  is a (fixed) subset of  $R^d$  (here we consider  $d = 2$ ). The actual data can be then represented by a collection of



**Fig. 1** Map of the situation of the Las Tablas de Daimiel National Park

observations  $y = \{y(s_1), \dots, y(s_n)\}$ , where the set  $(s_1, \dots, s_n)$  indicates the spatial units at which the measurements are taken. Depending on  $D$  being a continuous surface or a countable collection of  $d$ -dimensional spatial units, the problem can be specified as a spatially continuous or discrete random process, respectively (Gelfand et al. 2010).

In our case, we can consider a collection of data points with presence/absence obtained from the inventory; the sampled points are the set  $(s_1, \dots, s_n)$  of  $n$  points;  $y_s$  is the presence of each specie in each point and it is specified as

$$y_s \sim \text{Bernoulli}(\pi_s)$$

where  $\pi_s$  is the probability of the species being present.

Then, on the  $\text{logit}(\pi_s)$ , a linear model is specified including covariates  $\mathbf{x}_1$  (elevation),  $\mathbf{x}_2$  (ambient) and a spatial field  $\xi_s$

$$\text{logit}(\pi_s) = b_s + x_{1s}\beta + x_{2s}\gamma + \xi_s$$

where a discretely indexed spatial random process (see Lindgren et al. 2011) is included to approximate the continuous process:

$$\xi_s = \sum \varphi_g(s) \tilde{\xi}_g$$

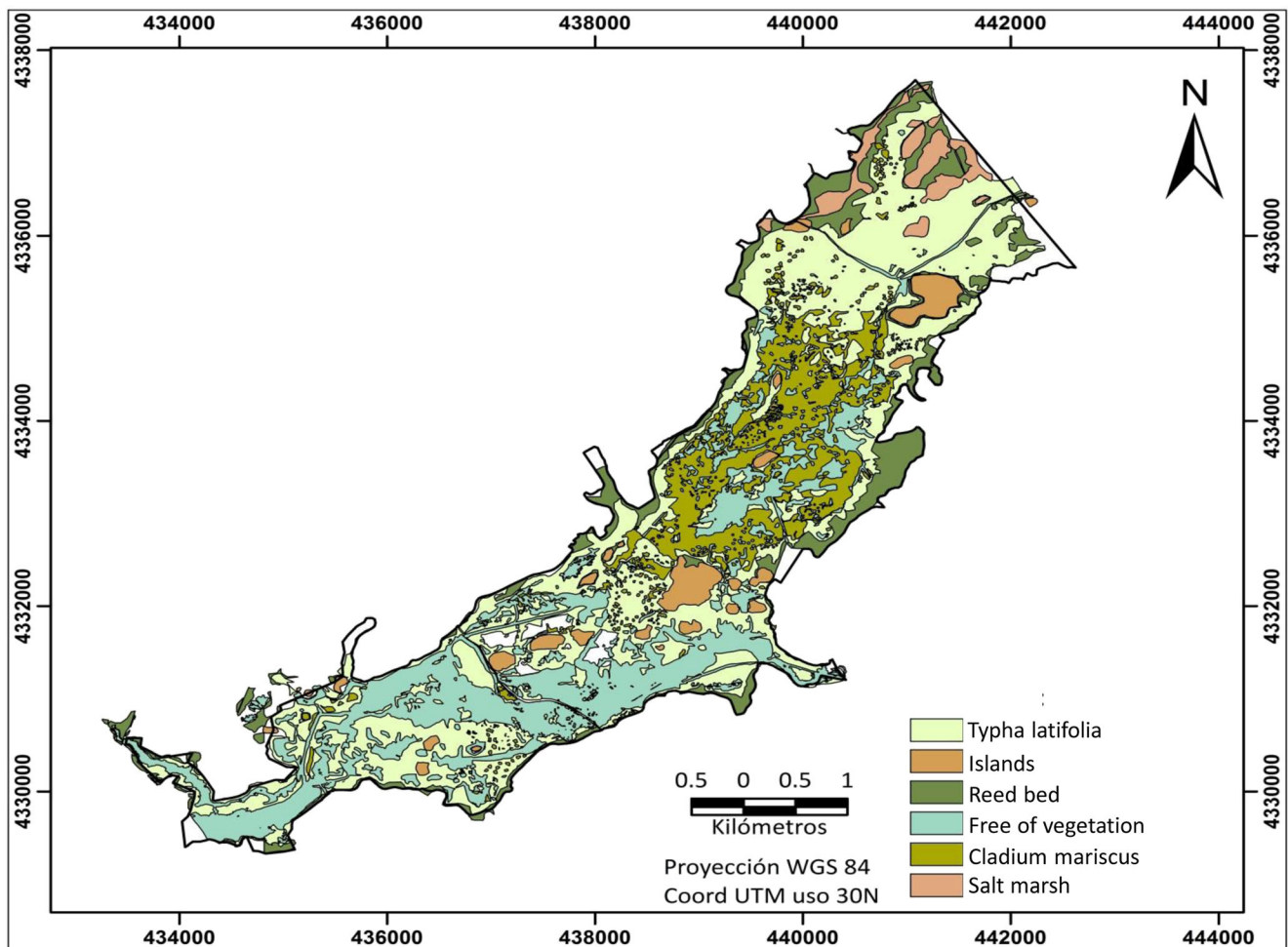
In practice, the discretisation is done dividing the study region in triangles and writing  $\xi_s$  as a linear combination of basic functions  $\varphi_g$  weighted by some zero mean terms  $\tilde{\xi}_g$  (for more details see Blangiardo and Cameletti 2015).

## Implementation, inference and evaluation

The statistical inference has been carried out through the integrated nested Laplace approximation (INLA) implemented in the R-INLA within R statistical software. In R-INLA, the first step required to run the geostatistical spatial model through SPDE is the triangulation of the considered spatial domain. We use the `inla.mesh.create` specifying the spatial coordinates used for estimation. This function performs a constrained refined Delaunay triangulation for a set of spatial locations: firstly the triangle vertices are placed at the observation locations and then further vertices are added in order to satisfy triangulation quality constraints (Lindgren et al. 2011).

A natural way to estimate out-of-sample prediction error is cross-validation (see Geisser and Eddy 1979, and Vehtari and Lampinen 2002, for a Bayesian perspective), but researchers





**Fig. 2** Map of the distribution of the ambients in the Las Tablas de Daimiel National Park

have always sought alternative measures, as cross-validation requires repeated model fits and can run into trouble with sparse data (Gelman and Shalizi 2013). In a comparative perspective (e.g. to evaluate which model fits the data best), the most used index is the DIC (Spiegelhalter et al. 2002; van der

Linde 2005) which similarly to AIC consists of two components, a term that measures goodness of fit and a penalty term for increasing model complexity.

More recently, the WAIC (Watanabe 2010) has been proposed as a suitable alternative for estimating the out-of-sample

**Table 2** Posterior estimates for the models with elevation and with elevation and ambient

	Parameter	Mean	St. Dev.	2.5%	50%	97.5%
Model with elevation						
Amphibia WAIC = 21.17/LCPO = 1.759/AUC = 0.762	Intercept	−4.69	30.36	−64.30	−4.69	54.86
	Elevation	0.01	0.05	−0.10	0.01	0.10
Model with elevation and ambient						
Amphibia WAIC = 24.53/LCPO = 1.898/AUC = 0.735	Intercept	−4.70	30.51	−64.62	−4.70	55.16
	Elevation	−0.03	0.05	−0.13	−0.03	0.08
	Ambient.Reedbed	0.30	6.34	−21.97	0.75	13.04
	Ambient.Islands	0.71	8.16	−15.96	0.62	15.32
	Ambient.Freeofvegetation	0.82	6.74	−16.49	0.75	20.86
	Ambient.Cladiummariscus	−0.72	8.37	−25.06	0.69	7.13
	Ambient.Saltmarsh	−1.17	13.09	−33.29	−2.19	45.36

**Table 3** WAIC, and LCPO and AUC comparison in species models

		E			E&A		
		WAIC	LCPO	AUC	WAIC	LCPO	AUC
Amphibia	<i>Bufo bufo</i>	1.83	1.965	0.726	13.03	2.260	0.712
	<i>Bufo calamita</i>	14.04	1.521	0.783	13.03	1.531	0.752
	<i>Pelobates cultripes</i>	1.103	1.894	0.83	3.27	2.265	0.758
	<i>Pelodytes punctatus</i>	10.79	2.523	0.693	5.43	3.914	0.656
	<i>Triturus pygmaeus</i>	4.61	2.327	0.725	3.38	2.986	0.718

expectation in a fully Bayesian approach. This approach starts with the computed log pointwise posterior predictive density and then adds a correction for the effective number of parameters to adjust for overfitting (Gelman and Shalizi 2013). WAIC operates on predictive probability density of observed variables rather than on model parameter; hence, it can be applied in singular statistical models (i.e. models with non-identifiable parameterization, see Li et al. 2015).

We have also calculated the conditional predictive ordinate (CPO) (Pettit 1990) to evaluate model assessment. The conditional predictive ordinate (CPO) is based on leave-one-out-cross-validation. CPO estimates the probability of observing a value after having already observed the others. The mean logarithmic score (LCPO) was calculated as a measure of the predictive quality of the model (Gneiting and Raftery 2007; Roos and Held 2011). High LCPO values suggest possible outliers, high-leverage and influential observations.

Finally, we have used an AUC (Area Under operating Curve score) approach to calculate the predictive accuracy of each method by comparing the validation data with the predicted presence value. AUC represents a commonly used and adequately performing measure of predictive accuracy (Huang and Ling 2005) and works by calculating the relative numbers of correctly and incorrectly identified predictions across all possible classification threshold values of the binomial response, with an AUC value equal to or below 0.5 indicating a predictive ability equal to random expectation and 1 a perfect predictive ability (Qiao et al. 2015).

## Results

Table 2 presents the main results of the analyses for the Amphibia, characterised by more data sparsity (a fewer presences). As we can see, the first model obtained using only the elevation has a better fit (WAIC = 21.17/LCPO = 1.759/AUC = 0.762) than the model with elevation and ambient (WAIC = 24.53/LCPO = 1.898/AUC = 0.735). However, based on LCPO, the model without ambient has fewer outliers. Also, we have compared performance of the different models based in AUC, these analysis shows similar results than LCPO, obtaining better values in models without ambient.

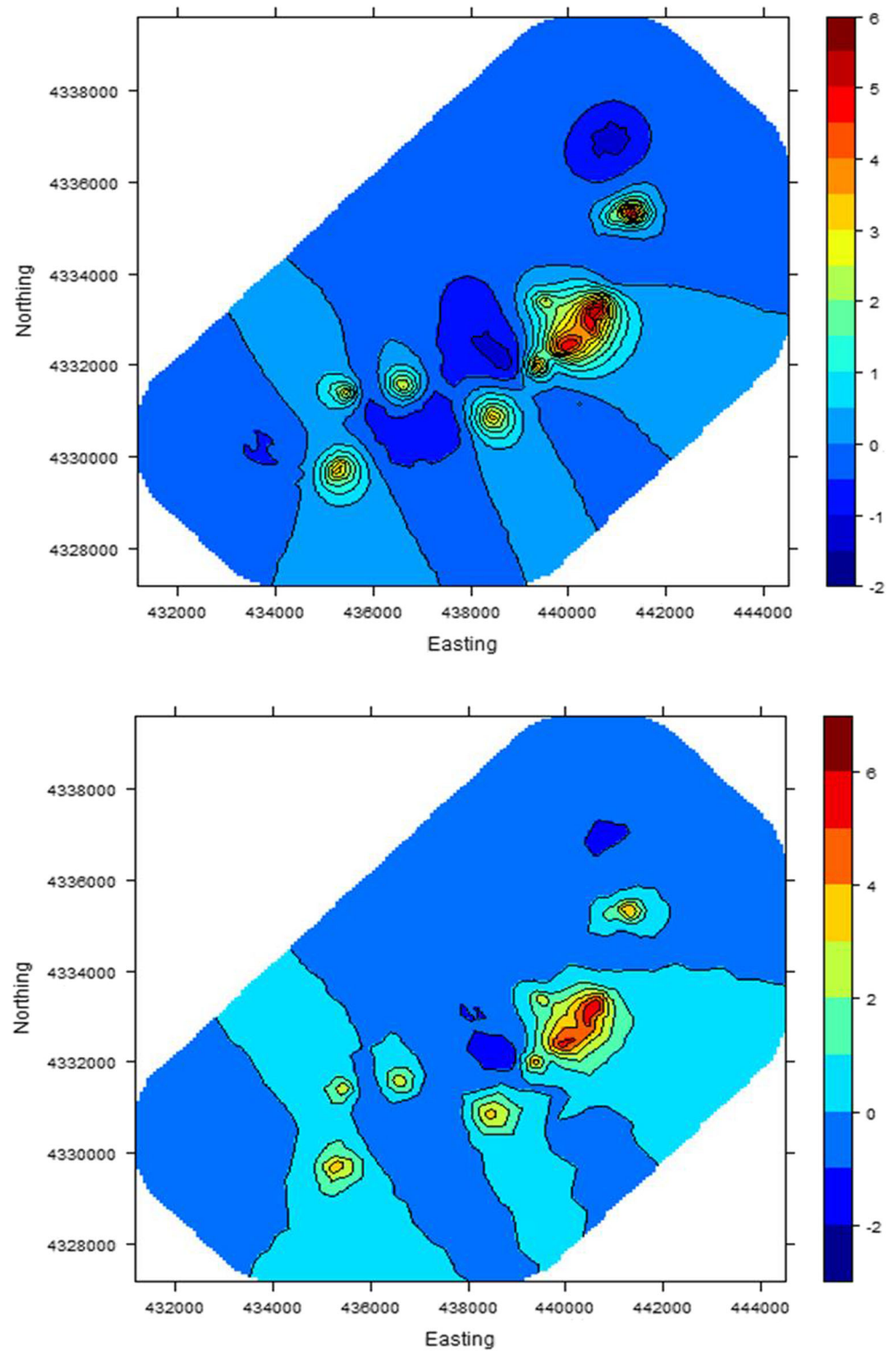
In Table 3, we can see the summary of the WAIC, LCPO and AUC values obtained in the different models for each species (model E with only Elevation and model E&A with Elevation and Ambient); this shows that, looking at the WAIC, most of the species have a better fit for the model with vegetation (except *Bufo bufo* and *Pelobates cultripes*), but looking at LCPO values Ambient seems to increase the number of outliers.. Also, looking at AUC values, models with ambient have lower predictability.

From an ecological perspective, amphibian as group (Table 2) does not have any preference between the different ambients analysed, with small negative point estimates, but with a credibility intervals including zero. Analysing the different species (Table 4), we can summarise as follows: *Bufo calamita* has all the variables have point estimates positives and excluding zero except areas with *Cladium mariscus* where the point estimate includes zero; *Pelodytes punctatus* and *Triturus*

**Table 4** Posterior estimates for the best models for the different species

	<i>Bufo calamita</i>		<i>Pelodytes punctatus</i>		<i>Triturus pygmaeus</i>		<i>Bufo bufo</i>		<i>Pelobates cultripes</i>	
	Mean	St.Dev.	Mean	St.Dev.	Mean	St.Dev.	Mean	St.Dev.	Mean	St.Dev.
Intercept	− 0.0034	11.9523	0.06	11.95	0.01	11.95	− 4.06E-04	5.20E-03	1.84E-05	0.011
Elevation	0.0001	0.0001	− 7.36E-05	6.27E-05	− 8.49E-06	1.32E-05	3.57E-06	9.62E-06	9.87E-06	1.93E-05
Ambient.Reedbed	1.39	14.07	0.45	1.72	− 6.08	23.91	178.91	24.36	143.15	15.72
Ambient.Islands	6.06	16.21	− 0.05	1.53	2.99	21.48	− 8.94	28.69	29.20	20.88
Ambient.Freeofvegetation	5.52	20.68	− 1.96	2.99	− 4.05	11.35	− 10.08	29.06	25.71	21.21
Ambient.Cladiummariscus	− 5.24	13.76	0.19	2.87	1.50	9.54	− 2.57	30.52	13.99	22.98
Ambient.Saltmarsh	7.62	15.22	− 1.14	1.41	− 8.64	30.52	− 1.36	31.13	7.00	24.64

**Fig. 3** Maps of posterior mean for the models with only Elevation as covariate (top); Elevation and Ambient (bottom)



*pygmaeus* have a different relationship with the variables, with all the point estimates including zero; however, *Bufo bufo* has predictability for areas with Reed bed as is the only variable with point estimate positive and the zero is not included; finally, *Pelobates cultripes* has better relationship with Reed bed, Islands and Free of vegetation areas.

In Fig. 3, we can see the maps of posterior mean for class Amphibia of models developed based on the covariates used. In both cases, we can see that the distribution are complementary: due to the nature of amphibians and reptiles, there are no points with both classes present at the same time. On the bottom, the map shows the distribution of the model with

elevation and ambient. As we can see, the maps obtained with elevation and with elevation and ambient are very similar. Also, the model with only elevation as covariate has more details than the model with Ambient.

## Conclusions

Hierarchical models are commonly used in ecology (Clark 2005; Cressie et al. 2009). The hierarchical modelling framework has been useful when implementing process models that include ecological theory or when modelling the data collection process (Hooten and Wikle 2008; Hooten et al. 2007, Wikle 2003). The hierarchical modelling framework has also been useful and commonly implemented for count and presence/absence data (MacKenzie et al. 2006; Royle and Dorazio 2008), but has not been readily used for presence-only data (Dorazio 2014; Fithian et al. 2015). However, we can affirm that absences are giving important information to define species distribution. Also, the hierarchical species distribution modelling approach can be readily extended to include multiple species, as we have done with amphibians, and possibly interacting species (Hui 2016; Ovaskainen and Soininen 2011; Warton et al. 2015). The spatio-temporal Poisson point process model is currently used for the analysis of species movement data captured using telemetry devices (Brost et al. 2015; Johnson et al. 2013; Russell et al. 2016). Data models developed for telemetry data will have a similar use for species distribution models (Brost et al. 2015). The methodology used to account for repeated measurements (i.e., locations) of the same individual(s) developed for telemetry data will have analogous use for species distribution models that are used to model count, presence-absence, and presence-only data that includes multiple observations of the same individuals (Hefley and Hooten 2016).

In this work, we have specified a Bayesian spatial model for studying species distribution. We have evaluated the inclusion of two variables (elevation and ambient).

The main advantage of the Bayesian model formulation is the computational ease in model fit and prediction compared to classical geostatistical methods. The main goal of this study has been to predict the occurrence of species with a relatively small number of data points, but the data was useful to show the power of this kind of process and the options of the model construction. To do so, instead of MCMC, we have used the novel integrated nested Laplace approximation approach. More precisely, we have applied the work of Lindgren et al. (2011), which provides a link between Gaussian Fields and Gaussian Markov Random Fields through the Stochastic Partial Differential Equation (SPDE) approach. The SPDE approach can be easily implemented providing results in reasonable computing time (comparing with MCMC). We showed how SPDE is as useful tool in the analysis of species

distribution. This modelling could be expanded to the spatio-temporal domain by incorporating an extra term for the temporal effect, using parametric or semiparametric constructions to reflect linear, non-linear, autoregressive or more complex behavior.

On the other hand, we have concluded some interesting points from an ecological perspective. Amphibians are easy species to model: due to their dependence to the water, most of the species live in freshwater aquatic ecosystems. The relationship between the amphibians and the water is stronger than the relationship with the ambient, probably because most of the amphibians use the water as ambient for the reproductive habits. At the same time, vegetation and environment are less important than the elevation (as a distance of the water surface). Wetlands are essential breeding habitats for many amphibian species. Pond-breeding amphibians require aquatic habitats for breeding, and embryonic and larval development, whereas terrestrial habitats are used for foraging and aestivation, and as migration and dispersal routes. Both aquatic and terrestrial habitats are used for hibernation. Hence, pond-breeding amphibians can be susceptible to changes in the availability and quality of both local-scale (aquatic habitats) and landscape-scale habitat characteristics (Piha et al. 2007). Currently, habitat loss is considered one of the greatest threats to the world's amphibian species, one-third of which are threatened (Stuart et al. 2004). Pond-breeding amphibians may be particularly influenced by the loss and increased isolation of important habitat types caused by agricultural intensification in our case due to the water lost from the agricultural ponds.

We understand that extending this framework to situations characterised by environmental changes, there is the possibility to experience climatic changes between points. And this framework could benefit by the inclusion of meteorological variables. However, Las Tablas de Daimiel National Park has a really small extension with only one meteorological station (IGME 2017), so we have avoided the use of climatic features in order to not include estimated data into the model selection.

Also, we can see that hierarchical models are particularly useful when data are sparse or species are similar. In our case, Amphibia model has different relationship with the environmental variables than the individual species included in the class model.

As we have explained about the study area, the National Park is a water area of about 1600 ha. As we are working with amphibians that develop most of their cycle in water and also, as there is no way to introduce the distance to the water due to the variation in water level, we have introduced elevation as a proxy. We assume that elevation can be affected by spatial autocorrelation. However, all methods assume spatial stationarity, i.e. spatial autocorrelation does only depend on distance between point locations, and there are very few methods to deal with non-stationarity in this context (Osborne et al. 2007).



Finally, we can conclude that due to the low level of observations, CPO is more robust than WAIC due to the presence of influential observations.

We conclude that SPDE and INLA are promising tools to work with species distribution model as they save in computational times and are easy to specify and to implement also for non-statisticians when we work with a large data set.

Summarising, R-INLA can be a complementary tool for ecologists. The major strength of R-INLA is that it allows to perform Bayesian inference, based on highly accurate approximations of posterior distributions, where models are specified using a syntax that should be familiar to R users, and where data are formatted in a straightforward way with relatively few lines of code. The straightforward model syntax and data format could help remove barriers to the adoption of N-mixture models for biologists. The substantial decrease in computation time should also facilitate the use of a wider variety of model and variable selection techniques (e.g. cross-validation and model averaging) that are not commonly used in an MCMC context due to practical issues related to computing time (Kery and Schaub 2011).

Limitations of R-INLA are mostly related to the more restricted set of N-mixture models that can be specified. R-INLA does not handle site survey covariates, employs only Poisson-Binomial and Negative Binomial-Binomial mixtures, and handles random effects (exchangeable, spatially and temporally structured) for  $p$  only. In cases where site survey covariates are particularly significant and not otherwise controlled in the sampling design, R-INLA will not be the suitable tool (Meehan et al. 2017).

## References

- Acreman M, Almagro J, Alvarez J, Bouraoui F, Bradford R, Bromley J, Croke B, Crooks S, Cruces J, Dolz J, Dunbar M, Estrela T, Fernandez-Carrasco P, Fornes J, Gustard G, Haverkamp R, De La Hera A, Hernández-Mora N, Llamas R, Martínez CL, Papamatorakis J, Ragab R, Sánchez M, Vardavas I, Webb T (2000) Groundwater and river resources programme on a European scale (GRAPES). Technical report to the European Union ENV4-CT95-0186. Institute of Hydrology, Wallingford
- Bishop CA, Gendron AD (1998) Reptiles and amphibians: shy and sensitive vertebrates of the Great Lakes basin and St. Lawrence river. *Environ Monit Assess* 53:225–244
- Blangiardo M, Cameletti M (2015) Spatial and spatio-temporal Bayesian models with R-INLA. John Wiley & Sons, Chichester
- Brost BM, Hooten MB, Hanks EM, Small RJ (2015) Animal movement constraints improve resource selection inference in the presence of telemetry error. *Ecology* 96:2590–2597
- Burger J, Snodgrass J (1998) Heavy metals in bullfrog (*Rana catesbeiana*) tadpoles: effects of depuration before analysis. *Environ Toxicol Chem* 17:2203–2209
- Busby JR (1991) BIOCLIM: a bioclimatic analysis and predictive system. In: Margules C, Austin M (eds) *Nature conservation: cost effective biological surveys and data analysis*. CSIRO, Canberra, pp 64–68
- Chakraborty A, Gelfand AE, Wilson AM, Latimer AM, Silander JA (2010) Modeling large scale species abundance with latent spatial processes. *Ann Appl Stat* 4(3):1403–1429
- Clark JS (2005) Why environmental scientists are becoming Bayesians. *Ecol Lett* 8(1):2–14
- Corn SP (2005) Climate change and amphibians. *Anim Biodivers Conserv* 28:59–67
- Coronado R, Del Portillo F, Sáez-Royuela R (1974) *Tablas de Daimiel National Park Guide*. ICONA, Madrid
- Cosandey-Godin A, Teixeira Krainski E, Worm B, Mills Flemming J (2015) Applying Bayesian spatiotemporal models to fisheries bycatch in the Canadian Arctic. *Can J Fish Aquat Sci* 72:1–12
- Cots F, David Tabara J, Werners S, McEvoy D (2007) Climate change and water adaptive management through transboundary cooperation. The case of the Guadiana river basin. Paper presented to the first International Conference on Adaptive and Integrative Water Management (CAIWA), Basel, Switzerland, November 2007
- Crase B, Liedloff AC, Wintle BA (2012) A new method for dealing with residual spatial autocorrelation in species distribution models. *Ecography* 35(10):879–888
- Cressie N, Calder CA, Clark JS, Hoef JMV, Wikle CK (2009) Accounting for uncertainty in ecological analysis: the strengths and limitations of hierarchical statistical modeling. *Ecol Appl* 19:553–5701
- Cummins CP (2003) UV-B radiation, climate change and frogs—the importance of phenology. *Ann Zool Fenn* 40:61–67
- Dorazio RM (2014) Accounting for imperfect detection and survey bias in statistical analysis of presence-only data. *Glob Ecol Biogeogr* 23(12):1472–1484
- Elith J, Burgman MA (2002) Predictions and their validation: rare plants in the Central Highlands, Victoria, Australia. In: Scott JM, Heglund PJ, Morrison ML, Raphael MG, Wall WA, Samson FB (eds) *Predicting Species Occurrences: Issues of Accuracy and Scale*. Island Press, Covelo, pp 303–314
- Elith J, Leathwick JR (2009) Species distribution models: ecological explanation and prediction across space and time. *Annu Rev Ecol Evol Syst* 40:677–697
- Fellers GM, McConnell LL, Pratt D, Datta S (2004) Pesticides in mountain yellow legged frogs (*Rana muscosa*) from the Sierra Nevada Mountains of California, USA. *Environ Toxicol Chem* 23:2170–2177
- Fithian W, Elith J, Hastie T, Keith DA (2015) Bias correction in species distribution models: pooling survey and collection data for multiple species. *Methods Ecol Evol* 6(4):424–438
- Geisser S, Eddy W (1979) A predictive approach to model selection. *J Am Stat Assoc* 74:153–160
- Gelfand AE, Silander JA, Wu SJ, Latimer AM, Rebelo PLAG, Holder M (2006) Explaining species distribution patterns through hierarchical modeling. *Bayesian Anal* 1(1):41–92
- Gelfand AE, Diggle P, Fuentes M, Guttorp P (eds) (2010) *Handbook of spatial statistics*. Chapman & Hall, Boca-Raton
- Gelman A, Shalizi C (2013) Philosophy and the practice of Bayesian statistics (with discussion). *Br J Math Stat Psychol* 66:8–80
- Gendron AD, Marcogliese DJ, Barbeau S, Christin MS, Brousseau P, Ruby S, Cyr D, Fournier M (2006) Exposure of leopard frogs to a pesticide mixture affects life history characteristics of the lungworm *Rhabdias ranae*. *Oecologia* 135:469–476
- Gibbons JW, Stangel PW (eds) (1999) *Conserving amphibians and reptiles in the new millenium*. Proceedings of the Partners in Amphibian and Reptile Conservation (PARC). Conference; 2–4 June 1999; Atlanta (GA). Aiken (SC): Savannah River Ecology Laboratory. *Herp Outreach Publication* #2
- Gibbons JW, Scott DE, Ryan TJ, Buhlmann KA, Tuberville TD, Metts BS, Greene JL, Mills T, Leiden Y, Poppy S, Winne CT (2000) The global decline of reptiles, Déjà vu amphibians. *BioScience* 50(8):653–666
- Gneiting T, Raftery AE (2007) Strictly proper scoring rules, prediction, and estimation. *J Am Stat Assoc* 102(477):359–378

- Golding N, Purse BV (2016) Fast and flexible Bayesian species distribution modelling using Gaussian processes. *Methods Ecol Evol* 7(5):598–608
- Guisan A, Thuiller W (2005) Predicting species distribution: offering more than simple habitat models. *Ecol Lett* 8:993–1009
- Guisan A, Edwards TC, Hastie T (2002) Generalized linear and generalized additive models in studies of species distributions: setting the scene. *Ecol Model* 157:89–100
- Hatch AC, Blaustein AR (2003) Combined effects of UV-B radiation and nitrate fertilizer on larval amphibians. *Ecol Appl* 13:1083–1093
- Hefley TJ, Hooten MB (2016) Hierarchical species distribution models. *Curr Landscape Ecol Rep* 1(2):87–97
- Henry PFP (2000) Aspects of amphibian anatomy and physiology. In: Sparling DW, Linder G, Bishop CA (eds) *Ecotoxicology of amphibians and reptiles*. SETAC Press, Pensacola, pp 71–110
- Hijman R, Graham C (2006) The ability of climate envelope models to predict the effect of climate change on species distributions. *Glob Chang Biol* 12(12):2272–2281
- Hijmans RJ, Elith J (2015) Species distribution modelling with R. <http://cran.r-project.org/web/packages/dismo/vignettes/sdm.pdf>, The R foundation for statistical computing
- Hooten MB, Wikle CK (2008) A hierarchical Bayesian non-linear spatio-temporal model for the spread of invasive species with application to the Eurasian collared-dove. *Environ Ecol Stat* 15(1):59–70
- Hooten MB, Wikle CK, Dorazio RM, Royle JA (2007) Hierarchical spatiotemporal matrix models for characterizing invasions. *Biometrics* 63(2):558–567
- Huang J, Ling CX (2005) Using AUC and accuracy in evaluating learning algorithms. *IEEE Trans Knowl Data Eng* 17(3):299–310
- Hui FK (2016) Boral–Bayesian ordination and regression analysis of multivariate abundance data in R. *Methods Ecol Evol* 7(5):744–750
- Hurlbert SH (1984) Pseudoreplication and the design of ecological field experiments. *Ecol Monogr* 54:187–211
- IGME (Instituto Geológico y Minero de España) (2017) [http://www.igme.es/zonas\\_humedas/daimiel/medio\\_fisico/clima.htm](http://www.igme.es/zonas_humedas/daimiel/medio_fisico/clima.htm). Date visited: 01/03/2017
- Illian JB, Martino S, Sørbye SH, Gallego-Fernández JB, Zunzunegui M, Esquivias MP, Travis JMJ (2013) Fitting complex ecological point process models with integrated nested Laplace approximation. *Methods Ecol Evol* 4:305–315. <https://doi.org/10.1111/2041-210X.12017>
- Johnson DS, Hooten MB, Kuhn CE (2013) Estimating animal resource selection from telemetry data using point process models. *J Anim Ecol* 82(6):1155–1164
- Kéry M, Schaub M (2011) Bayesian population analysis using WinBUGS: a hierarchical perspective. Academic Press, Burlington
- Latimer AM, Wu SS, Gelfand AE, Silander JA (2006) Building statistical models to analyze species distributions. *Ecol Appl* 16(1):33–50
- Leathwick JR, Rowe D, Richardson J, Elith J, Hastie T (2005) Using multivariate adaptive regression splines to predict the distributions of New Zealand's freshwater diadromous fish. *Freshw Biol* 50: 2034–2052
- Lecours V (2017) On the use of maps and models in conservation and resource management (warning: results may vary). *Front Mar Sci* 4:288
- Li L, Qiu S, Zhang B, Feng CX (2015) Approximating cross-validated predictive evaluation in Bayesian latent variable model with integrated IS and WAIC. *Stat Comput* 26:881–897. <https://doi.org/10.1007/s11222-015-9577-2>
- Lindgren F, Rue H, Lindström J (2011) An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach [with discussion]. *J R Stat Soc B* 73(4):423–498
- Loarie SR, Carter BE, Hayhoe K, McMahon S, Moe R, Knight CA, Ackerly DD (2008) Climate change and the future of Californian endemic flora. *PLoS One* 3(6):e2502
- MacKenzie DI, Nichols JD, Lachman GB, Droege S, Royle JA, Langtimm CA (2002) Estimating site occupancy rates when detection probabilities are less than one. *Ecology* 83:2248–2255
- MacKenzie DI, Nichols JD, Royle JA, Pollock KH, Bailey LL, Hines JE (2006) Occupancy estimation and modeling inferring patterns and dynamics of species occurrence. Academic Press, Burlington
- Meehan TD, Michel NL, Rue H (2017) Estimating animal abundance with N-mixture models using the R-INLA package for R arXiv preprint arXiv:1705.01581
- Midgley GF, Thuiller W (2007) Potential vulnerability of Namaqualand plant diversity to anthropogenic climate change. *J Arid Environ* 70:615–628
- Muñoz F, Pennino MG, Conesa D, López-Quílez A, Bellido JM (2013) Estimation and prediction of the spatial occurrence of fish species using Bayesian latent Gaussian models. *Stoch Environ Res Risk Assess* 27:1171–1180
- Navarro V, García B, Sánchez D, Asensio L (2011) An evaluation of the application of treated sewage effluents in Las Tablas de Daimiel National Park, Central Spain. *J Hydrol* 401:53–64
- Osborne PE, Foody GM, Suárez-Seoane S (2007) Non-stationarity and local approaches to modelling the distributions of wildlife. *Divers Distrib* 13(3):313–323
- Ovaskainen O, Soininen J (2011) Making more out of sparse data: hierarchical modeling of species communities. *Ecology* 92(2):289–295
- Pettit LI (1990) The conditional predictive ordinate for the normal distribution. *J R Stat Soc Ser B* 52(1):175–184
- Piha H, Luoto M, Merilä J (2007) Amphibian occurrence is influenced by current and historic landscape characteristics. *Ecol Appl* 17(8):2298–2309
- Pressey RL, Cabeza M, Watts EM, Cowling RM, Wilson KA (2007) Conservation planning in a changing world. *Trends Ecol Evol* 22: 583–592
- Qiao H, Soberón J, Peterson AT (2015) No silver bullets in correlative ecological niche modelling: insights from testing among many potential algorithms for niche estimation. *Methods Ecol Evol* 6(10): 1126–1136. <https://doi.org/10.1111/2041-210X.12397>
- R Core Team (2016) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. <http://www.R-project.org/>
- Rivera ORD, López-Quílez A (2017) Development and comparison of species distribution models for forest inventories. *ISPRS Int J Geo-Inf* 6(6):176
- Roos M, Held L (2011) Sensitivity analysis in Bayesian generalized linear mixed models for binary data. *Bayesian Anal* 6(2):259–278
- Royle JA (2004) N-mixture models for estimating population size from spatially replicated counts. *Biometrics* 60:108–115
- Royle JA, Dorazio RM (2008) Hierarchical modeling and inference in ecology: the analysis of data from populations, metapopulations and communities. Elsevier, Amsterdam
- Royle JA, Kéry M, Gautier R, Schmidt H (2007) Hierarchical spatial models of abundance and occurrence from imperfect survey data. *Ecol Monogr* 77:465–481
- Rue H, Martino S (2007) Approximate Bayesian inference for hierarchical Gaussian Markov random field models. *J Stat Plan Inference* 137(10):3177–3192
- Rue H, Martino S, Chopin N (2009) Approximate Bayesian inference for latent Gaussian models using integrated nested Laplace approximations (with discussion). *J R Stat Soc Ser B* 71:319–392
- Russell JC, Hanks EM, Haran M (2016) Dynamic models of animal movement with spatial point process interactions. *J Agric Biol Environ Stat* 21(1):22–40
- Sánchez-Carrillo S, Angeler D G, Álvarez-Cobelas M, Sánchez-Andrés R (2010) Freshwater wetland eutrophication. In *Eutrophication: causes, consequences and control*. Springer, Dordrecht, pp 195–210
- Sánchez-Ramos D, Sánchez-Emeterio G, Florín Beltrán M (2015) Changes in water quality of treated sewage effluents by their receiving environments in Tablas de Daimiel National Park, Spain. *Environ Sci Pollut Res* 23:6082–6090. <https://doi.org/10.1007/s11356-015-4660-y>

- Simpson D, Lindgren F, Rue H (2011) Fast approximate inference with INLA: the past, the present and the future. Technical report at [arxiv.org](https://arxiv.org)
- Sower SA, Reed KL, Babbitt KJ (2000) Limb malformations and abnormal sex hormone concentrations in frogs. *Environ Health Perspect* 108:1085–1090
- Spiegelhalter DJ, Best NG, Carlin BP, van der Linde A (2002) Bayesian measures of model complexity and fit (with discussion). *J R Stat Soc Ser B* 64(4):583–616
- Stuart S, Chanson JS, Cox NA, Young BE, Rodrigues ASL, Fishman DL, Waller RW (2004) Status and trends of amphibian declines and extinctions worldwide. *Science* 306:1783–1786
- Underwood AJ (1981) Techniques of analysis of variance in marine biology and ecology. *Oceanogr Mar Biol Annu Rev* 19:513–605
- van der Linde A (2005) DIC in variable selection. *Statistica Neerlandica* 59(1):45–56
- Vehtari A, Lampinen J (2002) Bayesian model assessment and comparison using cross validation predictive densities. *Neural Comput* 14:2439–2468
- Wade PR (2000) Bayesian methods in conservation biology. *Conserv Biol* 14(5):1308–1316. <https://doi.org/10.1046/j.1523-1739.2000.99415.x>
- Warton DI, Blanchet FG, O'Hara RB, Ovaskainen O, Taskinen S, Walker SC, Hui FK (2015) So many variables: joint modeling in community ecology. *Trends Ecol Evol* 30(12):766–779
- Watanabe S (2010) Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *J Mach Learn Res* 11:3571–3594
- Wikle CK (2003) Hierarchical Bayesian models for predicting the spread of ecological processes. *Ecology* 84(6):1382–1394
- Wintle BA, McCarthy MA, Volinsky CT, Kavanagh RP (2003) The use of Bayesian model averaging to better represent uncertainty in ecological models. *Conserv Biol* 17(6):1579–1590. <https://doi.org/10.1111/j.1523-1739.2003.00614.x>
- Wisz MS, Hijmans RJ, Li J, Peterson AT, Graham CH, Guisan A (2008) Effects of sample size on the performance of species distribution models. *Divers Distrib* 14:763–773
- Yustres A, Navarro V, Asensio L, Candel M, García B (2013) Groundwater resources in the Upper Guadiana Basin (Spain): a regional modelling analysis. *Hydrogeol J* 21(5):1129–1146