Modelling species presence-absence in the ecological niche theory framework using shape-constrained generalized additive models

L. Citores[1,2], L. Ibaibarriaga[1], D.-J. Lee[2], M.J. Brewer[3], M. Santos[4], and G. Chust[1]

[2]*Basque Center for Applied Mathematics, Alameda de Mazarredo, 14, 48009 Bilbao, Spain*

[1]*AZTI, Txatxarramendi ugartea z/g, 48395 Sukarrieta, Spain*

[3]*Biomathematics and Statistics Scotland, Craigiebuckler, Aberdeen, AB15 8QH,UK*

[4]*AZTI, Herrera kaia - Portualdea z/g, 20110 Pasaia, Spain*

Abstract

According to ecological niche theory, species response curves are unimodal with respect to environmental gradients. A variety of statistical methods have been developed for species distribution modelling. A general problem with most of these habitat modelling approaches is that the estimated response curves can display biologically implausible shapes which do not respect ecological niche theory. This work proposes using shape-constrained generalized additive models (SC-GAMs) to build species distribution models under the ecological niche theory framework, imposing concavity constraints in the linear predictor scale. Based on a simulation study and a real data application, we compared performance with respect to other regression models without shape-constraints (such as standard GLMs and GAMs with varying degrees of freedom) and also to models based on so-called “Plateau” climate-envelopes. The imposition of concavity for response curves resulted in a good balance between the goodness of fit (GOF) and agreement with ecological niche theory. The approach has been applied to fit distribution models for three fish species given several environmental variables.

Keywords: Ecological niche theory; GAMs; GLMs; Plateau method; shape-constrained GAMs; Species distribution models

Preprint accepted in Ecological Modelling, December 2019

Highlights

- Shape-constrained GAMs are proposed for niche-based species distribution modelling.
- Methods are tested and compared via Monte Carlo simulation and using real data.
- Results show good balance between GOF and agreement with ecological niche theory.
- Shape-restricted and non-restricted explanatory variables can be combined.
- Spawning habitats of 3 pelagic species are modelled as a multivariate case study.

1 Introduction

Species distribution models (SDM) relate species occurrence or abundance with information on environmental conditions and spatial characteristics of locations where the species was found (Elith and Leathwick, 2009). These models can be used to predict or to have a better understanding of the species distribution (Halvorsen, 2012; Petitpierre et al., 2017). They are widely used in several fields, such as ecology, evolutionary biology and conservation (Guisan et al., 2013; Peterson et al., 2011; Zimmermann et al., 2009). In recent years there has been an increased interest in understanding how future environmental changes may impact species distribution; the most highly cited papers have focused on developing novel methods to better predict environmental suitability for species and to improve model performance (Barbosa and Schneck, 2015).

A variety of statistical methods have been applied to species distribution modelling (e.g. Guisan and Zimmermann, 2000; Merow et al., 2014) such as regression-based models (Guisan et al., 2002; Hastie and Tibshirani, 1990), environmental envelopes (BIOCLIM, Busby (1991), Cerdeira et al. (2018)), mechanistic approaches (CLIMEX, Kriticos et al. (2015)), neural networks (Pearson et al., 2002) and maximum entropy models (MAXENT, Phillips et al. (2006)). However, most of these methods can result in species responses along environmental gradients that are convex or multimodal, and consequently not ecologically meaningful or otherwise difficult to interpret (see below for further discussion).

It has been often claimed that species distribution models need a stronger theoretical background (see Austin, 2002; Elith and Leathwick, 2009; Jiménez-Valverde et al., 2008, for a detailed review). Recently, several authors have attempted to clarify the relationship between species distribution models and the concept of ecological niche (Kearney, 2006; Peterson et al., 2011; Pulliam, 2000; Soberon and Nakamura, 2009). Although the debate is still open (Halvorsen, 2012), it is agreed that the resulting statistical model should be ecologically plausible (Elith and Leathwick, 2009). According to ecological niche theory, species distributions should provide unimodal relationships with respect to environmental gradients (Hutchinson, 1957). When environmental conditions become less favourable, various stages of the life cycle (feeding, growth and reproduction) are affected, resulting in lower presence of the species (Austin, 1987; Helaouet and Beaugrand, 2009). Hutchinson (1957) defined the niche as an “n-dimensional hypervolume”, where the dimensions are environmental states within which a species is able to survive. Hutchinson (1957) also distinguished "fundamental" from "realized" niches, to define the conditions under which species could survive and those where they actually live, respectively: (1) the fundamental niche is determined by the physiological range of tolerance of the species to environmental factors in the absence of biotic interactions (e.g. competition, predation or parasitism), and (2) the realized niche is the part of the fundamental niche actually occupied by the species, given factors such as the presence of competitors/predators and dispersal limitations of the species (Soberón and Arroyo-Peña, 2017). As a result, the realized niche tends to be smaller than the fundamental niche (Soberón and Arroyo-Peña, 2017). Although the fundamental niche should be unimodal, the realized niche can be bimodal when the centre of the niche gradient is affected by interspecific competition or when the species is not occupying the most suitable habitat due to dispersal limitation (Austin, 2002). However, scarce species data and the heterogeneous distribution of species occurrence along gradients are the most problematic situation leading to multimodal and ecologically non-meaningful relationships with environmental variables. For instance, data on occurrence of a fish species which spawns in two river mouths, separated by a latitudinal distance, can easily lead to a bimodal distribution along a temperature gradient.

The concept of niche has evolved after the 80s and incorporates the impacts of the organism on environmental factors (Chase and Leibold, 2003) to better explain competition and species coexistence (Pocheville, 2015). To the pragmatic purpose of modelling species distribution, this can include several types of variables, as well as those defining the niche namely, direct variables, resource variables, and indirect variables (Austin and Smith, 1990; Guisan and Zimmermann, 2000; Huston, 1994). Direct variables are those environmental factors having a direct physiological impact on the species but are not consumed, typical examples being pH affecting plant growth or temperature affecting fish growth. Indirect variables do not have a direct physiological impact, but might be highly correlated with the species through the combination of related factors effects. For example, elevation can affect species presence through the combined effect of atmospheric pressure, temperature and UV radiation, and have ecophysiological implications. Resource variables refer to limiting factors (i.e. essential resources consumed by the species, such as food and oxygen) and biotic interactions (competition, predation or mutualism). The first two types (direct and indirect) of variables are within the group of variables that do not interact dynamically with the species and hence are not affected by species abundance. These were termed "scenopoetic" variables by Hutchinson (1978). In contrast, resource variables interact with the species and are affected by species presence and abundance. In the context of species distribution models, several authors (Austin, 1980, 2007; Austin and Smith, 1990) have discussed the shape of response curves and how this depends on the variable type. While there is no theoretical expectation regarding the shape with respect to indirect variables, they advocated that the fundamental niche as a function of direct variables should be unimodal (symmetric or not), and for limiting factors should be logistic or Michaelis-Menten saturation curves. SDMs based on non-scenopoetic variables might require more elaborate mathematical methods to include species interaction (Peterson et al., 2011). Thus, species distribution models need to combine environmental variables that are expected to meet the ecological niche theory with other explanatory variables having no shape restrictions.

Commonly used methods to build species distribution models in the ecological niche theory framework include regression-based methods, such as Generalized Linear Models (GLMs) and Generalized Additive Models (GAMs). They have been well-documented, both theoretically and empirically (Coudun and Gegout, 2006; Guisan et al., 2002; Lehmann et al., 2002; Scott et al., 2002). Generalized Linear Models (Guisan et al., 2002; McCullagh and Nelder, 1989) are widely used in statistical ecology as a simple parametric technique that may allow symmetric bell-shaped ecological response curves (Coudun and Gegout, 2006; Jamil and Ter Braak, 2013). However, this can be too restrictive as often non-symmetric responses have been observed (Austin, 2007; Huisman et al., 1993). Generalized Additive Models are also very popular as semi-parametric and more flexible regression-like approaches (Austin, 2002; Heikkinen and Makipaa, 2010). Pedersen et al. (2019) proposed an extension of GAMS called hierarchical GAMs (HGAMs) to model intergroup variability in ecology; these models allow smooth functions to vary between groups and can be used to test if the smooth functions are common across groups. In general, GAMs and related extensions allow flexible non-symmetric shapes, but they can result in implausible response curves, contrary to the ecological niche theory framework. Current practice tends to use low degree smoothing functions, such as splines with a low number of knots, in order to obtain response curves in agreement with niche theory (Chust et al., 2014). However, restrictions on the number of knots and/or the degrees of freedom (by altering the smoothing parameter within GAMs, say) do not guarantee this aim, and a visual evaluation of resulting fitted curves is still required.

Other attempts to build species distribution models under ecological niche theory include Beta functions (Minchin, 1987) and Huisman-Olff-Fresco (HOF) curves (Huisman et al., 1993), fitting unimodal and monotonic response curves with or without symmetry. A simulation study by Oksanen and Minchin (2002) concluded that HOF curves obtained better results than Beta functions and Gaussian response models which provided biased or inappropriate models. However, they are only allowed for single-variable analysis. Alternatively, the “Plateau” method proposed by Brewer et al. (2016) is an environmental envelope model based on a concave piece-wise polynomial function. While providing an ecologically meaningful method (unimodal even if not symmetric), this approach can be easily extended to multiple environmental variables accounting for potential interactions between the climatic variables.

Shape-constrained generalized additive models (or simply SC-GAMs, Pya and Wood, 2014) are based on the same statistical framework as GLMs and GAMs regression methods, but they allow us to incorporate monotonicity and concavity shape-constraints in the component functions of the linear predictor of the GAMs. Imposing concavity constraints should be an effective alternative to fitting non-symmetric parametric response curves, while retaining the unimodality constraint, required by ecological niche theory, for direct variables and limiting factors. Recently, several successful applications of shape-constrained models to incorporate prior knowledge about the shape of the response curve along variables of interest have been found related to animal activity, pollution mortality, tree height-diameter relationships or petroleum engineering (Guevara et al., 2018; Hofner et al., 2016; Schmidt et al., 2018).

The objective of this work is to assess the performance of SC-GAMs in fitting species distribution models under the ecological niche theory in comparison with other approaches. We considered two different implementations of SC-GAMs: the maximum likelihood implementation from the *scam* R R Core Team (2018) package (Pya, 2018); and the component-wise boosting approach from the *mboost* R package (Hothorn et al., 2018). First, we conducted a simulation study to assess performance in terms of goodness-of-fit and agreement with ecological niche theory—comparisons with respect to GLMs, GAMs with different degrees of smoothness, and the "Plateau" method. All methods were evaluated within a real case study, modelling the probability of presence of sardine eggs in the Bay of Biscay as a function of sea surface temperature. Secondly, SC-GAMs were used to model egg distribution at the spawning of three pelagic species as a function of several environmental gradients, combining direct and indirect variables, and accounting for model selection and validation.

2 Methods

2.1 Regression models for presence-absence data

We considered six different approaches for fitting species distribution models. In all of them species presence-absence data were modelled as a function of an environmental variable x. Let Y be the response variable coming from a Binomial distribution with probability of presence $p(x)$. The *logit* transformation of $p(x)$ is a function of the environmental variable x (presented here using a single explanatory variable for simplicity):

$$\log\left(\frac{p(x)}{1-p(x)}\right) = g(x). \tag{1}$$

The simplest model is a binary logistic generalized linear model (GLM, McCullagh and Nelder 1989,Oaksenen et al. 2001; Ter Braak and Looman 1986) where the linear predictor is a second order polynomial of the environmental variable:

$$g(x) = \beta_0 + \beta_1 x + \beta_2 x^2. \tag{2}$$

For $\beta_2 < 0$, this results in a unimodal and symmetric relationship between the species response and the environmental variable.

Generalized Additive Models (GAMs, Hastie and Tibshirani 1990; Wood 2017) are a generalization of GLMs, where the linear predictor is a smooth function of the explanatory variable. In a binary regression model with logit link, we have the form:

$$g(x) = \beta_0 + f(x), \tag{3}$$

where $f(x)$ is a smooth function. There are several ways to represent $f(x)$, from kernel smoothing or local linear methods to splines-based regression methods. We describe the latter approaches, where $f(x)$ is given by a sum of some basis functions. Hence for a single covariate x, we have:

$$f(x) = \sum_{k=1}^{K} \theta_k B_k(x), \tag{4}$$

where θ_k are the regression coefficients and $B_k(x)$ a basis function of x. There are several choices for the basis functions (e.g. polynomials of a certain order, natural splines, cubic splines or B-splines). Splines are flexible tools for smoothing in general. A spline of degree d is a function formed by connecting polynomial segments of degree d so that the function is continuous, the function has $d-1$ continuous derivatives, and the dth derivative is constant between knots. B-splines (de Boor, 1972) are a popular choice given that they are easy to compute and they have good numerical properties. In regression splines, estimated regression coefficients, $\hat{\theta}_j$, are obtained by least squares (i.e. by minimizing the residual sum of squares) and hence the shape of a spline can be controlled by carefully choosing the number of knots and their exact locations in order to allow flexibility (e.g. fix the locations of k knots at quantiles of x), and avoid overfitting where the trend changes little. However, in many situations, choosing the number of knots and their locations is a very difficult problem to solve.

Alternatively, smoothing splines find the solution of f which minimizes:

$$\sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda \int f''(x)^2 dx,$$

where the minimizer $f(x)$ is a natural cubic spline, with knots at each sample point $x_1, ..., x_n$, and $\lambda \int f''(x)^2 dx$ is the roughness penalty. The parameter λ, controls the amount of smoothness and takes values $0 < \lambda < \infty$, for $\lambda \to \infty$; large values of λ result in strong penalisation (a straight line in the limit) and for values λ close to 0 the resulting fit is a wiggly function. The selection of λ can be performed by (generalized) cross-validation or information criteria such as Akaike or Bayesian information criteria (Akaike, 1974; Schwarz et al., 1978). However, the main drawback of smoothing splines is the dimensionality for large n (Green and Silverman, 1993).

In contrast to smoothing splines, low-rank approximations have been proposed in the literature (see Ruppert et al., 2003, for a complete overview), which are called penalized regression splines. For instance, thin plate regression splines (Wood, 2003) are constructed by a simple transformation and truncation of the basis that arises from the solution of the thin plate spline smoothing problem. P-splines (Eilers and Marx, 1996) are also a low-rank approximation and a simpler alternative to smoothing splines. They consider moderately large B-spline basis functions of a size smaller than the observations and modify the penalty term by a discrete order difference penalty on adjacent coefficients, i.e. the difference operator acts on the regression coefficients, $\Delta\theta_j = \theta_j - \theta_{j-1}$, $\Delta^2\theta_j = \Delta(\Delta\theta_j) = \theta_j - 2\theta_{j-1} + \theta_{j-2}$ and in general $\Delta^d\theta_j = \Delta(\Delta^{d-1}\theta_j)$ (see Eilers et al., 2015, for further details).

The R package *mgcv* (Wood, 2019) is the most popular R package to fit GAMs. The use of GAMs has already been proposed in the literature on habitat modelling and ecological niche theory (Chust et al., 2014). Generally, the species response curve is not constrained to a particular shape, but instead is controlled by limiting the flexibility of the model by selecting the number of knots. GCV (Generalized Cross Validation) criterion is used for smoothing parameters estimation as default method in the used *mgcv* package.

The methods proposed in this paper, SC-GAMs, are based on generalized additive models, allowing us to impose shape-constraints on the linear predictor function. In Bollaerts et al. (2006) or Eilers (2017) an algorithm based on asymmetric penalties in an iterative procedure is proposed. A similar approach is considered in Pya and Wood (2014) using shape constraints (monotonicity, concavity/convexity or mixed-typed constraints) with B-splines on the first or second derivates of the smooth terms. The latter methods are implemented in the R package *scam* in a more general framework, e.g. including bivariate tensor product smooths (Pya, 2018).
For fitting species distribution models in agreement with ecological niche theory, we imposed concavity constraints in the linear predictor scale ($f''(x) \leq 0$) for which the condition $\theta_j \leq \theta_{j-1}$ suffices (see Pya and Wood, 2014, for further details). As proved in Annex C, this implies unimodal probability response curves. The implementation of the method allows for an automatic selection of the smoothing parameters by calling the *gam* function in the R package *mgcv*. However, we found the algorithm fails to converge in some situations. This issue is discussed in the next section.

Another method we considered is the so-called model-based boosting. Boosting is a gradient descent algorithm for optimizing general risk functions using component-wise penalized least squares for fitting GAMs (see Bühlmann and Hothorn, 2007; Hothorn et al., 2010, for further details). Boosting is a popular ensemble method in machine learning, where multiple learners (usually known as base learners) are trained to solve the same problem. In the particular case of modelling species distributions, shape constraints are implemented in the package *mboost* through the base-learner *bmono*, based on P-spline base-learners with an additional symmetric penalty in second order differences on the linear predictor scale, as in Bollaerts et al. (2006). The optimal number of boosting iterations can be achieved via cross-validated estimation of the empirical risk for hyper-parameter selection. For more technical details about theoretical aspects and software implementation, see Hothorn et al. (2018) or Hofner et al. (2014).

Finally, the “Plateau” method, proposed by Brewer et al. (2016), performs climate envelope fitting via an explicitly defined concave shape on the linear predictor scale. This shape consists of an increasing slope, a possible plateau, and a decreasing slope. In the univariate case, the envelope function is defined as a piece-wise function:

$$g(x) = \begin{cases} \alpha_1 + \beta_1 x & x \leq -\alpha_1/\beta_1 \\ \beta_0 & -\alpha_1/\beta_1 < x < \alpha_2/\beta_2 \\ \alpha_2 + \beta_2 x & x \geq \alpha_2/\beta_2 \end{cases} \tag{5}$$

where $\beta_1 > 0$, $\beta_2 < 0$ are increasing and decreasing slopes, α_1, α_1 are intercepts and β_0 is the *plateau* value.

2.2 Simulation

In order to evaluate and compare the performance of the proposed approaches for fitting species distribution models we carried out a simulation study.

First, four different theoretical response curves depending on a single environmental variable were generated within the simulation model, which are considered as the true curves for performance statistics computation. Afterwards sampling and observation errors were introduced and presence-absence data sets were generated based on the underlying theoretical probability curves. The simulated data sets were then fitted according to the proposed models. Finally, the goodness-of-fit and the concordance of the fitted model with ecological niche theory were measured by means of several performance statistics, described in section 2.2.6 below.

2.2.1 Environmental gradient

The real environmental variable used for data simulation was the sea surface temperature (SST) in the Atlantic Ocean in 1999 (Edwards et al., 2012). These data are arranged on a grid with a spatial resolution of 1x1 degrees (1489 data points) covering the region between 40° and 63° in latitude and $-70°$ and 2° in longitude. The average SST in the selected data is 9.92°C with a standard deviation of 5.35°C and minimum and maximum values of -2°C and 20.6°C respectively.

2.2.2 Species responses

Theoretical species response curves along the environmental gradient of SST (x) followed the generalized Beta function proposed by Minchin (1987):

$$b(x) = \begin{cases} \frac{P_0}{d}\left(\frac{x-m}{r}+b\right)^{\alpha}\left(1-(\frac{x-m}{r}+b\right)^{\gamma} & m-rb < x < m+r(1-b) \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

where m is the location of the optimum, P_0 is the maximum probability of occurrence at the mode, r is the range of occurrence along the gradient and α and γ are shape parameters. The additional parameters b and d depend only on α and γ and are introduced to reduce the complexity of the formula ($b = \alpha/(\alpha+\gamma)$ and $d = b^{\alpha}(1-b)^{\gamma}$). Combining different values for the shape parameters, we generated 4 distinct curves representing different plausible scenarios: a symmetric curve (denoted as *curve1*, with $\alpha = 4$, $\gamma = 4$), a platykurtic curve (denoted as *curve2*, with $\alpha = 0.1$, $\gamma = 0.4$), a left skewed curve (denoted as *curve3*, with $\alpha = 1.5$, $\gamma = 0.5$) and a right-skewed curve (denoted as *curve4*, with $\alpha = 1$, $\gamma = 4$) (Figure 1). All scenarios were generated with the same maximum probability of occurrence ($P_0 = 1$), location of optima ($m = 6.95$ °C) and range of occurrence ($r = 10$ °C).

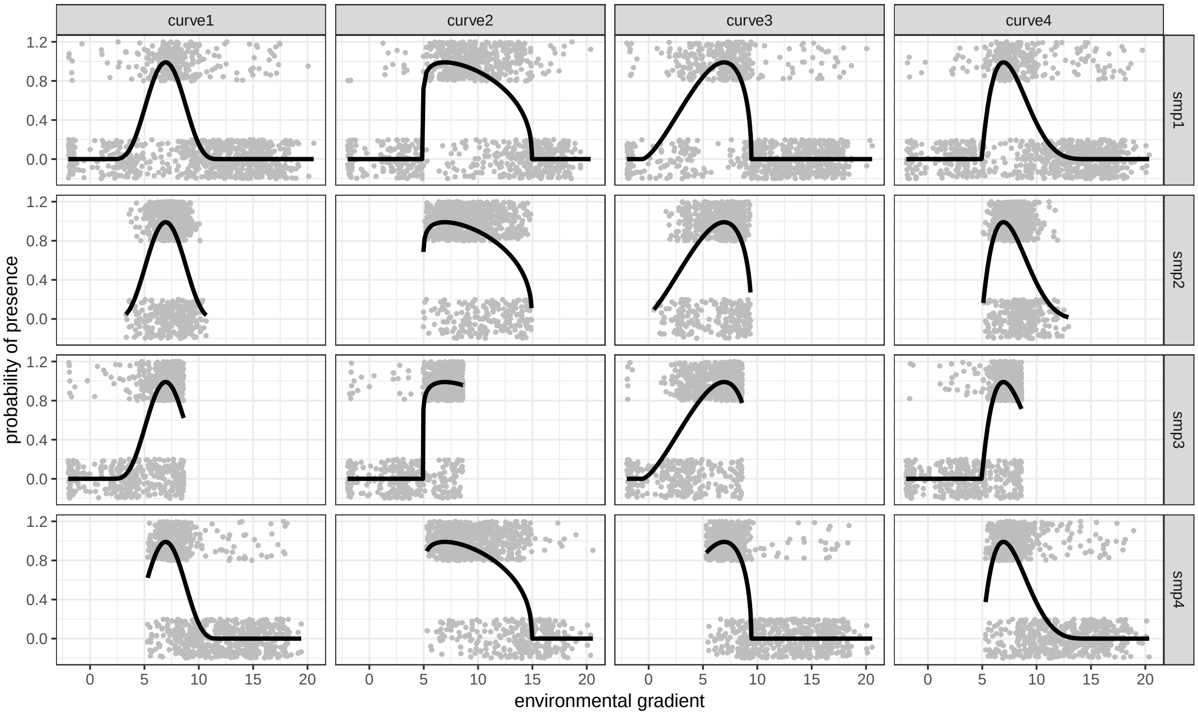

Figure 1: Columns are the true curves (*curve1*-*curve4*) and rows are the generated presence-absence data (grey dots) by sampling scenarios (*smp1*–*smp4*) for a single replicate.

2.2.3 Sampling

For each type of curve we generated a sample of 1000 observations according to four different sampling schemes. As a first sampling option (*smp1*) samples were generated randomly along the whole range of the environmental gradient with the same probability at all locations. In the second sampling scheme (*smp2*), the sampling probability is proportional to $b(x)$, so that the probability of sampling locations is higher around the theoretical response curve mode than in the tails. The last two options (*smp3* and *smp4*) account for the cases where the whole range of the environmental gradient is not observed, having a sampling probability of zero above (or below) a specific value of the gradient (see the rows of Figure 1).

2.2.4 Presence-absence data

The presence-absence data y was generated via a Bernoulli distribution with probability of occurrence $p(x)$, which is a noisy version of $b(x)$ in Eq (6). In order to mimic the effect of (unobserved or unmodelled) environmental variables other than x, $p(x)$ was draw from a beta inflated distribution (*BEINF*, allowing for zero and one inflation) implemented in the R package *gamlss.dist* (Rigby and Stasinopoulos, 2005; Stasinopoulos et al., 2019):

$$p(x) = \mathrm{BEINF}(b(x), \sigma, \nu, \tau), \tag{7}$$

where the theoretical occurrence probability $b(x)$ is the mean of the distribution, $\sigma = 0.1$ is the scale parameter and $\nu = \tau = 0.1$ are parameters modelling the probabilities of zero

Method	Functional form	Constraints	Basis	Penalty	R package	R function
GLM	$\beta_0 + \beta_1 x + \beta_2 x^2$	$\beta_2 < 0$	2nd order polynomial	No	`stats`	`glm`
GAMlk	$\beta_0 + \sum_{j=1}^{K} f_j(x)$	$K = 3$	tprs	Yes	`mgcv`	`gam`
GAMhk	$\beta_0 + \sum_{j=1}^{K} f_j(x)$	$K = 10$	tprs	Yes	`mgcv`	`gam`
SCAMfixSP	$\beta_0 + \sum_{j=1}^{K} f_j(x)$	$f''(x) \leq 0$	B-splines with concavity	No	`scam`	`scam`
boost	$\beta_0 + \sum_{j=1}^{K} f_j(x)$	$f''(x) \leq 0$	concavity constraint	Yes	`mboost`	`gamboost`
Plateau	See Eq. (5)	$\beta_1 > 0, \beta_2 < 0$	piece-wise parametric	No	`plateau`	`fit.glm.env`

Table 1: Summary of the six approaches considered.

and one respectively ($\nu = p_0/p_2, \tau = p_1/p_2$, where $p2 = 1 - p_0 - p_1$ and p_0 and p_1 are probabilities of zero and one respectively). for further details on *BEINF* parametrization see Stasinopoulos et al. (2019) .

2.2.5 Model fit

For each type of curve and each sampling scheme, 100 replicated data sets were generated (a total of 1600 data sets). Each generated data set, with 1000 observations each, was fitted using the proposed methods. Table 1 summarizes the six approaches considered (namely *"GLM", "GAMhk", "GAMlk", "SCAMfixSP", "boost"* and *"Plateau"*) and includes: the functional form of the model; constraints (if any); type of basis function (or base learner in the case of *boost*); penalty (yes or no); and finally the corresponding R packages and specific functions.

It is important to state some options we fixed in performing the simulations: i) for GAM methods we consider a low number of knots ($K = 3$, in *GAMlk*) and a higher number of knots ($K = 10$, in *GAMhk*), following Chust et al. (2014) for illustrative purposes; ii) for the SC-GAM's implementation in the R package *scam*, we found several convergence problems in the current implementation (*scam* version 1.2-4), and hence we decided to remove the penalty from the model by fixing the smoothing parameter (with the argument `sp`) to 10^{-4} and controlling the smoothness with a fixed number of knots in the construction of the model bases; iii) boosting is a computationally more expensive method but overcomes the convergence problems in *scam* (See Annex B for implementation details and code).

For each sampling scenario, fitted values were obtained along the corresponding sampled environmental gradient interval while predictions were computed for the whole gradient interval. Analyses were performed using the computing environment R (R Core Team, 2018).

2.2.6 Performance statistics

The goodness-of-fit of each method was evaluated in terms of the Root Mean Squared Error (RMSE):

$$RMSE = \sqrt{\frac{\sum (p - \hat{p})^2}{n}}, \tag{8}$$

where p is the real, theoretical probability, $\hat{p}$ is the estimated probability and n is the sample size ($n = 1000$ in this case).

The level of agreement with ecological niche theory was evaluated in terms of the concavity constraint. Second derivatives along the environmental gradient were approximated via finite differentiation. Negative second derivatives for the predicted curves along the whole environmental gradient indicate that the concavity restriction is respected on the linear predictor scale, while positive values at some point would indicate that a non-concave shape has been estimated. When concavity is held, we looked at the first derivatives, computing the number of changes of sign of the fitted curve, to evaluate whether the method was capable of estimating a global maximum, as defined in the theoretical curve, or not.

Uncertainty around estimated curves was compared by means of estimated variances of predicted values along the whole range of each curve. Coverage probabilities were computed as the percentage of theoretical values along the whole gradient that fell inside the estimated 90% confidence intervals in each replicate (Morris et al., 2019).

3 Results

The six modelling approaches were applied to each replicated data set for each type of curve and sampling scheme. The proposed shape-constrained GAM methods ("SCAMfixSP", "boost") as well as the "Plateau" method do satisfy the concavity restriction, resulting in unimodal response curves, and show closer estimated probabilities to the true theoretical response curve compared to the rest of the methods, as illustrated in Figure 2 for a single replicate and single scenario. In contrast, the "GAMhk" method, the most flexible option, does not fulfill the concavity restriction and neither of the "GLM" and "GAMlk" methods are capable of detecting the maximum. Estimated probabilities with these last two methods are far from the theoretical curve, mainly for unsampled environmental gradient values (Figure 2).

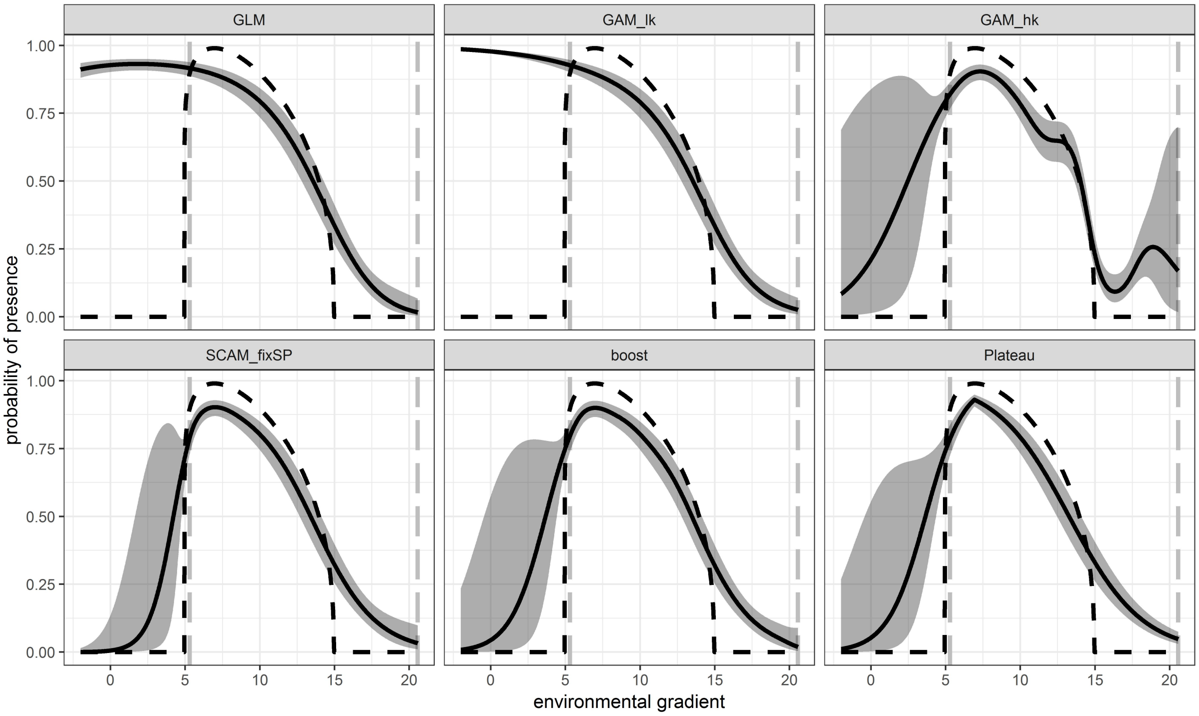

Figure 2: Predicted response curves by method for *curve2* and *smp4* for a single replicate. Dashed black lines represent the true theoretical response curves and solid lines represent obtained fitted curves with their corresponding 90% confidence intervals in gray. Vertical gray dashed lines represent the sampling range. Each panel corresponds to a particular method.

In order to summarize the performance statistics for all scenarios and methods, median and 0.1, 0.25, 0.75 and 0.9 percentile values across the 100 replicates were computed. The SC-GAM methods (“SCAMfixSP”, “boost”) and the “Plateau” method all satisfy concavity restrictions in all cases, assuring unimodal response curves are estimated in every scenario. They are able to detect a single global maxima in more than 80% of the replicates in most of the scenarios, with the “boost” method having the highest success percentages on detecting global maxima for all scenarios (Table 2). Furthermore, SC-GAMs result in better performance in terms of RMSE, giving lower values than the rest of the methods, except for the most flexible “GAMhk” method, which gives the lowest RMSE values (Figure 3). However, when using “GAMhk”, estimated curves almost never satisfy the concavity restriction (only 40 fitted curves out of 1600 simulations are concave). The “GLM” and “GAMlk” methods are able to fit concave curves only for sampling options *smp1* and *smp2* and result in worse RMSE values than the shape-constrained methods. For the rest of sampling options (*smp3* and *smp4*), these methods are not able to always fit concave curves, and when concavity does hold, global maxima are not detected in most cases (Table 2).

	curve1 smp1	curve1 smp2	curve1 smp3	curve1 smp4	curve2 smp1	curve2 smp2	curve2 smp3	curve2 smp4	curve3 smp1	curve3 smp2	curve3 smp3	curve3 smp4	curve4 smp1	curve4 smp2	curve4 smp3	curve4 smp4
GLM																
concave %	100	100	85	0	100	100	15	100	100	100	100	0	100	100	19	0
max detected %	100	100	47	0	100	100	8	100	100	100	100	0	100	100	6	0
GAM lk																
concave %	100	100	88	0	100	100	15	100	100	100	100	0	100	100	18	1
max detected %	100	100	2	0	100	57	0	12	100	100	2	0	100	74	0	0
GAM hk																
concave %	0	2	0	0	0	31	0	2	0	6	0	0	0	0	0	0
max detected %	0	2	0	0	0	20	0	1	0	6	0	0	0	0	0	0
SCAM fix SP																
concave %	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
max detected %	100	100	100	70	100	100	94	69	100	100	100	35	100	100	100	100
Boost																
concave %	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
max detected %	100	100	100	99	100	98	93	82	100	100	100	87	100	100	100	100
Plateau																
concave %	100	100	100	100	100	100	98	100	100	100	100	100	100	100	100	100
max detected %	100	100	98	92	97	44	18	68	100	93	74	78	100	89	63	87

Table 2: Percentage of replicates for each scenario and method for which estimated response curves are concave in the linear predictor scale and percentage of fitted curves that detect a single global maximum.

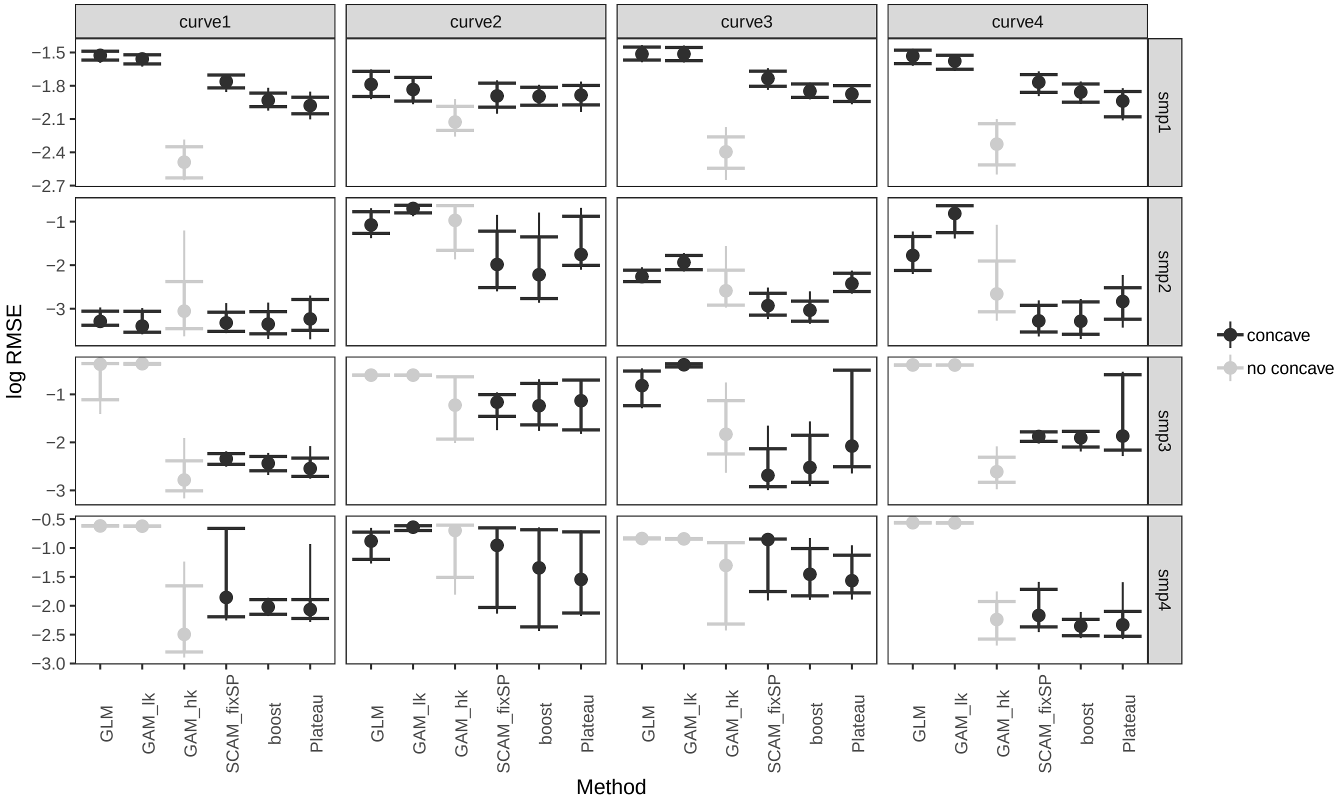

Figure 3: log of Root Mean Square Error for all curves and sampling scenarios. Points are median RMSE values across replicates and vertical thick lines represent the 75% interquantile range, while thin lines represent the 90% interquantile range. Black elements: all obtained fitted curves are concave; grey elements: not all fitted curves are concave. Each column corresponds to a curve type and each row to a sampling scenario.

Concerning uncertainty indicators, standard deviations were computed for each data-point and were used to compute 90% confidence intervals around the estimated curves. Among the methods that are able to estimate concave shapes, SC-GAM methods have higher coverage percentages (percentage of true theoretical values that fall inside these confidence intervals, see Figure 4) in comparison to "GLM" and "GAMlk" methods. The most flexible method, "GAMhk", shows the highest coverage percentages. However, we have noted that the underlying fitted curves are often not concave. Although in most scenarios, the "Plateau" method and proposed shape-constrained GAM methods show similar results (overlapping intervals), the "Plateau" method presents higher variability in results, while the "boost" method shows more stable interquantile ranges across replicates (see Figures 3 and 4). Note that obtained coverage percentages are low in all cases due to the introduced zero and one inflated error, making the estimated maximum probability lower than the theoretically fixed value ($P_0 = 1$), and estimated curve tails greater than 0 (see Figure2).

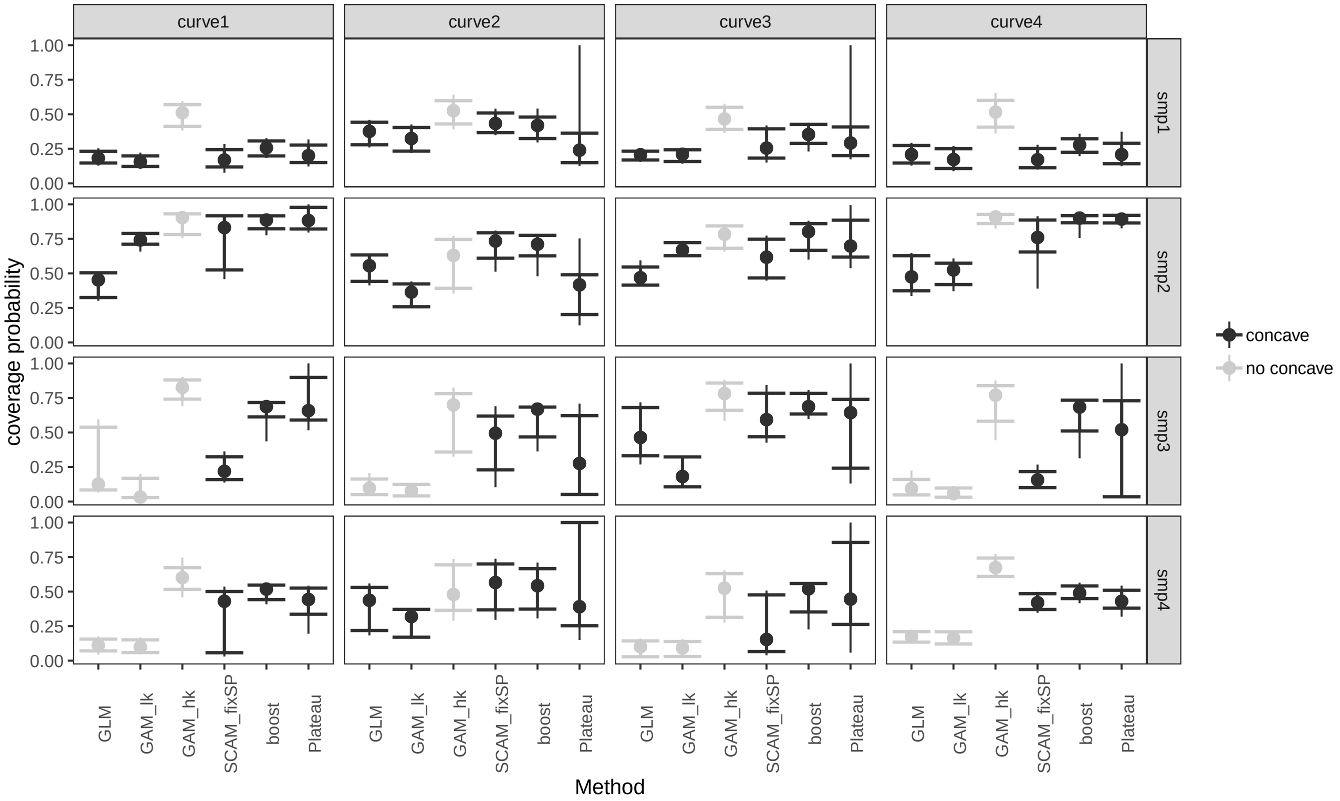

Figure 4: Coverage percentages for each scenario and method. Thick lines represent the 75% interquantile range, while thin lines represent the 90% interquantile range. Points represent median values. Black color represents that all obtained fitted curves are concave and grey color means that not all estimated curves are concave. Each column corresponds to a curve and each row to a sampling scenario.

4 Case studies

The proposed methods were also used to model the spawning habitat of some fish species in two different case studies. In the first case study, the six modelling approaches were tested and compared in an univariable analysis, modelling the occurrence of sardine (*Sardine pilchardus*) eggs in the Bay of Biscay as a function of Sea Surface Temperature (SST). In the second case study, the use of the proposed shape constrained methods was extended to more than one variable. An illustration of the use of concavity restrictions for some variables in a more complex and realistic case study is provided.

4.1 Thermal niche for sardine eggs

The European sardine (*Sardine pilchardus*) is a small pelagic fish distributed along the Northeast Atlantic and the Mediterranean Sea (Parrish et al., 1989). Several studies have attempted to identify the main environmental variables and timing that determine sardine spawning and found that temperature was an important factor (e.g. (Bernal et al., 2007; Planque et al., 2007).

We analysed the presence of sardine eggs as a function of sea surface temperature (SST) using data collected in the BIOMAN survey (Santos et al., 2018). This survey is conducted yearly in May in order to estimate the spawning stock biomass of anchovy in the Bay of Biscay by the Daily Egg Production Method (DEPM, Lasker 1985; Parker 1980). In addition, in some years the DEPM is also used to estimate the spawning stock biomass of sardine (see ICES, 2017, technical report). We compiled data from years 1999, 2002, 2008, 2014 and 2017, for which the full DEPM was applied for sardine. At each sampling location, presence-absence data of sardine eggs, geographical position (longitude and latitude), and environmental variables such as SST, were recorded. In total, 3472 data points were used for the model fitting. The presence-absence data distribution along the environmental gradient for this case study is similar to *smp2* scenarios in the simulation study, with overlapping distributions of presences and absences (see Figures A.1 and A.2 in Annex A for presence-absence data densities).

From the six proposed methods, “GLM” and “GAMhk” result in a convex and a multimodal response curve respectively, that are incompatible with the niche theory. Shape-constrained methods give concave unimodal curves which do agree with the niche theory (see Figure 5). The GAM method with fewer degrees of freedom (“GAMlk”) results in a monotone decreasing function. When predicting for temperatures lower than observed the predicted probabilities of presence continue to increase, being far from the expected bell-shaped response curve.

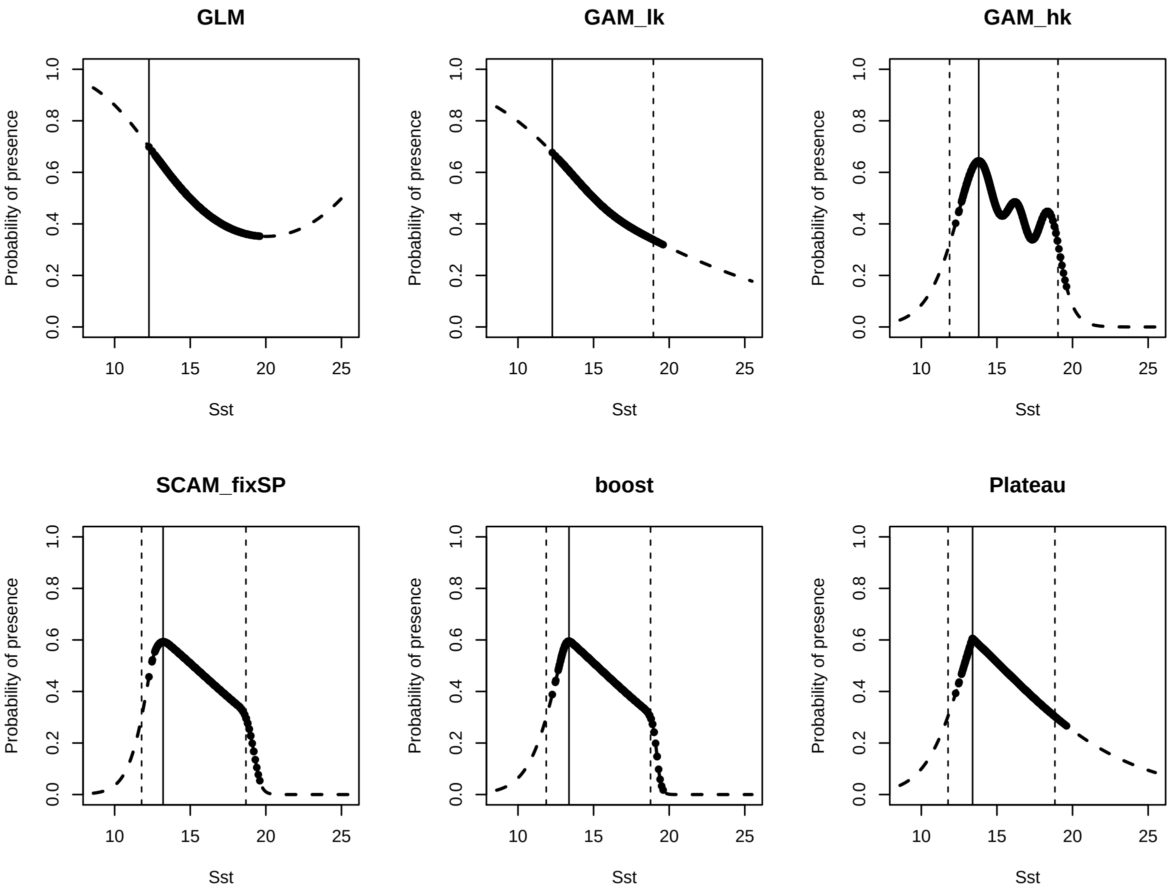

Figure 5: Sardine egg real presence-absence data (grey points), fitted response curves (in black), and predicted curves (dashed curves) along the SST environmental gradient. Vertical lines represent the optimum and dashed vertical lines tolerance limits. Each panel corresponds to a specific method.

For each method we computed the optimum temperature and the range of tolerance temperatures. The optimum was the value of the gradient with the highest estimated probability of presence and the tolerance was determined by the range of gradient where the predicted probability of species occurrence was higher than half of the maximum value for predicted probability (Schröder et al., 2005). For "GLM" and "GAMlk" the lower limit of the tolerance range could not be computed, or can be considered $-\infty$ given that the estimated curve is monotonically increasing for decreasing values of the gradient. The optimum SST is estimated around 12.5°C for these two methods while for the rest of the methods it located around 13.5°C. The obtained tolerances with shape-constrained methods and "GAMhk" methods are very similar giving a range from around 12 to 18°C.

4.2 Spawning habitat of three pelagic species

Often when fitting species distribution models, the spatio-temporal coverage of the data is limited and does not cover the range of the environmental gradient that determines the biogeographic species area (Austin, 2007). In those cases, the species response is truncated and cannot be modelled adequately. The ample coverage of the ICES triennial mackerel egg survey makes it an exception. Since 1977, the survey has been conducted every three years between January and July and covers a large area from southern Spain to the north of Scotland, with the aim of estimating the total annual egg production of the western Atlantic mackerel stock ICES (2018); Lockwood et al. (1981). The egg presence-absence and abundance data collected during the survey have been used to characterize the spawning habitat of mackerel: see Borchers et al. (1997); Bruge et al. (2016); Brunel et al. (2018). Within the framework of an EU programme (INDICES, EU Study 97/017), the samples collected during the 1998 triennial survey were reanalyzed and eggs and larvae of other fish species were quantified Ibaibarriaga et al. (2007). We applied SC-GAMs to model the egg distribution of three of these species: European anchovy (*Engraulis encrasicolus*); sardine (*Sardine pilchardus*); and Atlantic mackerel (*Scomber scombrus*). Their performance was compared with respect to the other methods considered. For each sampling location of presence-absence of eggs, we compiled environmental and depth data. Environmental data were extracted from the NCEP Global Ocean Data Assimilation System, GODAS (Derber and Rosati, 1989), which provides gridded 4D data with a monthly temporal resolution and a vertical resolution of 10m on 0.333°x1° latitude-longitude grid points of: sea surface temperature (SST), salinity (SSS), temperature at 205 m (temp205), difference between surface temperature and temperature at 205 m (temp dif), oceanic mixed layer (dbss obml). Depth data were obtained from the bathymetric database ETOPO1 from NOAA using the package *marmap* (Pante and Simon-Bouhet, 2013) in R (R Core Team, 2018) and introduced in log scale (logbathy).

We applied the "SCAMfixSP' method, which allows constructing models as a combination of shape constrained variables and non-restricted variables. Among the variables available for these case study, all of them were treated as direct variables (Austin, 2007), and therefore introduced with shape constraints, except for bathymetry, which was considered to be an indirect variable, and so introduced without shape restriction. Variable selection was based on AICc, as defined in Barton (2009), selecting for each species the model with the lowest AICc, after removing the variables that were not significant in univariable analysis. Depth was selected for all species models. Additionally, salinity, surface temperature and temp_dif were also selected for anchovy, obtaining a model fit with 61.1% of explained deviance. In the sardine model salinity and temperature at 205m were included obtaining 33.7% of explained deviance, while for mackerel salinity and temp_dif were selected for the final model with 29.97% of explained deviance. All selected variables and AICc values for each species are shown in Table 3.

All used variables except for depth were introduced in the models with the concavity restriction on the linear prediction scale, assuring this way that the ecological niche theory was met. These variables' response curves of the selected direct variables (sea surface temperature (SST), salinity (SSS), temperature at 205 m (temp205), difference between surface temperature and temperature at 205 m (temp_dif) and the oceanic mixed layer (dbss obml)) are monotonic or unimodal, presenting a single optimum at most. The optimum salinity value was estimated at 35.3 psu for anchovy and 35.5 psu for mackerel, while for sardine the whole range could not be captured, resulting in a monotonic decreasing response curve (Figure 6). The optimum along the Temp205 variable was estimated at 12.2°C for sardine. All marginal response curves for these variables and each species can be found in the supplementary material (Annex A, Figures A.3, A.4, A.5).

This proposed SC-GAM approach was also compared with other unrestricted methods for this multivariate case study. Presence-absence data for these three species with the same selected explanatory variables were also fitted using more common GAM approaches ("GAMlk" and "GAMhk"), showing that marginal response curves are not in agreement with ecological niche theory—some estimated response curves do not satisfy the unimodality condition (Figure 6).

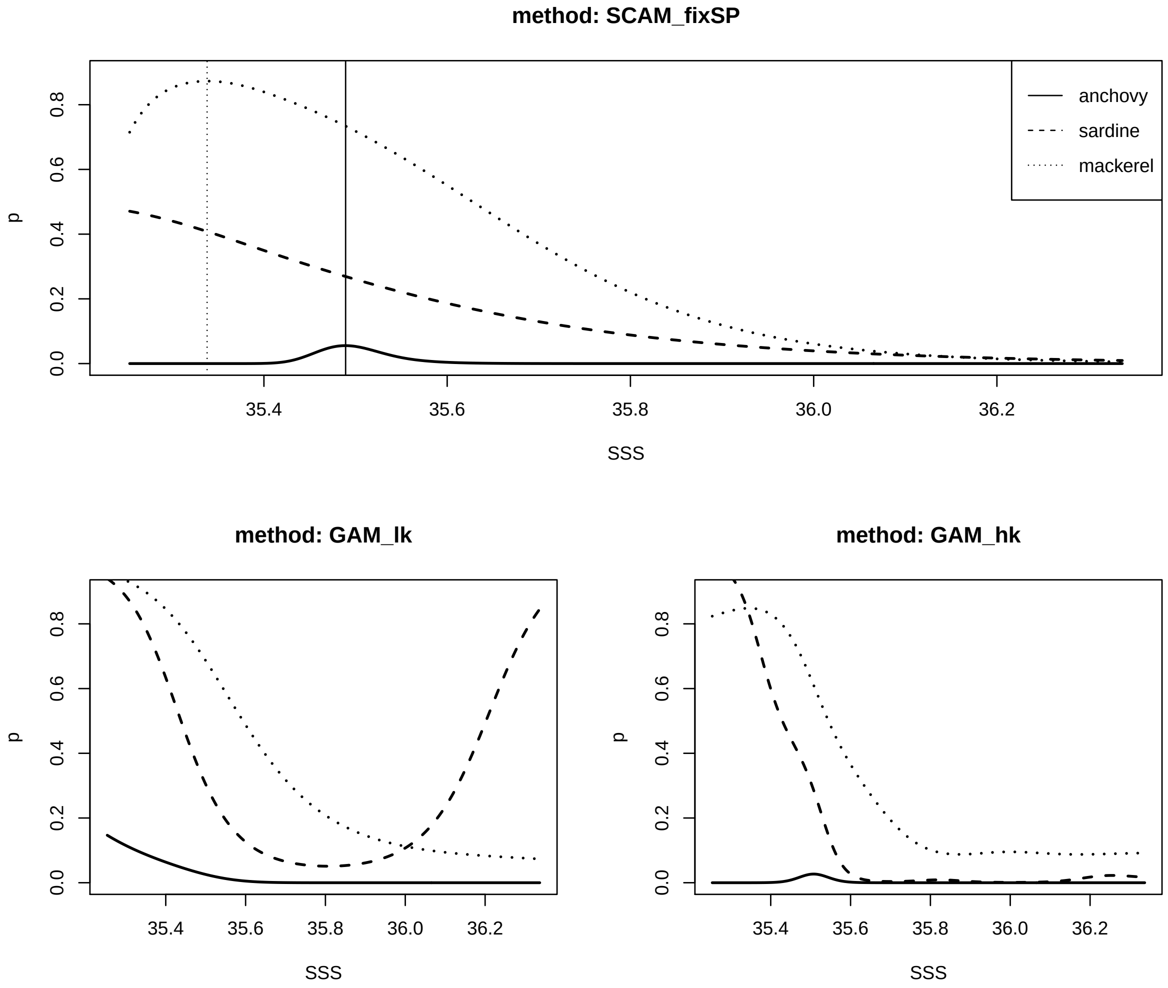

Figure 6: Predicted marginal response curves along salinity variable (SSS) for anchovy, sardine and mackerel fitted with three different methods; a proposed SC-GAM method (“SCAMfixSP’) and no restricted GAM approaches with k=3 (“GAMlk’) and k=10 (“GAMhk’).

Validation for these models was conducted via k-fold cross-validation (with $k = 5$). The data set was divided into k equally sized groups (Hijmans, 2012), using 80% of randomly selected observations to run the model and the remaining 20% for validation, iteratively for each fold. Accuracy indicators, such as AUC (Area Under the Receiver Operating Characteristic—ROC—curve) (Fielding and Bell, 1997; Raes and ter Steege, 2007), sensitivity (true predicted presences) and specificity (true predicted absences) were computed for each k random subsets and then averaged. The threshold for presence-absence classification for each species was obtained as the values maximizing sensitivity plus specificity. Obtained AUC, sensitivity, and specificity indicators are above 70% for the three species (Table 3) and are similar to the values obtained when using all data without a cross-validation process, showing good out-of-sample performance of the models.

					All data (%)		CV (%)	
Species	**Selected variables**	**AICc**	**expl.dev (%)**	**AUC**	**spec**	**sens**	**spec**	**sens**
Anchovy	SSS,SST,temp_dif,logbathy	375.75	61.10	0.92	90	93	90	93
Sardine	SSS,temp205,logbathy	899.35	33.70	0.80	79	82	79	82
Mackerel	SSS,temp_dif,logbathy	1322.62	29.90	0.77	73	81	73	81

Table 3: For each species, selected variables in the final model (using method "SCAMfixSP"), AICc, explained deviance (%), AUC, specificity and sensitivity (%) derived from the whole data set (All data), and specificity and sensitivity (%) derived from the cross-validation process (CV).

Predicted occurrence probabilities for each species have been mapped, using for prediction environmental variables from GODAS for June 1998. Apart from optimum detection for each explanatory environmental variable, extrapolated maps allow us to identify the spawning distribution of each species (Figure 7). For mackerel, the north-west part of the map shows a high probability area, although presences were not collected in this area during 1998. However, it has been reported that these species do lay in this area in recent years (Bruge et al., 2016), which confirms the reliability of the model in this area. For anchovy, areas close to the coast in the Bay of Biscay are detected as locations with high probability of presence, while for sardine, this area is wider, extending it along the Portuguese coast and up to the Celtic Sea (Figure 7).

This case study data set was also analyzed using the "boost" method, which is also capable of dealing with restricted and unrestricted variables. Results were similar to those described and can be found in the supplementary material (Annex A, Figure A.6).

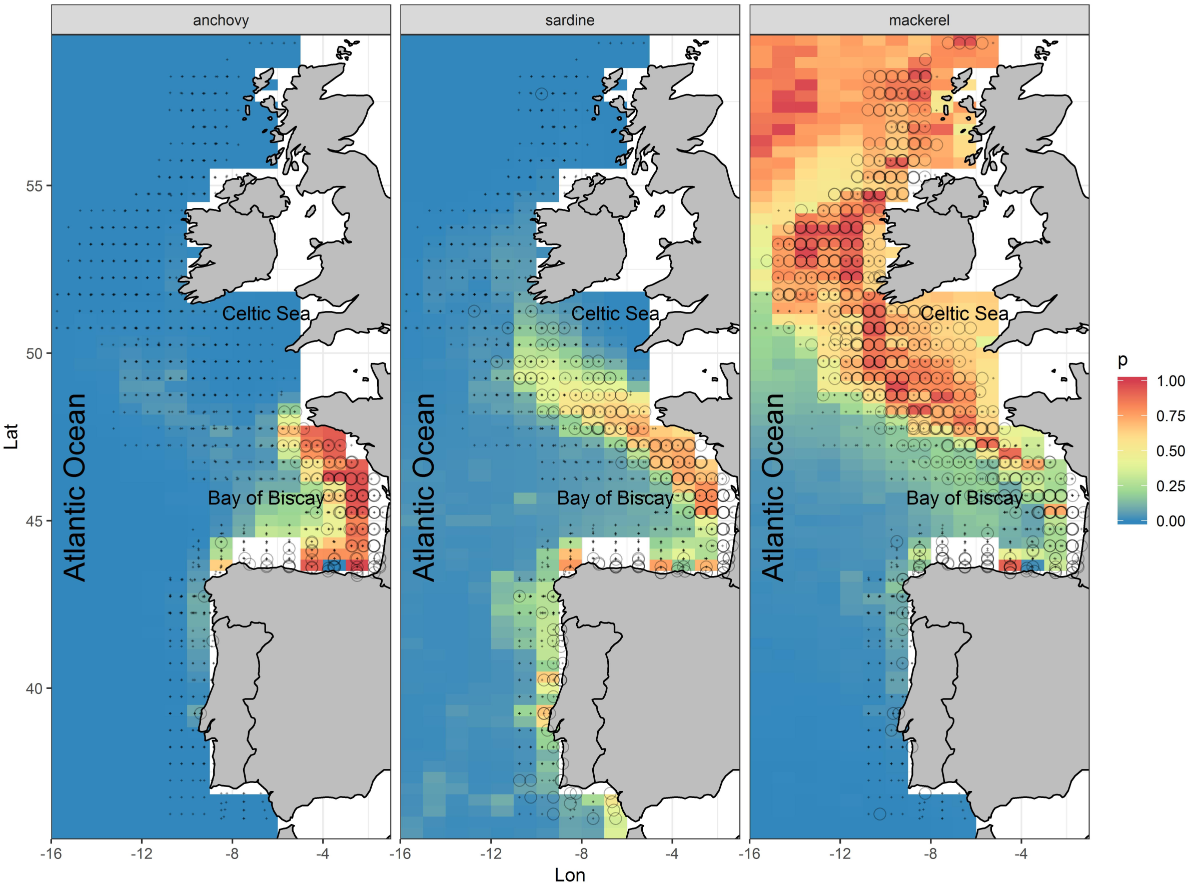

Figure 7: Predicted occurrence probabilities (p) in each map cell along with presences (circles) and absences (small dots) for each species in the north-east Atlantic.

5 Discussion

This study proposes SC-GAMs for species distribution models under the ecological niche theory framework. This emerges as a new approach in the centre ground between pure statistical fitting and process-based (or mechanistic) models that apply physiological thresholds (Martínez et al., 2015) or take into account factors affecting spatial population dynamics such as species interactions, reproduction, mortality and migration rate; see the comparison in Melle et al. (2014); Robinson et al. (2011). Our proposed model has been tested by simulation for various types of theoretical curves and sampling schemes and have been applied successfully to real case studies. The performance has been compared to other regression models without shape-constraints (GLMs and GAMs with different degrees of freedom (Guisan et al., 2002; Hastie and Tibshirani, 1990; McCullagh and Nelder, 1989)) and to models based on climate envelopes such as “Plateau” (Brewer et al., 2016).

SC-GAMs are based on the same statistical framework as GLMs and GAMs that are commonly used to fit species distribution models (Guisan et al., 2002). According to the simulation results, in several scenarios, mainly when the range of the environmental gradient was not fully covered, “GLM” and “GAMlk” methods were not able to approximate correctly the underlying theoretical niche model. Increasing the degrees of freedom of the GAM (“GAMhk”) helped to estimate curves that were closer to the true theoretical curve; however, due to random noise, fitted curves were mostly multimodal and not concave which renders them implausible under the ecological niche theory framework. An essential challenge when modelling the relationship between species occurrence and environmental drivers is to capture the signal and to differentiate it from sampling and environmental noise (Burnham and Anderson, 2003). Therefore, for all models in general, and for GAMs in particular, determining the appropriate model complexity is critical both for robust inference and for accurate prediction. Excessive flexibility can lead to overfitted models where resulting patterns can be spurious and affected by noise, and predictions based on such models can be biased and unreliable (Burnham and Anderson, 2003). Many authors have favoured simpler versus more complex models (Merow et al., 2014, and references therein), suggesting that researchers should constrain the complexity of their models based on the study objective, attributes of the data, and an understanding of how these interact with the underlying biological processes. Austin (2002) suggested that complex functions produced by GAMs could be replaced by an equivalent parametric function, simpler and ecologically easier to interpret. In practice, other authors have manually changed the degrees of freedom of the smoothing functions to achieve simpler curves (e.g., Bruge et al. (2016); Brunel et al. (2018)) or even unimodal or monotonic shapes following the ecological niche theory (Chust et al., 2014). In that context, SC-GAMs automatically provide response curves in agreement with the niche theory. In the simulations, obtained fits were closer to the underlying theoretical curves in comparison to “GLM” and “GAMlk” approaches, and in scenarios where the sampling did not cover the whole environmental range, results were similar or even better than those obtained with the most flexible GAM.

SC-GAMs were also compared to “Plateau” (Brewer et al., 2016) which is a regression model based on climatic envelopes. “Plateau” can provide the correct shape with variance estimates from the hessian in a fast way. The extension to the multivariable functions is straightforward and more variables and their potential interactions can be readily incorporated. The simulations indicated that there were no differences regarding the performance in terms of agreement with the ecological niche theory. Both the “Plateau” and the SC-GAMs satisfied the concavity restrictions and estimated the maximum correctly. However, the simulation results showed that shape-constrained models were more robust across replicates, with less uncertainty in point estimation and coverage probabilities.

The two SC-GAMs implementations tested in this study present statistically sound methods that allow for robust estimation, model comparison, and prediction. However, they exhibited some differences in terms of uncertainty estimation, computing time and ease of use. The "boost" approach seemed to be more robust to the generated uncertainties and showed more stable and narrower intervals for RMSE values and for coverage probabilities. Variance estimation in this approach is performed through bootstrapping which implies a high computational cost. Alternatively, the "SCAM" approach builds on the framework of unconstrained generalized additive models (Wood, 2006), being computationally efficient (Pya and Wood, 2014). In addition, it uses almost the same syntax as in *mgcv* R package which facilitates its use.

SC-GAMs provide a unified framework to deal with different types of variables in species distribution models. Direct variables and limiting factors are expected to have unimodal shape (symmetric or not), whereas there is no theorectical expectation regarding direct variables. However, sometimes, there might be exceptions in which the realized niche is not unimodal with respect to environmental gradients (Austin, 2002). In those cases, comparison between shape-constrained and unconstrained methods could help to better disentangle the factors defining the ecological niche of the species. When modelling species distribution based only on niche theory, results are limited by the strong assumptions such as unlimited dispersal of species, and not consideration of competition processes between species, population dynamics and adaptation of the species (*sensu* population fitness).

The extent and resolution of the data are crucial to obtain an adequate characterization of the niche of a species (Peterson et al., 2011). If the range of the environmental gradient does not cover the limits of the species, the species response is truncated and determining the actual shape of the response will be difficult (Austin, 2007). Thuiller et al. (2004) found that this could be especially problematic on the tails of the species response curves, yielding spurious projections. In our simulations, the performance of the shape-unconstrained methods was worse when the range of the environmental gradient was not fully observed. In most of the cases they were not able to fit concave curves, the single maximum was not found and presented high RMSE values. However, shape-constrained methods performed similarly regardless the type of sampling. Therefore, adding the shape constraints warranted that the species distribution model was ecologically meaningful within the observed range of the environmental variable, and facilitating its subsequent use for extrapolation and prediction.

Methods have been also tested in two different real case studies. The first case study shows that shape-constrained methods can solve issues arising with the other methods, as concluded with the simulation study. Optimum SST values and tolerance ranges obtained by SC-GAMs in the presented real case study are very similar to those reported in Bernal et al. (2007). They compiled data from all the available ichthyoplankton surveys in the Northeast Atlantic and found that spawning is restricted mainly to the shelf area and in a range of temperatures between 12°C and 17°C. Stratoudakis et al. (2007) detected that spawning seasonality vary with latitude following temperature gradients. The preferred temperatures for spawning were identified between 14 and 15°C, while temperatures below 12°C and above 16°C were avoided. In the Bay of Biscay, thermal preference at surface was found between 12°C and 15°C (Planque et al., 2007). The second case study involves the incorporation of several variables in order to find species probabilities of occurrence combining different types of variables. Results are similar to those reported in other studies that needed manual selection of smoother parameters such as the probability of presence of anchovy eggs along salinity or sea surface temperature reported in Erauskin-Extramiana et al. (2019) or the estimated optima for mackerel spawning along the salinity gradient in Brunel et al. (2018). It is shown that the framework of SC-GAMs enables us to fit both unconstrained and shape-constrained shapes for each of the included variables depending on the type and prior knowledge. It also allows us to test the shape of each predictor consistently with the expected ecological theory as suggested in Austin (2007).

We consider that proposed SC-GAMs can be readily applied for fitting distribution models and are useful tools for modelling communities of large number of species, as they result in a good balance between goodness of fit and agreement with ecological niche theory. They can incorporate multiple explanatory variables with or without interaction, both shape-constrained and unconstrained, depending on the nature of the variables involved. Thus, SC-GAMs offer the possibility of investigating, for example, the effect of climate change on multiple species without requiring sophisticated and time-consuming mechanistic models that depend on detailed knowledge of vital rates and life traits for each species. Future applications of SC-GAMs in the context of ecological models could go beyond the examples shown in this work. Bivariable smooths with concavity restrictions would allow better understanding of the interactions between environmental variables, as in Brewer et al. (2016). SC-GAMs could also be extended to include response shapes varying per grouping level as in HGAMs (Pedersen et al., 2019). In this case several species could be modeled together including interactions between the explanatory variables and the species as a factor obtaining a common effect and different response curves for each species. Multivariate adaptive regression splines (MARS, (Friedman et al., 1991)) are also claimed to have strong performance for multiresponse species distribution models (Leathwick et al., 2006). Shape constraints could be also introduced, for unimodality condition in the response curve, to obtain comparable results with SC-GAMs. SDMs can be also fitted in a Bayesian framework, allowing to incorporate prior knowledge of species ecology (Golding and Purse, 2016) or prior information on response curve shapes ((Fraaije et al., 2015) - Appendix3) using INLA as a tool to perform full Bayesian analysis of latent Gaussian models using Integrated Nested Laplace Approximation (Rue et al., 2014).

Acknowledgments

This research is supported by AZTI under a PhD grant and also by the Basque Government through the BERC 2018-2021 program and the CLIPES project. It has been also funded by the Spanish Ministry of Science, Innovation and Universities: BCAM Severo Ochoa accreditation SEV-2017-0718 and through the MTM2017-82379-R project funded by (AEI/FEDER, UE) and acronym “AFTERAM”. M.J. Brewer was funded by the Scottish Government’s Rural & Environment Science & Analytical Services Division. Data for case studies was funded by the European Commission within the frame of the National Sampling Programme and The General Secretariat of Sea for BIOMAN survey and by the EU funded programme INDICES (EU Study 97-017). Thanks are due to Xabier Irigoien that provided valuable comments to improve the manuscript. We also thank reviewers and the editor for their thoughtful reviews and constructive comments.

References

Akaike, H. (1974). A new look at the statistical model identification. In *Selected Papers of Hirotugu Akaike*, pages 215–222. Springer.

Austin, M. (1980). Searching for a model for use in vegetation analysis. *Vegetatio*, 42(1-3):11–21.

Austin, M. (1987). Models for the analysis of species' response to environmental gradients. *Vegetatio*, 69:35–45. doi: 10.1007/BF00038685.

Austin, M. (2002). Spatial prediction of species distribution: an interface between ecological theory and statistical modelling. *Ecological Modelling*, 157:101–118. doi: 10.1016/S0304-3800(02)00205-3.

Austin, M. (2007). Species distribution models and ecological theory: A critical assessment and some possible new approaches. *Ecological Modelling*, 200:1–19. doi: 10.1016/j.ecolmodel.2006.07.005.

Austin, M. and Smith, T. (1990). A new model for the continuum concept. In *Progress in theoretical vegetation science*, pages 35–47. Springer. doi: 10.1007/BF00031679.

Barbosa, F. G. and Schneck, F. (2015). Characteristics of the top-cited papers in species distribution predictive models. *Ecological Modelling*, 313:77–83. doi: 10.3152/147154403781776645.

Barton, K. (2009). Mumin: multi-model inference. r package version 1. 0. 0. *http://r-forge. r-project. org/projects/mumin/*.

Bernal, M., Stratoudakis, Y., Coombs, S., Angelico, M., De Lanzós, A. L., Porteiro, C., Sagarminaga, Y., Santos, M., Uriarte, A., Cunha, E., Valdés, L., and Borchers, D. (2007). Sardine spawning off the european atlantic coast: characterization of and spatio-temporal variability in spawning habitat. *Progress in Oceanography*, 74(2-3):210–227. doi: 10.1016/j.pocean.2007.04.018.

Bollaerts, K., Eilers, P., and van Mechelen, I. (2006). Simple and multiple p-splines regression with shape constraints. *British Journal of Mathematical and Statistical Psychology*, 59:451–469.

Borchers, D., Buckland, S., Priede, I., and Ahmadi, S. (1997). Improving the precision of the daily egg production method using generalized additive models. *Canadian Journal of Fisheries and Aquatic Sciences*, 54(12):2727–2742. 10.1139/f97-134.

Brewer, M. J., O'Hara, R. B., Anderson, B. J., and Ohlemüller, R. (2016). Plateau: a new method for ecologically plausible climate envelopes for species distribution modelling. *Methods in Ecology and Evolution*, 7(12):1489–1502. doi: 10.1111/2041-210X.12609.

Bruge, A., Alvarez, P., Fontán, A., Cotano, U., and Chust, G. (2016). Thermal niche tracking and future distribution of atlantic mackerel spawning in response to ocean warming. *Frontiers in Marine Science*, 3:86.

Brunel, T., Van Damme, C. J., Samson, M., and Dickey-Collas, M. (2018). Quantifying the influence of geography and environment on the northeast atlantic mackerel spawning distribution. *Fisheries oceanography*, 27(2):159–173.

Bühlmann, P. and Hothorn, T. (2007). Boosting algorithms: regularization, prediction and model fitting. *Statistical Science*, 22(4):477–505.

Burnham, K. P. and Anderson, D. R. (2003). *Model selection and multimodel inference: a practical information-theoretic approach.* Springer Science & Business Media. doi: 10.1007/b97636.

Busby, J. (1991). Bioclim-a bioclimate analysis and prediction system. *Plant protection quarterly (Australia).* doi: 10.1371/journal.pone.0046283.

Cerdeira, J. O., Monteiro-Henriques, T., Martins, M. J., Silva, P. C., Alagador, D., Franco, A. M., Campagnolo, M. L., Arsénio, P., Aguiar, F. C., and Cabeza, M. (2018). Revisiting niche fundamentals with tukey depth. *Methods in Ecology and Evolution.* doi: 10.1111/2041-210X.13074.

Chase, J. M. and Leibold, M. A. (2003). *Ecological niches: linking classical and contemporary approaches.* University of Chicago Press.

Chust, G., Castellani, C., Licandro, P., Ibaibarriaga, L., Sagarminaga, Y., and Irigoien, X. (2014). Are calanus spp. shifting poleward in the north atlantic? a habitat modelling approach. *ICES Journal of Marine Science*, 71(2):241–253. doi: 10.1111/j.1467-2979.2008.00315.x.

Coudun, C. and Gegout, J.-C. (2006). The derivation of species response curves with gaussian logistic regression is sensitive to sampling intensity and curve characteristics. *Ecological Modelling*, 199(2):164–175. doi: 10.1016/j.ecolmodel.2006.05.024.

de Boor, C. (1972). *A Practical Guide to Splines.* Springer.

Derber, J. and Rosati, A. (1989). A global oceanic data assimilation system. *Journal of Physical Oceanography*, 19(9):1333–1347.

Edwards, K., Barciela, R., and Butenschön, M. (2012). Validation of the nemo-ersem operational ecosystem model for the north west european continental shelf. *Ocean Sciences Discussions*, 8(6):983–1000. doi: 10.5194/os-8-983-2012.

Eilers, P. and Marx, B. (1996). Flexible smoothing with B-splines and penalties. *Stat. Sci.*, 11:89–121.

Eilers, P., Marx, B., and Durbán, M. (2015). Twenty years of P-splines. *SORT-Statistics and Operations Research Transactions*, 39(2):149–186.

Eilers, P. H. (2017). Uncommon penalties for common problems. *Journal of Chemometrics*, 31(4):e2878. DOI 10.1002/cem.2878.

Elith, J. and Leathwick, J. R. (2009). Species distribution models: ecological explanation and prediction across space and time. *Annual review of ecology, evolution, and systematics*, 40. doi: 10.1146/annurev.ecolsys.110308.120159.

Erauskin-Extramiana, M., Alvarez, P., Arrizabalaga, H., Ibaibarriaga, L., Uriarte, A., Cotano, U., Santos, M., Ferrer, L., Cabré, A., Irigoien, X., and Chust, G. (2019). Historical trends and future distribution of anchovy spawning in the bay of biscay. *Deep Sea Research Part II: Topical Studies in Oceanography*, 159:169–182.

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

Fraaije, R. G., ter Braak, C. J., Verduyn, B., Breeman, L. B., Verhoeven, J. T., and Soons, M. B. (2015). Early plant recruitment stages set the template for the development of vegetation patterns along a hydrological gradient. *Functional Ecology*, 29(7):971–980.

Friedman, J. H. et al. (1991). Multivariate adaptive regression splines. *The annals of statistics*, 19(1):1–67.

Golding, N. and Purse, B. V. (2016). Fast and flexible bayesian species distribution modelling using gaussian processes. *Methods in Ecology and Evolution*, 7(5):598–608.

Green, P. J. and Silverman, B. W. (1993). *Nonparametric regression and generalized linear models: a roughness penalty approach*. Chapman and Hall/CRC.

Guevara, J., Zadrozny, B., Buoro, A., Lu, L., Tolle, J., Limbeck, J., Wu, M., and Hohl, D. (2018). A hybrid data-driven and knowledge-driven methodology for estimating the effect of completion parameters on the cumulative production of horizontal wells. In *SPE Annual Technical Conference and Exhibition*. Society of Petroleum Engineers. doi: 10.2118/191446-MS.

Guisan, A., Edwards Jr, T. C., and Hastie, T. (2002). Generalized linear and generalized additive models in studies of species distributions: setting the scene. *Ecological modelling*, 157(2-3):89–100. doi: 10.1016/S0304-3800(02)00204-1.

Guisan, A., Tingley, R., Baumgartner, J. B., Naujokaitis-Lewis, I., Sutcliffe, P. R., Tulloch, A. I. T., Regan, T. J., Brotons, L., McDonald-Madden, E., Mantyka-Pringle, C., Martin, T. G., Rhodes, J. R., Maggini, R., Setterfield, S. A., Elith, J., Schwartz, M. W., Wintle, B. A., Broennimann, O., Austin, M., Ferrier, S., Kearney, M. R., Possingham, H. P., and Buckley, Y. M. (2013). Predicting species distributions for conservation decisions. *Ecology letters*, 16(12):1424–1435. doi: 10.1111/ele.12189.

Guisan, A. and Zimmermann, N. E. (2000). Predictive habitat distribution models in ecology. *Ecological modelling*, 135(2-3):147–186. doi: 10.1016/S0304-3800(00)00354-9.

Halvorsen, R. (2012). A gradient analytic perspective on distribution modelling. *Sommerfeltia*, 35(1):1–165. doi: 10.2478/v10208-011-0015-3.

Hastie, T. J. and Tibshirani, R. J. (1990). Generalized additive models.

Heikkinen, J. and Makipaa, R. (2010). Testing hypotheses on shape and distribution of ecological response curves. *Ecological Modelling*, 221(3):388–399. doi: 10.1016/j.ecolmodel.2009.10.030.

Helaouet, P. and Beaugrand, G. (2009). Physiology, ecological niches and species distribution. *Ecosystems*, 12(8):1235–1245. doi: 10.1007/s10021-009-9261-5.

Hijmans, R. J. (2012). Cross-validation of species distribution models: removing spatial sorting bias and calibration with a null model. *Ecology*, 93(3):679–688.

Hofner, B., Kneib, T., and Hothorn, T. (2016). A unified framework of constrained regression. *Statistics and Computing*, 26(1-2):1–14. doi: 10.1007/s11222-014-9520-y.

Hofner, B., Mayr, A., Robinzonov, N., and Schmid, M. (2014). Model-based boosting in R: A hands-on tutorial using the R package mboost. *Computational Statistics*, 29(3–35).

Hothorn, T., Buehlmann, P., Kneib, T., Schmid, M., and Hofner, B. (2010). Model-based boosting 2.0. *Journal of Machine Learning Research*, 11:2109–2113.

Hothorn, T., Buehlmann, P., Kneib, T., Schmid, M., and Hofner, B. (2018). *mboost: Model-Based Boosting.* https://CRAN.R-project.org/package=mboost.

Huisman, J., Olff, H., and Fresco, L. (1993). A hierarchical set of models for species response analysis. *Journal of Vegetation Science*, 4(1):37–46. doi: 10.2307/3235732.

Huston, M. A. (1994). *Biological diversity: the coexistence of species.* Cambridge University Press. doi: 10.1016/0169-5347(95)90033-0.

Hutchinson, G. (1957). Concluding remarks cold spring harbor symposia on quantitative biology. *GS SEARCH*, (22):415–427. doi: 10.1101/SQB.1957.022.01.039.

Hutchinson, G. E. (1978). An introduction to population ecology.

Ibaibarriaga, L., Irigoien, X., Santos, M., Motos, L., Fives, J., Franco, C., Lago de Lanzós, A., Acevedo, S., Bernal, M., Bez, N., Eltink, G., Faniha, A., Hammer, C., Iversen, S., Milligan, S., and Reid, G. (2007). Egg and larval distributions of seven fish species in north-east atlantic waters. *Fisheries Oceanography*, 16(3):284–293.

ICES (2017). Report of the working group on acoustic and egg surveys for sardine and anchovy in ices areas 7, 8, and 9.

ICES (2018). Report of the working group on mackerel and horse mackerel egg surveys.

Jamil, T. and Ter Braak, C. J. (2013). Generalized linear mixed models can detect unimodal species-environment relationships. *PeerJ*, 1:e95. doi: 10.7717/peerj.95.

Jiménez-Valverde, A., Lobo, J. M., and Hortal, J. (2008). Not as good as they seem: the importance of concepts in species distribution modelling. *Diversity and Distributions*, 14(6):885–890. doi: 10.1111/j.1472-4642.2008.00496.x.

Kearney, M. (2006). Habitat, environment and niche: what are we modelling? *OIKOS*, 115(1):186–191. doi: 10.1111/j.2006.0030-1299.14908.x.

Kriticos, D. J., Maywald, G. F., Yonow, T., Zurcher, E. J., Herrmann, N. I., and Sutherst, R. (2015). Exploring the effects of climate on plants, animals and diseases. *CLIMEX Version*, 4:184.

Lasker, R. (1985). An egg production method for estimating spawning biomass of pelagic fish: Application to the northern anchovy, engraulis mordax.

Leathwick, J., Elith, J., and Hastie, T. (2006). Comparative performance of generalized additive models and multivariate adaptive regression splines for statistical modelling of species distributions. *Ecological modelling*, 199(2):188–196.

Lehmann, A., Overton, J. M., and Austin, M. (2002). Regression models for spatial prediction: their role for biodiversity and conservation. *Biodiversity & Conservation*, 11(12):2085–2092. doi: 10.1023/A:1021354914494.

Lockwood, S., Nichols, J., and Dawson, W. A. (1981). The estimation of a mackerel (scomber scombrus l.) spawning stock size by plankton survey. *Journal of Plankton Research*, 3(2):217–233. 10.1093/plankt/3.2.217.

Martínez, B., Arenas, F., Trilla, A., Viejo, R. M., and Carreño, F. (2015). Combining physiological threshold knowledge to species distribution models is key to improving forecasts of the future niche for macroalgae. *Global change biology*, 21(4):1422–1433. doi: 10.1111/gcb.12655.

McCullagh, P. and Nelder, J. A. (1989). *Generalized linear models*, volume 37. CRC press.

Melle, W., Runge, J., Head, E., Plourde, S., Castellani, C., Licandro, P., Pierson, J., Jonasdottir, S., Johnson, C., Broms, C., Debes, H., Falkenhaug, T., Gaard, E., Gislason, A., Heath, M., Niehoff, B., Nielsen, T. G., Pepin, P., Stenevik, E. K., and Chust, G. (2014). The north atlantic ocean as habitat for calanus finmarchicus: Environmental factors and life history traits. *Progress in Oceanography*, 129:244–284. doi: 10.1016/j.pocean.2014.04.026.

Merow, C., Smith, M. J., Edwards Jr, T. C., Guisan, A., McMahon, S. M., Normand, S., Thuiller, W., Wüest, R. O., Zimmermann, N. E., and Elith, J. (2014). What do we gain from simplicity versus complexity in species distribution models? *Ecography*, 37(12):1267–1281. doi: 10.1111/ecog.00845.

Minchin, P. R. (1987). Simulation of multidimensional community patterns: Towards a comprehensive model. *Vegetatio*, 71:145–156. doi: 10.1007/BF00039167.

Morris, T., R White, I., and Crowther, M. (2019). Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, pages 1–29. doi: 10.1002/sim.8086.

Oaksenen, J., LääRä, E., Tolonen, K., and Warner, B. G. (2001). Confidence intervals for the optimum in the gaussian response function. *Ecology*, 82:1191–1197. doi: 10.1890/0012-9658(2001)082[1191:CIFTOI]2.0.CO;2.

Oksanen, J. and Minchin, P. R. (2002). Continuum theory revisited: what shape are species responses along ecological gradients? *Ecological Modelling*, 157:119–129. doi: 10.1111/aec.12463at.

Pante, E. and Simon-Bouhet, B. (2013). marmap: a package for importing, plotting and analyzing bathymetric and topographic data in r. *PLoS One*, 8(9):e73051.

Parker, K. (1980). A direct method for estimating northern anchovy engraulis mordax spawning biomass.

Parrish, R., Serra, R., and Grant, W. (1989). The monotypic sardines, sardina and sardinops: their taxonomy, distribution, stock structure, and zoogeography. *Canadian Journal of Fisheries and Aquatic Sciences*, 46(11):2019–2036. 10.1139/f89-251.

Pearson, R. G., Dawson, T. P., Berry, P. M., and Harrison, P. (2002). Species: a spatial evaluation of climate impact on the envelope of species. *Ecological modelling*, 154(3):289–300. doi: 10.1016/S0304-3800(02)00056-X.

Pedersen, E., Miller, D., Simpson, G. L., and Ross, N. (2019). Hierarchical generalized additive models: an introduction with mgcv. *PeerJ*. doi: 10.7717/peerj.6876.

Peterson, A. T., Soberón, J., Pearson, R. G., Anderson, R. P., Martínez-Meyer, E., Nakamura, M., and Araújo, M. B. (2011). *Ecological niches and geographic distributions (MPB-49)*, volume 56. Princeton University Press. doi: 10.23943/princeton/9780691136868.001.0001.

Petitpierre, B., Broennimann, O., Kueffer, C., Daehler, C., and Guisan, A. (2017). Selecting predictors to maximize the transferability of species distribution models: lessons from cross-continental plant invasions. *Global Ecology and Biogeography*, 26(3):275–287. doi: 10.1111/geb.12530.

Phillips, S. J., Anderson, R. P., and Schapire, R. E. (2006). Maximum entropy modeling of species geographic distributions. *Ecological modelling*, 190(3-4):231–259. doi: 10.1016/j.ecolmodel.2005.03.026.

Planque, B., Bellier, E., and Lazure, P. (2007). Modelling potential spawning habitat of sardine (sardina pilchardus) and anchovy (engraulis encrasicolus) in the bay of biscay. *Fisheries Oceanography*, 16(1):16–30. 10.1111/j.1365-2419.2006.00411.x.

Pocheville, A. (2015). The ecological niche: history and recent controversies. pages 547–586. 10.1007/978-94-017-9014-7_26.

Pulliam, H. R. (2000). On the relationship between niche and distribution. *Ecology letters*, 3(4):349–361. doi: 10.1046/j.1461-0248.2000.00143.x.

Pya, N. (2018). *scam: Shape Constrained Additive Models.* https://CRAN.R-project.org/package=scam.

Pya, N. and Wood, S. N. (2014). Shape constrained additive models. *Statistics and Computing*, 25(3):543–559. doi: 10.1007/s11222-013-9448-7.

R Core Team (2018). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Raes, N. and ter Steege, H. (2007). A null-model for significance testing of presence-only species distribution models. *Ecography*, 30(5):727–736.

Rigby, R. A. and Stasinopoulos, D. M. (2005). Generalized additive models for location, scale and shape,(with discussion). *Applied Statistics*, 54:507–554. doi: 10.1111/j.1467-9876.2005.00510.x.

Robinson, L., Elith, J., Hobday, A., Pearson, R., Kendall, B., Possingham, H., and Richardson, A. (2011). Pushing the limits in marine species distribution modelling: lessons from the land present challenges and opportunities. *Global Ecology and Biogeography*, 20(6):789–802.

Rue, H., Martino, S., Lindgren, F., Simpson, D., Riebler, A., and Krainski, E. T. (2014). Inla: Functions which allow to perform full bayesian analysis of latent gaussian models using integrated nested laplace approximaxion. *R package version 0.0-1404466487, URL http://www. R-INLA. org.*

Ruppert, D., Wand, M., and Carroll, R. (2003). *Semiparametric Regression.* Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Santos, M., Uriarte, A., Boyra, G., and Ibaibarriaga, L. (2018). Anchovy depm surveys 2003–2012 in the bay of biscay (subarea8): Bioman survey series. *In Massé, J., Uriarte, A., Angélico, M. M., and Carrera, P. (Eds.). Pelagic survey series for sardine and anchovy in ICES subareas 8 and 9 –Towards an ecosystem approach. ICES Cooperative Research Report No. 332. 268.*

Schmidt, M., Breidenbach, J., and Astrup, R. (2018). Longitudinal height-diameter curves for norway spruce, scots pine and silver birch in norway based on shape constraint additive regression models. *Forest Ecosystems*, 5(1):9. doi: 10.1186/s40663-017-0125-8.

Schröder, H. K., Andersen, H. E., and Kiehl, K. (2005). Rejecting the mean: Estimating the response of fen plant species to environmental factors by non-linear quantile regression. *Journal of Vegetation Science*, 16(4):373–382. doi: 10.1111/j.1654-1103.2005.tb02376.x.

Schwarz, G. et al. (1978). Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464.

Scott, J., Heglund, P., Morrison, M., Haufler, J., Raphael, M., Wall, W., and Samson, F. (2002). Predicting species occurrences: issues of scale and accuracy. doi: 10.1644/1545-1542(2003)084¡0319:R¿2.0.CO;2.

Soberón, J. and Arroyo-Peña, B. (2017). Are fundamental niches larger than the realized? testing a 50-year-old prediction by hutchinson. *PloS one*, 12(4):e0175138. doi: 10.1371/journal.pone.0175138.

Soberon, J. and Nakamura, M. (2009). Niches and distributional areas: concepts, methods, and assumptions. *Proceedings of the National Academy of Sciences of the United States of America*, 106 Suppl 2:19644–50. doi: 10.1073/pnas.0901637106.

Stasinopoulos, M., Rigby, B., Akantziliotou, C., Heller, G., Ospina, R., and Stasinopoulos, M. M. (2019). *Package gamlss. dist.*

Stratoudakis, Y., Coombs, S., de Lanzós, A. L., Halliday, N., Costas, G., Caneco, B., Franco, C., Conway, D., Santos, M. B., Silva, A., et al. (2007). Sardine (sardina pilchardus) spawning seasonality in european waters of the northeast atlantic. *Marine Biology*, 152(1):201–212. 10.1007/s00227-007-0674-4.

Ter Braak, C. J. and Looman, C. W. (1986). Weighted averaging, logistic regression and the gaussian response model. *Vegetatio*, 65(1):3–11. doi:10.1007/BF00032121.

Thuiller, W., Brotons, L., Araújo, M. B., and Lavorel, S. (2004). Effects of restricting environmental range of data to project current and future species distributions. *Ecography*, 27(2):165–172.

Wood, N. (2003). Thin plate splines regression. *Journal of the Royal Statistical Society*, 65(1):95–114.

Wood, S. (2019). "mgcv" mixed gam computation vehicle with automatic smoothness estimation. *Version 1.8-29.* https://cran.r-project.org/web/packages/mgcv/mgcv.pdf.

Wood, S. N. (2006). On confidence intervals for generalized additive models based on penalized regression splines. *Australian and New Zealand Journal of Statistics*, 48(4):445–464. doi: 10.1111/j.1467-842X.2006.00450.x.

Wood, S. N. (2017). *Generalized additive models: an introduction with R.* Chapman and Hall/CRC.

Zimmermann, N. E., Yoccoz, N. G., Edwards, T. C., Meier, E. S., Thuiller, W., Guisan, A., Schmatz, D. R., and Pearman, P. B. (2009). Climatic extremes improve predictions of spatial patterns of tree species. *Proceedings of the National Academy of Sciences*, 106(2):19723–19728. doi: 10.1073/pnas.