Modelling environmental DNA data; Bayesian variable selection accounting for false positive and false negative errors.

Jim E. Griffin

Department of Statistical Science, University College London, UK.

E-mail: j.griffin@ucl.ac.uk

Eleni Matechou

School of Mathematics, Statistics and Actuarial Science, University of Kent, Canterbury, UK.

Andrew S. Buxton

Durrell Institute of Conservation and Ecology, School of Anthropology and Conservation, University of Kent, Canterbury, UK.

Dimitrios Bormpoudakis

Durrell Institute of Conservation and Ecology, School of Anthropology and Conservation, University of Kent, Canterbury, UK.

Richard A. Griffiths

Durrell Institute of Conservation and Ecology, School of Anthropology and Conservation, University of Kent, Canterbury, UK.

Summary. Environmental DNA (eDNA) is a survey tool with rapidly expanding applications for assessing presence of a species at surveyed sites. eDNA methodology is known to be prone to false negative and positive errors at the data collection and laboratory analysis stage. Existing models for eDNA data require augmentation with additional sources of information to overcome identifiability issues of the likelihood function and do not account for environmental covariates that predict the probability of species presence or the proba-

bilities of error. We present a novel Bayesian model for analysing eDNA data by proposing informative prior distributions for logistic regression coefficients that allow us to overcome parameter identifiability, while performing efficient Bayesian model-selection. Our methodology does not require the use of trans-dimensional algorithms and provides a general framework for performing Bayesian variable selection under informative prior distributions in logistic regression models.

Keywords: Informative prior distributions, known presences, likelihood symmetries, logistic regression, occupancy probability, Pólya-Gamma scheme

1. Introduction

Since the initial proof of concept by Ficetola et al. (2008), the use of environmental DNA (eDNA) for the assessment of aquatic biodiversity has been rapidly expanding. In essence, the eDNA survey method isolates DNA that has become separated from an organism and suspended within the water column, to identify the recent presence of that species within a waterbody (Jane et al., 2015). Surveyors opt for eDNA over traditional survey methods for two reasons. First, eDNA offers a rapid assessment tool with potential cost (Rees et al., 2014) and logistical savings, allowing large-scale monitoring programs to be implemented, that would be too onerous using traditional methods such as trapping or electrofishing (Jerde et al., 2011; Biggs et al., 2015). Second, some studies have indicated a decrease in the probability of a false negative error over traditional methods, (Jerde et al., 2011; Biggs et al., 2015), particularly for rare and cryptic species that are difficult to detect (Sigsgaard et al., 2015).

Nevertheless, eDNA methodology is not error-free and both false positive and false negative errors are possible in the two stages of an eDNA survey: the data collection stage (Stage 1) and laboratory analysis stage (Stage 2) (Biggs et al., 2014; Tréguier et al., 2014). (see Fig. 1 for a schematic representation of the stages of eDNA sampling).

The recently developed model for eDNA data by Guillera-Arroita et al. (2017) estimates the probability of species presence at a site and the (conditional) probabilities of error in the two survey stages. However, the model requires augmenting eDNA data with two additional sources of information and does not take into account site covariates

that affect the probability of species presence or the probabilities of error in either of the two stages. Clearly, additional sources of information may not always be available, especially in large scale surveys targeting several species. Additionally, it is well-known that probabilities of error may be influenced by environmental and waterbody characteristics (Ficetola et al., 2015). Regarding Stage 1, several pond characteristics such as dense mats of vegetation or wide shallow drawdown zones around ponds, may prevent the thorough mixing of eDNA into the water column, potentially resulting in a failure to collect target DNA (Biggs et al., 2014). Similarly, water flows between ponds may allow for the transport of eDNA from one pond to another, or the removal of eDNA from a survey area (Biggs et al., 2014). Regarding Stage 2, errors may result from components within the water reducing the efficiency of DNA extraction, such as sediment or organic matter, from poor lab practices or from DNA becoming airborne, allowing amplification through contamination.

eDNA surveys are now being enshrined within policy and commercial practice. Commercial and political decision-making has started to rely solely on results from eDNA surveys to assess species presence at surveyed sites, whether this be in management decisions around the introduction of invasive species of Asian carp in the USA (Jerde et al., 2011) or development mitigation decisions surrounding protected species such as the great crested newt in the UK (Natural England, 2017). However, neither the reliability of eDNA methodology with regard to estimating species presence nor the effect of environmental covariates on the probabilities of error in either Stage have been assessed. As a result, decisions with prominent commercial and political consequences are being made with unknown levels of confidence in the results. The ability to identify the degree of error from eDNA surveys when assessing species presence and to link probabilities of error to environmental covariates would be hugely valuable in demonstrating the accuracy of the technique and assigning confidence in individual samples (Barnes et al., 2014; Barnes and Turner, 2016; Willoughby et al., 2016).

In this paper we present a Bayesian formulation of the Guillera-Arroita et al. (2017) model, but with all model parameters as functions of categorical and continuous covariates within a logistic regression framework. We propose a set of prior distributions for

the regression coefficients that overcomes the identifiability issues of the model, introduced by the likelihood function, and enables us to estimate site-specific probabilities of species presence without requiring additional sources of information. We show how information on verified species presences for a number of sites, when that is available, can be incorporated in the model. We present an efficient algorithm for performing Bayesian variable selection (George and McCulloch, 1997; Chipman et al., 2001; O'Hara and Sillanpää, 2009) elegantly, even when the number of possible models to be considered is large. We exploit the Pólya-Gamma (Polson et al., 2013) data augmentation scheme for logistic regression models, which allows us to efficiently update the model with regression coefficients marginalized out. This approach avoids using trans-dimensional algorithms, such as reversible jump MCMC (Green, 1995), that require careful tuning.

We present our Bayesian model in section 2 and we give details on computational aspects in section 3. Section 4 presents a simulation study, assessing the effect on the accuracy of the variable selection process when varying the number of sites, number of samples, baseline probability of species presence or the proportion of sites with associated verified species presence. We apply our proposed model to a data set commissioned by Natural England, collected and extracted using a precipitation in ethanol eDNA methodology, with 12 technical polymerase chain reaction (qPCR) replicates in the eDNA analysis phase as per Biggs et al. (2014). The results of our model are presented in section 5 and the paper concludes in section 6. The associated online supplementary material presents further details on our model and algorithm and convergence diagnostics.

The eDNA data that motivated the work can be requested by Natural England (Peter.Brotherton@naturalengland.org.uk) while the methods developed in the paper have been implemented in an Rshiny app (Chang et al., 2019) `https://seak.shinyapps.io/eDNA/`.

2. Modelling eDNA data

eDNA data on a species result from visiting S independent sites and collecting M independent water samples from each site. Each of the SM water samples is subsequently analysed in K independent eDNA qPCR replicates, with each replicate leading to either

a negative or a positive result for eDNA presence of the species. We denote the number of positive results (from the K qPCR replicates) in the m-th water sample collected at the s-th site by y_{sm}. We may also observe confirmed species presences at some sites where the species is present following incidental observations of the target species, their larvae or eggs during eDNA sample collection, and we denote a confirmed species presence at site s by $k_s = 1$ (and $k_s = 0$ otherwise).

Guillera-Arroita et al. (2017) proposed a model for eDNA data that allows for errors at the data collection (Stage 1) and laboratory analysis (Stage 2) stages. The model is an extension of the Royle and Link (2006) mixture model that only allows for errors at the detection stage in occupancy studies (MacKenzie et al., 2002). We aim to identify the site-specific covariates that affect the probability of species presence as well as the probabilities of error at the two survey stages and so extend the Guillera-Arroita et al. (2017) model to allow for site-specific model parameters. We define $z_s = 1$ if the s-th site is occupied (*i.e.* the species is present) and $z_s = 0$ otherwise, and $w_{sm} = 1$ if eDNA of the species is present in the m-th sample of the s-th site and $w_{sm} = 0$ otherwise. We assume that the probability of species presence at site s is ψ_s and there exists a common probability π of a confirmed species presence at an occupied site. In Stage 1, the probability of eDNA presence in a water sample from site s is θ_{11s} if the site is occupied (Stage 1 true positive) and θ_{10s} otherwise (Stage 1 false positive). In Stage 2, the probability of a positive qPCR replicate is p_{11s} if eDNA of the species is present in a water sample from site s (Stage 2 true positive) and p_{10s} otherwise (Stage 2 false positive). The model assumes that a positive qPCR replicate is only directly affected by eDNA presence and not species presence. We assume conditional independence between replicates, samples and sites and further assume that all K qPCR replicates are identically distributed. We note that we cannot have confirmed species presence at unoccupied sites, which motivates modelling k_s conditional on species presence, as the data are missing not-at-random. A schematic representation of the stages of eDNA surveys and our corresponding model is given in Fig. 1.

The model implies that the marginal distribution of y_{sm} is a two component mixture model -with components $\text{Bi}(K, p_{11s})$ and $\text{Bi}(K, p_{10s})$ with $\psi_s\theta_{11s}+(1-\psi_s)\theta_{10s}$ the weight

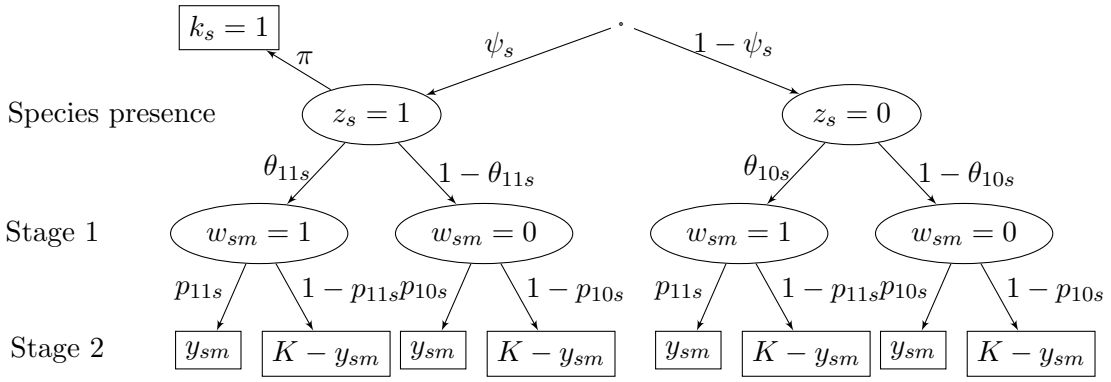


Fig. 1. Schematic representation of the hierarchical model defined in (1). Unobservable states are represented by ellipses and data by rectangles.

on the first component- and is represented in hierarchical form as

$$\begin{array}{lccc}
\text{Species presence} & & z_s \sim \text{Bernoulli}(\psi_s), & \\
\text{Confirmed species presence} & k_s|z_s = 1 \sim \text{Bernoulli}(\pi), & & \mathbb{P}(k_s = 1|z_s = 0) = 0, \\
\text{Stage 1} & w_{sm}|z_s = 1 \sim \text{Bernoulli}(\theta_{11s}), & & w_{sm}|z_s = 0 \sim \text{Bernoulli}(\theta_{10s}), \\
\text{Stage 2} & y_{sm}|w_{sm} = 1 \sim \text{Binomial}(K, p_{11s}), & & y_{sm}|w_{sm} = 0 \sim \text{Binomial}(K, p_{10s}).
\end{array} \tag{1}$$

The model leads to the following expression for the likelihood function

$$\begin{aligned}
&L(\psi, \theta_{11}, \theta_{10}, p_{11}, p_{10}, \pi|y, k) \\
&\propto \prod_{s=1}^{S} \left[\psi_s(1-\pi) \left\{ \prod_{m=1}^{M} \left(\theta_{11s} p_{11s}^{y_{sm}} (1-p_{11s})^{K-y_{sm}} + (1-\theta_{11s}) p_{10s}^{y_{sm}} (1-p_{10s})^{K-y_{sm}} \right) \right\} \right. \\
&\left. + (1-\psi_s) \left\{ \prod_{m=1}^{M} \left(\theta_{10s} p_{11s}^{y_{sm}} (1-p_{11s})^{K-y_{sm}} + (1-\theta_{10s}) p_{10s}^{y_{sm}} (1-p_{10s})^{K-y_{sm}} \right) \right\} \right]^{1-k_s} \\
&\times \left[\psi_s \pi \prod_{m=1}^{M} \left(\theta_{11s} p_{11s}^{y_{sm}} (1-p_{11s})^{K-y_{sm}} + (1-\theta_{11s}) p_{10s}^{y_{sm}} (1-p_{10s})^{K-y_{sm}} \right) \right]^{k_s}.
\end{aligned} \tag{2}$$

where $\psi = \{\psi_s\}_{s=1,\ldots,S}$, $\theta_{11} = \{\theta_{11s}\}_{s=1,\ldots,S}$, $\theta_{10} = \{\theta_{10s}\}_{s=1,\ldots,S}$, $p_{11} = \{p_{11s}\}_{s=1,\ldots,S}$, and $p_{10} = \{p_{10s}\}_{s=1,\ldots,S}$. The model is only locally identifiable (Cole et al., 2010) since there are four equally supported solutions in terms of the model parameters which give rise to the same likelihood function value (see Table 1 in Guillera-Arroita et al. (2017)

for details, also reproduced here in Table 1.) When the data include confirmed species

Table 1. Parameter values equally supported by the likelihood of equation (2) in the case of no site-specific covariates and no confirmed species presences, leading to $\psi_s = \psi$, $\theta_{11s} = \theta_{11}$, $\theta_{10s} = \theta_{10}$, $p_{11s} = p_{11}$, $p_{10s} = p_{10}$, $k_s = 0$, $\forall s$, and $\pi = 0$.

Solution	ψ_s	θ_{11s}	θ_{10s}	p_{11s}	p_{10s}
1	a	b	c	d	e
2	a	$1-b$	$1-c$	e	d
3	$1-a$	c	b	d	e
4	$1-a$	$1-c$	$1-b$	e	d

presences, the number of solutions with the same support by the likelihood function reduces to two, since we can now distinguish between solutions (1,2) and (3,4) in Table 1.

This lack of identifiability can be addressed in several ways. Guillera-Arroita et al. (2017) suggest introducing at least two additional sources of information and consider aural surveys and laboratory calibration experiments in addition to eDNA data in their data analysis. Alternatively, as the model can be seen as a product of mixture models, one could use ad-hoc methods for identifying mixture models, such as label-switching methods in a Bayesian analysis (see Papastamoulis, 2016, for a review of label-switching methods and an R package). We consider these inappropriate in this case since they assume exchangeability of the mixture parameters, which does not hold for the model introduced here. For a discussion on identifiability of mixtures of binomial models see Grün and Leisch (2008). Instead, we address the identifiability issue by introducing an informative prior distribution for the model parameters, which reflects our prior knowledge about the reliability of each of the eDNA survey stages, and show how this prior distribution can be extended to allow for variable selection of site-specific covariates. Additionally, we note that this prior can be used with or without additional data, such as confirmed species presences.

In our case study, we want to understand the effects of covariates on the parameters ψ, θ_{11}, θ_{10}, p_{11} and p_{10}, which can be accomplished using Bayesian variable selection (BVS) within a logistic regression model for each parameter. The model including all

potential covariates will most likely be unnecessarily complicated, therefore the use of BVS will address the danger of over-fitting given the limited amount of data available for each site and account for the possibility that different covariates may be important in each of the five logistic regression models. For $\xi \in \{\psi, \theta_{11}, \theta_{11}, p_{11}, p_{10}\}$, we assume that there are D^ξ available covariates and introduce γ^ξ, a D^ξ–dimensional vector for which $\gamma_k^\xi = 1$ if the k-th available covariate is included in the linear predictor for parameter ξ and 0 otherwise. We define $d^\xi = \sum_{j=1}^{D^\xi} \gamma_j^\xi$ to be the number of included covariates for parameter ξ and X^ξ to be an $(S \times d^\xi)$-dimensional design matrix of the included covariates for parameter ξ. The logistic regression model for ξ is

$$\text{logit}(\xi_s) = \eta_s^\xi = \mu^\xi + \sum_{j=1}^{d^\xi} X_{s,j}^\xi \beta_j^\xi.$$

We assume that all continuous covariates have been centred and are measured on the same scale (for example, by standardizing the covariates so that the sample variance is 1) and dummy variables for categorical covariates were defined relative to a baseline class. The prior distribution for μ^ξ, β^ξ and γ^ξ has the standard form

$$f(\mu^\xi, \beta^\xi, \gamma^\xi) = f(\mu^\xi, \beta^\xi | \gamma^\xi) f(\gamma^\xi).$$

The prior distribution on the included covariates γ^ξ follows the suggestion of Ley and Steel (2009),

$$\gamma_k^\xi \overset{i.i.d.}{\sim} \text{Bernoulli}(\tau^\xi) \text{ and } \tau^\xi \sim \text{Be}\left(1, \frac{D^\xi - \bar{d}^\xi}{\bar{d}^\xi}\right). \tag{3}$$

This implies that the prior mean of d^ξ is $\bar{d}^\xi$, which can be chosen to reflect prior beliefs about the number of important covariates. The structure implies a beta-binomial prior distribution for d^ξ, which robustifies the analysis to misspecification of $\bar{d}^\xi$. The prior distribution for the intercept and regression coefficients is

$$\mu^\xi \sim \text{N}\left(\mu_0^\xi,\ \Delta^\xi \phi_\alpha^\xi\right), \qquad \beta^\xi \sim \text{N}\left(0_{d^\xi},\ \Delta^\xi \phi_\beta^\xi C^\xi\right) \tag{4}$$

where 0_m represents an $(m \times 1)$-dimensional vector of 0's, Δ^ξ is a scaling hyperparameter defined below, and C^ξ are prior correlation matrices for the included regression coefficients in each case. The hyperparameters ϕ_α^ξ and ϕ_β^ξ control the relative prior variance of the intercept and the regression coefficients. There are several commonly used prior

covariance structures in BVS. The g-prior and various mixtures of g priors and their use in generalized linear models are reviewed by Li and Clyde (2018). Here, we will use a prior distribution that is independent across covariates. For a given model γ^{ξ} with d_1 continuous covariates and d_2 categorical covariates (where the k-th categorical covariate has $L_k + 1$ levels), we use

$$C^{\xi} = \begin{pmatrix} I_{d_1} & 0_{d_1 \times L_1} & \cdots & 0_{d_1 \times L_{d_2}} \\ 0_{L_1 \times d_1} & \frac{1}{2}\left(J_{L_1} + I_{L_1}\right) & \cdots & 0_{L_1 \times L_{d_2}} \\ \vdots & \vdots & \ddots & \vdots \\ 0_{L_{d_2} \times d_1} & 0_{L_{d_2} \times L_1} & \cdots & \frac{1}{2}\left(J_{L_{d_2}} + I_{L_{d_2}}\right) \end{pmatrix}$$

where J_k is a k-dimensional matrix of 1's, I_k is the k-dimensional identity matrix and $O_{k \times m}$ is a $(k \times m)$-dimensional matrix of 0's. This choice of prior correlation matrix for the regression coefficients associated with categorical covariates makes the prior distribution invariant to the choice of the baseline class (Fearn et al., 1999).

In the absence of any prior information on ψ, we will use $\mu_0^{\psi} = 0$ and $\Delta^{\psi} = 1$. For the parameters θ_{11}, θ_{11}, p_{11} and p_{10}, an informative prior distribution will be used to overcome the likelihood symmmetries. Table 1 shows that it is natural to think about the parameters in pairs (θ_{11}, θ_{10}) and (p_{11}, p_{10}) since the likelihood is symmetric with respect to these pairs. Our proposed prior distributions will encode the idea that the true positive probabilities in both stages (θ_{11} and p_{11}) are highly likely to be larger than their corresponding false positive probabilities (θ_{10} and p_{10}), which was shown to be true in all cases considered by Guillera-Arroita et al. (2017). We will describe a prior distribution for the pair p_{11} and p_{10} but the same idea is also used with the pair θ_{11} and θ_{10}. To introduce our proposed prior distribution, we first consider the model without covariates. We will say that p_{11} and p_{10} are *a priori* ϵ well-ordered if $\mathbb{P}(p_{11} < p_{10}) = \epsilon$. If we choose ϵ to be small, there will be negligible probability that $p_{10} > p_{11}$. In this case, an *a priori* ϵ well-ordered prior is

$$\text{logit}(p_{11}) \sim \text{N}\left(\text{logit}(a), \frac{(\text{logit}(a) - \text{logit}(b))^2}{2(\Phi^{-1}(\epsilon))^2}\right) \tag{5}$$

and

$$\text{logit}(p_{10}) \sim \text{N}\left(\text{logit}(b), \frac{(\text{logit}(a) - \text{logit}(b))^2}{2(\Phi^{-1}(\epsilon))^2}\right) \tag{6}$$

where $a > b$ and a and b are the prior medians of p_{11} and p_{10} respectively (further details are given in the Online Supplementary Material).

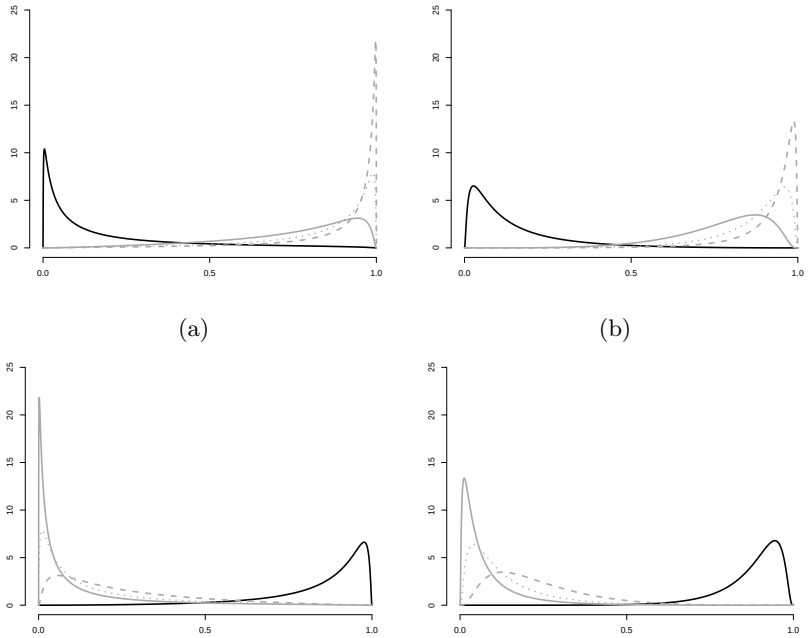


Fig. 2. First row: the prior density for p_{10} when $b = 0.1$ (black solid line) and for p_{11} when $a = 0.95$ (gray dashed line), $a = 0.9$ (gray dotted line) and $a = 0.8$ (gray solid line) with $\epsilon = 0.023$ (a) and $\epsilon = 0.0013$ (b). Second row: the prior density for p_{11} when $a = 0.9$ (black solid line) and for p_{10} when $b = 0.2$ (gray dashed line), $b = 0.1$ (gray dotted line) and $b = 0.05$ (gray solid line) with $\epsilon = 0.023$ (c) and $\epsilon = 0.0013$ (d).

Fig. 2 shows examples of this prior. For fixed a and b, as ϵ decreases, the overlap between the prior densities of p_{11} and p_{10} decreases. For fixed ϵ and a, as b increases, the median of the prior distribution for p_{10} shifts to the right. There is the opposite effect for p_{11} as a increases with fixed ϵ and b.

In the regression case, we define the prior distribution to be *a priori* ϵ well-ordered if the prior predictive distributions of the linear predictors $\eta_s^{p_{11}}$ and $\eta_s^{p_{10}}$ for a randomly chosen $X_s^{p_{11}}$ and $X_s^{p_{10}}$ have the same mean and variance as the priors in (5) and (6). We denote the covariance matrices of $X^{p_{11}}$ and $X^{p_{10}}$ by $\Sigma^{p_{11}}$ and $\Sigma^{p_{10}}$, respectively, which

suggests that $\mathrm{E}\left[\eta_s^{p_{11}}\right] = \mathrm{logit}(a)$, $\mathrm{E}\left[\eta_s^{p_{10}}\right] = \mathrm{logit}(b)$, and

$$\mathrm{V}\left[\eta_s^{\xi}\right] = \left(\phi_\alpha^\xi + \phi_\beta^\xi \sum_{i=1}^{d^\xi}\sum_{j=1}^{d^\xi} C_{ij}^\xi \Sigma_{ij}^\xi\right)\Delta^\xi$$

for $\xi \in \{p_{11}, p_{10}\}$. We obtain an *a priori* ϵ well-ordered prior by defining

$$\Delta^\xi = \frac{(\mathrm{logit}(a) - \mathrm{logit}(b))^2}{2(\Phi^{-1}(\epsilon))^2\left(\phi_\alpha^\xi + \phi_\beta^\xi \sum_{i=1}^{d^\xi}\sum_{j=1}^{d^\xi} C_{ij}^\xi \Sigma_{ij}^\xi\right)}.$$

This choice implies that the priors in (4) only depend on ϕ_α^ξ and ϕ_β^ξ through the ratio $r_0^\xi = \frac{\phi_\alpha^\xi}{\phi_\beta^\xi}$. Finally, the probability that an occupied site has a record of verified species presence associated with it, π, is given a uniform prior distribution on $(0, 1)$.

3. Computational Approach

Inference in the model in (2) can be made by employing Markov chain Monte Carlo methods using the hierarchical representation in (1), which treats z_s and w_{sm} as latent variables. For $\xi \in \{\psi, \theta_{11}, \theta_{10}, p_{11}, p_{10}\}$, we group together the regression coefficients as $\nu^\xi = (\mu^\xi, \beta^\xi)$. We group all regression coefficients as $\nu = \{\nu^\psi, \nu^{\theta_{11}}, \nu^{\theta_{10}}, \nu^{p_{11}}, \nu^{p_{10}}\}$, and all variable inclusion parameters as $\gamma = \{\gamma^\psi, \gamma^{\theta_{11}}, \gamma^{\theta_{10}}, \gamma^{p_{11}}, \gamma^{p_{10}}\}$. This leads to the following posterior distribution

$$f(\nu, \gamma, w, z|y, k) \propto f(y, w, z, k|\nu, \gamma)\, f(\nu|\gamma)\, f(\gamma) \tag{7}$$

$$\begin{aligned}
&\propto \prod_{s=1}^{S}\left[\pi^{k_s z_s}(1-\pi)^{(1-k_s)z_s}\frac{\exp(\eta_s^\psi)^{z_s}}{1+\exp(\eta_s^\psi)}\prod_{m=1}^{M}\left[\left\{\frac{\exp(\eta_s^{\theta_{11}})^{w_{sm}}}{1+\exp(\eta_s^{\theta_{11}})}\right\}^{z_s}\right.\right.\\
&\left.\left.\times\left\{\frac{\exp(\eta_s^{\theta_{10}})^{w_{sm}}}{1+\exp(\eta_s^{\theta_{10}})}\right\}^{(1-z_s)}\left\{\frac{\exp(\eta_s^{p_{11}})^{y_{sm}}}{(1+\exp(\eta_s^{p_{11}}))^K}\right\}^{w_{sm}}\left\{\frac{\exp(\eta_s^{p_{10}})^{y_{sm}}}{(1+\exp(\eta_s^{p_{10}}))^K}\right\}^{(1-w_{sm})}\right]\right]\\
&\times f(\nu|\gamma)\, f(\gamma)
\end{aligned} \tag{8}$$

where, from (4),

$$f(\nu|\gamma) \propto \prod_{\xi\in\{\psi,\theta_{11},\theta_{10},p_{11},p_{10}\}} \exp\left[-\frac{1}{2\Delta^\xi}\left\{\frac{(\mu^\xi - \mu_0^\xi)^T(\mu^\xi - \mu_0^\xi)}{\phi_\alpha^\xi} + \frac{{\beta^\xi}^T(C^\xi)^{-1}\beta^\xi}{\phi_\beta^\xi}\right\}\right].$$

and, from (3),

$$f(\gamma) \propto \prod_{\xi\in\{\psi,\theta_{11},\theta_{10},p_{11},p_{10}\}} \left(\frac{D^\xi - \bar{d}^\xi}{\bar{d}^\xi}\right)\frac{\Gamma(1+d^\xi)\Gamma\left(D^\xi - d^\xi + \frac{D^\xi - \bar{d}^\xi}{\bar{d}^\xi}\right)}{\Gamma\left(D^\xi + 1 + \frac{D^\xi - \bar{d}^\xi}{\bar{d}^\xi}\right)}.$$

This posterior distribution can be expressed as the product of the posterior distributions of five logistic regression models (for ψ, θ_{11}, θ_{10}, p_{11} and p_{10}, respectively). We combine the Pólya-Gamma sampling method for logistic models (Polson et al., 2013) with the standard Add-Delete-Swap Metropolis-Hastings sampling scheme for BVS (Brown et al., 1998; Chipman et al., 2001) to define a sampler that avoids using trans-dimensional algorithms, such reversible jump Markov chain Monte Carlo (Green, 1995).

The Pólya-Gamma sampling method uses the identity

$$\frac{(\exp\{x\})^a}{(1+\exp\{x\})^b} = 2^{-b}\int_0^\infty \exp\{-\omega(x^2-2\kappa x)/2\} f(\omega)\, d\omega \tag{9}$$

where $\kappa = a - b/2$, $\omega \sim \text{PG}(b,0)$ and $\text{PG}(b,0)$ represents the Pólya-Gamma distribution. This distribution is defined as an infinite sum, so that if $X \sim \text{PG}(b,c)$ then

$$X \overset{D}{=} \frac{1}{2\pi^2}\sum_{k=1}^{\infty}\frac{g_k}{(k-\frac{1}{2})^2+\frac{c^2}{4\pi^2}}$$

where $g_k \overset{i.i.d}{\sim} \text{Ga}(b,1)$. Polson et al. (2013) describe efficient methods for simulating draws from the Pólya-Gamma distribution that overcome the challenges of working with an infinite sum.

The identity in (9) allows us to write each element of (8) in terms of an integral. Specifically, for any parameter ξ, with $\xi \in \{\psi, \theta_{11}, \theta_{10}, p_{11}, p_{10}\}$,

$$\left[\frac{\exp(\eta_s^\xi)^a}{\left(1+\exp(\eta_s^\xi\right)^b}\right]^c = \left[2^{-b}\int_0^\infty \exp\left\{-\omega_s^\xi\left(\eta_s^{\xi\,2} - 2(a-b/2)\eta_s^\xi\right)/2\right\} f(\omega_s^\xi)\, d\omega_s^\xi\right]^c$$

where $\omega_s^{\xi} \sim \text{PG}(b, 0)$. This allows us to define the extended posterior density

$$
\begin{aligned}
f(\nu, \gamma, z, \omega | y) \propto \prod_{s=1}^{S} \Bigg[& \pi^{k_s z_s} (1-\pi)^{(1-k_s) z_s} 2^{-1} \exp\left\{ -\omega_s^{\psi} \left({\eta_s^{\psi}}^2 - 2(z_s - 1/2)\eta_s^{\psi} \right) /2 \right\} f(\omega_s^{\psi_s}) \\
& \times \prod_{m=1}^{M} \Bigg\{ \left[2^{-1} \exp\left\{ -\omega_s^{\theta_{11}} \left({\eta_s^{\theta_{11}}}^2 - 2(w_{sm} - 1/2)\eta_s^{\theta_{11}} \right) /2 \right\} f(\omega_s^{\theta_{11}}) \right]^{z_s} \\
& \times \left[2^{-1} \exp\left\{ -\omega_s^{\theta_{10}} \left({\eta_s^{\theta_{11}}}^2 - 2(w_{sm} - 1/2)\eta_s^{\theta_{10}} \right) /2 \right\} f(\omega_s^{\theta_{10}}) \right]^{z_s} \\
& \times \left[2^{-K} \exp\left\{ -\omega_s^{p_{11}} \left({\eta_s^{p_{11}}}^2 - 2(y_{sm} - K/2)\eta_s^{p_{11}} \right) /2 \right\} f(\omega_s^{p_{11}}) \right]^{w_{sm}} \\
& \times \left[2^{-K} \exp\left\{ -\omega_s^{p_{10}} \left({\eta_s^{p_{10}}}^2 - 2(y_{sm} - K/2)\eta_s^{p_{10}} \right) /2 \right\} f(\omega_s^{p_{10}}) \right]^{(1-w_{sm})} \Bigg\} \Bigg] \\
& \times f(\nu|\gamma) f(\gamma) \qquad (10)
\end{aligned}
$$

where $\omega = \{\omega^{\psi}, \omega^{\theta_{11}}, \omega^{\theta_{10}}, \omega^{p_{11}}, \omega^{p_{10}}\}$. The identity in (9) implies that integrating ω^{ψ}, $\omega^{\theta_{11}}$, $\omega^{\theta_{10}}$, $\omega^{p_{10}}$ and $\omega^{p_{10}}$ from the posterior density in (10) leads to the posterior density in (8). The linear predictors now enter this posterior distribution in a form that implies that the full conditionals of the regression parameters are normal distributions, allowing us to integrate out the regression coefficients and perform efficient covariate selection. The included covariates can be updated using a Metropolis-Hastings step where covariates are either added to the model, deleted from the model or a covariate currently included in the model is replaced by one currently excluded from the model. The steps of the Gibbs sampler are given in the Online Supplementary Material.

4. Simulation

We performed a simulation study to assess the ability of our BVS procedure to identify important predictors for each model parameter, while quantifying the effect on performance of changing q, which is the proportion of occupied sites with confirmed species presences, and M, which is the number of water samples from each site.

We considered M being set equal to 1 or 5, S (the number of sites) being set equal to 200 or 500, and q being set equal to 0, 0.2, 0.4, 0.6, or 0.8 with the baseline probability of species presence being 50% or 75%. Finally, K, the number of qPCR samples was set equal to 12. We note that for the newt data analysed in section 5, $M = 1$, $S = 189$,

$q = 0.0794$ and $K = 12$.

The comparison in each case is based on the covariate selection performance of the model by looking at the proportion of covariates that are correctly included for each of the model parameters. The data were simulated using 10 covariates, with five being important predictors for ψ, six for θ_{11} and θ_{10}, and seven for p_{11} and p_{10}. More details on the simulation study are presented in the Online Supplementary Material.

Fig. 3 presents the mean proportion of correctly identified species presences and the mean proportion of covariates, obtained over five simulation runs, that are correctly included in the model for each of the five parameters. The results suggest that when $M = 5$, inference on species presence is consistently good, regardless of q, while it improves considerably as q increases when $M = 1$. When the overall occupancy rate is around 50%, inference on ψ is good, regardless of q, while when occupancy rate is around 75%, inference improves as the proportion of sites with verified species presence data increases, especially in the $M = 1$, $S = 200$ case. Inference on p_{11} and p_{10} is consistently good when $M = 5$, regardless of q, while when $M = 1$ and especially in the $S = 200$ case, we can expect to identify around half of the important covariates for either p_{11} or p_{10}. Inference on θ_{11} and θ_{10} is more challenging, with the power to detect important covariates for the latter being lower when the baseline probability of species presence is high, since in that case the potential occurrence of a Stage 1 false positive error is low. The difficulty in inferring important predictors for the probabilities of Stage 1 error, compared to the probability of species presence, is due to the fact that occupancy status of sites remains unchanged throughout Stage 1, while eDNA presence can change between samples. Additionally, inference on species presence is based on data from all sites, while for θ_{11} and θ_{10} only on occupied and unoccupied sites, respectively. Similarly, compared to the probabilities of Stage 2 error, the difficulty in identifying important effects in Stage 1 error probabilities is due to the smaller number of repetitions, since $K = 12$ but M is either equal to 1 or 5 in this case.

As expected, increasing the number of sites, S, is also beneficial, but, it is interesting to note that increasing the number of samples collected from each site, M, is generally at least as beneficial as increasing S, which suggests that from a study design point of

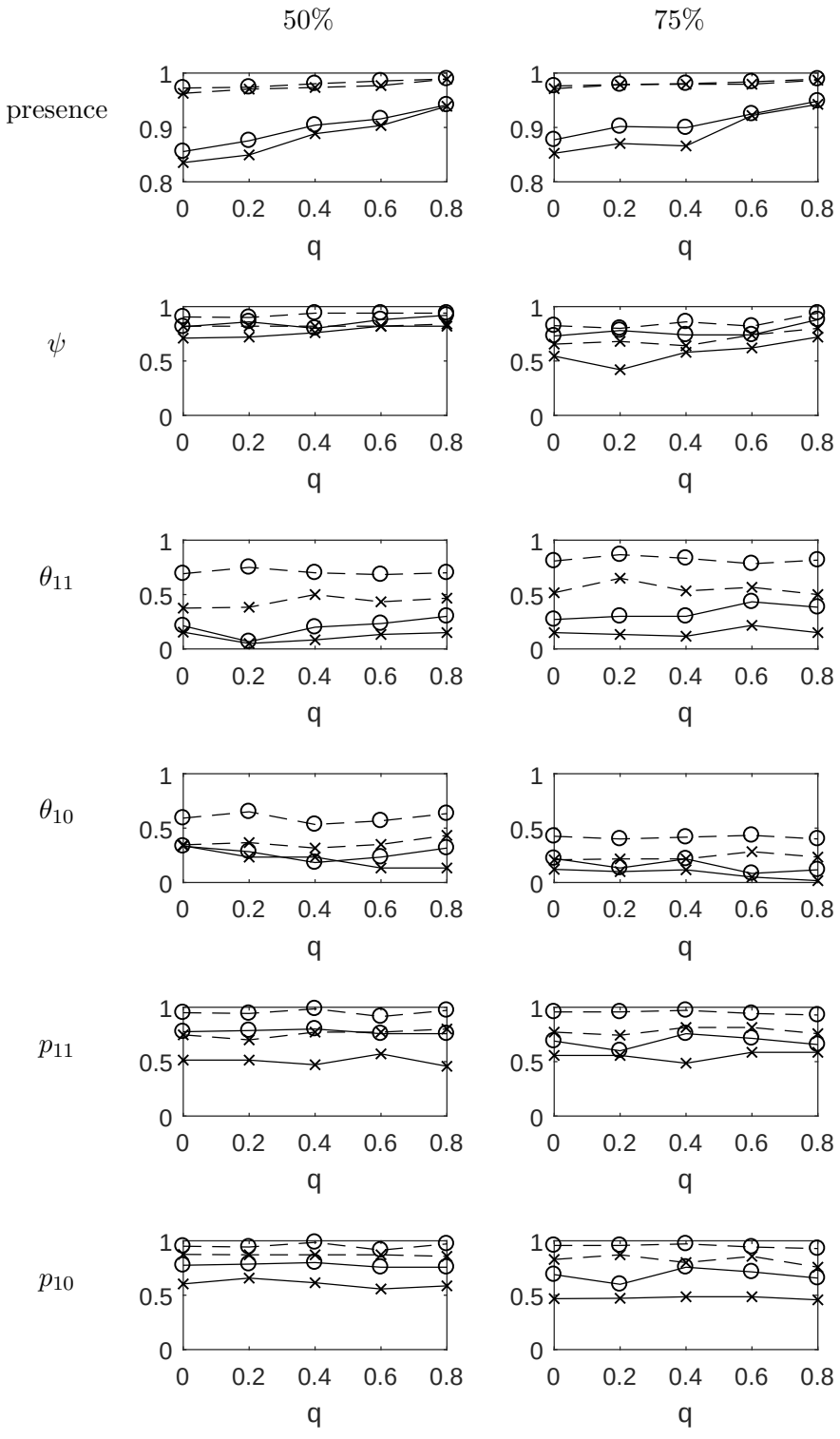


Fig. 3. The average proportion of correctly identified species presences and the average proportion of covariates correctly identified for ψ, θ_{11}, θ_{10}, p_{11}, p_{10} with $M = 1$ (solid lines) or $M = 5$ (dashed lines) and $S = 200$ (crosses) or $S = 500$ (circles). The baseline probability of occupancy is either 50% (left column) or 75% (right column). The proportion of confirmed species presence records in each case is denoted by q.

view, it is preferable to allocate resources to collect more water samples from each site rather than increase the number of sites sampled.

5. Great crested newt eDNA data

Samples were collected as part of a national distribution modelling assessment for great crested newts, commissioned by Natural England (Bormpoudakis et al., 2016). Sample collection and analysis followed a precipitation in ethanol protocol, exactly following those outlined in Biggs et al. (2014, 2015). Twelve quantitative real-time PCR (qPCR) replicates were performed per sample following the assay outlined in Biggs et al. (2014, 2015) using primers TCCBL and TCCBR, with hydrolysis probe TCCB developed by Thomsen et al. (2012). Appropriate positive, negative and inhibition control samples were included. An amplification replicate was considered to be positive if an exponential phase was observed during qPCR.

Surveyors were asked to collect information on additional pond-specific environmental covariates, which we list in Table S3 of the Online Supplementary Material. These pond-specific covariates were predominantly taken from the habitat suitability index for the species (Oldham et al., 2000), widely used for the prediction of the suitability of a pond for the target species, with some covariates, such as terrestrial habitat quality, expanded to give more detail, and some additional covariates, such as pond dimension and water flow also included. Here, we consider all available covariates as potential predictors for species presence as well as the probabilities of error at the two stages.

The following choices of hyperparameters were used. The prior variance of the intercept, ϕ_{μ}^{ψ}, was set to 4 and the prior variance of the regression coefficients, ϕ_{β}^{ψ}, was set to 0.25. The prior distribution of the intercept reflects a belief that the probability of species presence is roughly uniformly distributed and the prior on the regression coefficients represents a belief that the regression effects will be in $(-1, 1)$ with high probability. For the hyperparameters in Stage 1, we used $\mu_0^{\theta_{11}} = \text{logit}(0.8)$, $\mu_0^{\theta_{10}} = \text{logit}(0.2)$, $\epsilon = 0.025$, $r_0^{\theta_{11}} = 1$ and $r_0^{\theta_{10}} = 1$, while in Stage 2 we used $\mu^{p_{11}} = \text{logit}(0.9)$, $\mu_0^{p_{10}} = \text{logit}(0.1)$, $\epsilon = 0.001$, $r_0^{p_{11}} = 1$ and $r_0^{p_{10}} = 1$. These reflected the prior belief of our collaborators that false positives were more likely at data collection stage and that half the variation

at each level can be explained by the covariates. For all parameters, the prior expected number of included covariates was chosen to be 4.

To understand the potential of eDNA data in this case in discriminating between species presence and absence, we calculated the posterior conditional probability of species absence given x positive qPCR replicates at the modal combination of the available covariates (Table 2). As expected, the posterior probability of absence is high if

Table 2. Posterior conditional probability of species absence given x positive qPCR replicates, $1-\psi(x)$, (first row) and posterior conditional probability of x positive qPCR replicates given species presence, $q(x)$, (second row), at the modal combination of the available covariates.

x	0	1	2	3	4	5	6	7	8	9	10	11	12
$1-\psi(x)$	0.93	0.93	0.93	0.93	0.88	0.58	0.54	0.54	0.53	0.53	0.53	0.53	0.53
$q(x)$	0.159	0.093	0.026	0.005	0.001	0.004	0.014	0.039	0.087	0.151	0.192	0.161	0.069

there is a low number of positives in the sample, and decreases with the number of positives. However, it asymptotes at 53% and, so, even if all qPCR replicates return a positive result, the posterior probability of species presence is (just) below 50%. This is due to the overall low probability of species presence, with the posterior median estimated equal to 14% at the modal combination of available covariates (see Table 3 for posterior summaries of all model parameters at the modal combination of the available covariates). The posterior probability of 0 positive qPCR replicates given species presence is 16% (Table 2), which reflects the number of sites in the sample with confirmed species presence but no positive qPCR results. Specifically, the number of positives with their corresponding frequencies in parentheses are 0(5), 3(1), 5(1), 6(1), 11(1) and 12(6). Hence, there is a clear U-shape pattern, suggesting that the high number of occupied sites with 0 positives is mostly due to a Stage 1 false negative error instead of a Stage 2 error, as also supported by the results in Table 3.

The inference about covariate selection in terms of posterior inclusion probabilities (PIP's) is shown in Fig. S1 of the Online Supplementary Material. The results using our prior distribution are shown as dots with three other hyperparameter settings used to understand the sensitivity of our conclusions to the prior settings (the values are

Table 3. Posterior mean and 95% credible interval for all model parameters at the modal combination of the available covariates.

Parameter	Posterior mean	95% posterior credible interval
ψ	0.14	(0.04, 0.42)
θ_{11}	0.73	(0.45, 0.89)
θ_{10}	0.15	(0.05, 0.27)
p_{11}	0.81	(0.71, 0.90)
p_{10}	0.05	(0.03, 0.07)

given in Table S4 of the Online Supplementary Material). Firstly, we consider results using our prior distribution. We have not identified any covariates that are linked to the probability of species presence, ψ, or to the probabilities of a Stage 1 error, as they all have PIP well below 50%. This is not surprising given our simulation results presented in section 4, which suggested that when $M = 1$ and q is low, as is the case here, the average proportion of important predictors identified for ψ, θ_{11} and θ_{10} is low. On the other hand, four covariates with PIP $> 50\%$ have been identified for p_{11} (maximum pond depth, PIP: 1.00, and pond length, PIP: 0.63, presence of macrophytes, PIP: 0.71 and pond density, PIP: 0.91) and one for p_{10} (fish presence, PIP: 0.97). Summaries of the posterior distributions of the regression coefficients for p_{11} and p_{10} are presented in Fig. S2 of the Online Supplementary Material. Maximum pond depth and presence of macrophytes have a positive effect on Stage 2 true positive probability, while pond length and pond density have a negative effect. Finally, the presence of fish decreases the probability of a Stage 2 false positive result. We offer no ecological explanation for this at this Stage. However, these results are hugely important in an applied context. They suggest to practitioners that samples collected from ponds with low levels of macrophytes or shallow depths may have reduced Stage 2 true positive probabilities. Additionally high densities of fish may reduce the instances of Stage 2 false positive results. It is imperative that practitioners are armed with this information to allow them to take probabilities of error into account when interpreting eDNA data sets.

Considering the other hyperparameter settings, we find that many PIPs are robust to the choice of prior and our BVS results in terms of predictors with PIP > 0.5 are

unchanged in most cases, apart from cases where the PIP is around the 0.5 cut-off point (for example, Macrophytes in the model for ψ and Woodland, Rough Grass, Length and Area in the model for p_{11}).

Fig. S3 of the Online Supplementary Material shows the posterior probabilities that $p_{10} > p_{11}$ or $\theta_{10} > \theta_{11}$ at each site. On average under the prior distribution, these probabilities are 0.001 and 0.025 respectively. Clearly, there are similar average rates under the posterior distribution. There are two sites with high probabilities that $p_{10} > p_{11}$ (sites 77 and 179). In both cases, the posterior means of both p_{10} and p_{11} are extremely low leading to a large amount of "crossing".

6. Conclusions

We presented a general framework for modelling eDNA data using a Bayesian model that provides estimates of species presence and of the probabilities of error in the two Stages of eDNA surveys as functions of covariates. Our novel prior formulation overcomes identifiability issues introduced by symmetries in the likelihood function without the need to augment eDNA data with additional data sets. The use of the Pólya-Gamma sampler allows us to define an efficient MCMC algorithm for posterior inference where the models for all model parameters can be updated using a Metropolis-Hastings step.

Our simulation results demonstrated that using our modelling approach we can correctly assess species presence and identify important predictors for all model parameters using eDNA data, with inference improving when data on verified species presence are also incorporated for some of the sites. Additionally, the results highlighted that the added gain, in terms of identifying important predictors, from increasing the number of visited sites is generally smaller than that obtained from increasing the number of water samples collected from each site.

The probabilities of a false negative error are estimated as higher than we anticipated (posterior means equal to 0.27 and 0.19 in Stage 1 and 2, respectively), while the probability of a Stage 1 false positive error is estimated as three times as high as that in Stage 2 (posterior means equal to 0.15 and 0.05 in Stage 1 and 2, respectively). Therefore, our results clearly show that, like traditional methods, eDNA analyses are subject to

imperfect detection at different stages of the analytical protocol, and this needs to be taken into account when interpreting survey results. Although the model also shows how covariates of detection can be identified, further work is needed from a wider range of sampling sites to explore the ecological drivers of false positive and false negative errors.

Our proposed model can be simplified to the Royle and Link (2006) model where only Stage 2 errors are allowed (by setting the probability of a Stage 1 error equal to 0), and hence it is generally applicable to occupancy studies where the probability of a false species detection is greater than 0. In addition, our model can be extended using standard regression techniques, for example to account for heterogeneity caused by differences in water collection protocols or lab practices.

Acknowledgments

We thank Natural England for allowing us to analyse their eDNA survey data for the purposes of this paper.

References

Barnes, M. A. and Turner, C. R. (2016) The ecology of environmental DNA and implications for conservation genetics. *Conservation Genetics*, **17**, 1–17.

Barnes, M. A., Turner, C. R., Jerde, C. L., Renshaw, M. A., Chadderton, W. L. and Lodge, D. M. (2014) Environmental conditions influence eDNA persistence in aquatic systems. *Environmental Science & Technology*, **48**, 1819–1827.

Biggs, J., Ewald, N., Valentini, A., Gaboriaud, C., Dejean, T., Griffiths, R. A., Foster, J., Wilkinson, J. W., Arnell, A., Brotherton, P. and Williams, P. (2015) Using eDNA to develop a national citizen science-based monitoring programme for the great crested newt (*Triturus cristatus). Biological Conservation*, **183**, 19–28.

Biggs, J., Ewald, N., Valentini, A., Gaboriaud, C., Griffiths, R., Foster, J., Wilkinson, J., Arnett, A., Williams, P. and Dunn, F. (2014) Analytical and methodological development for improved surveillance of the great crested newt. Defra Project WC1067. *Tech. rep.*, Freshwater Habitats Trust, Oxford, UK.

Bormpoudakis, D., Foster, J., Gent, T., Griffiths, R. A., Russell, L., Starnes, T., Tzanopoulos, J. and Wilkinson, J. (2016) Developing models to estimate the occurrence in the English countryside of Great Crested Newts, a protected species under the Habitats Directive. Defra Project WC1108. *Tech. rep.*, DICE, University of Kent, Canterbury, UK. URL: `http://randd.defra.gov.uk/Default.aspx?Menu=Menu&Module=More&Location=None&ProjectID=19272&FromSearch=Y&Publisher=1&SearchText=WC1108&SortString=ProjectCode&SortOrder=Asc&Paging=10`.

Brown, P. J., Vannucci, M. and Fearn, T. (1998) Multivariate Bayesian variable selection and prediction. *Journal of the Royal Statistical Society, B*, **60**, 627–641.

Chang, W., Cheng, J., Allaire, J., Xie, Y. and McPherson, J. (2019) *shiny: Web Application Framework for R*. URL: `https://CRAN.R-project.org/package=shiny`. R package version 1.3.2.

Chipman, H., George, E. I. and McCulloch, R. E. (2001) The practical implementation of Bayesian model selection. In *Model Selection* (ed. P. Lahiri). Hayward.

Cole, D. J., Morgan, B. J. and Titterington, D. (2010) Determining the parametric structure of models. *Mathematical biosciences*, **228**, 16–30.

Fearn, T., Brown, P. J. and Haque, M. S. (1999) Logistic discrimination with many variables. *Revista de la Real Academia de Ciencias, Exactas, Fiscias y Naturales*, **93**, 372–342.

Ficetola, G. F., Miaud, C., Pompanon, F. and Taberlet, P. (2008) Species detection using environmental DNA from water samples. *Biology Letters*, **4**, 423–425.

Ficetola, G. F., Pansu, J., Bonin, A., Coissac, E., Giguet-Covex, C., De Barba, M., Gielly, L., Lopes, C. M., Boyer, F., Pompanon, F. et al. (2015) Replication levels, false presences and the estimation of the presence/absence from eDNA metabarcoding data. *Molecular Ecology Resources*, **15**, 543–556.

George, E. I. and McCulloch, R. E. (1997) Approaches for Bayesian variable selection. *Statistica Sinica*, **7**, 339–373.

Green, P. (1995) Reversible jump MCMC computation and Bayesian model determination. *Biometrika*, **82**, 711.

Grün, B. and Leisch, F. (2008) Finite mixtures of generalized linear regression models. In *Recent advances in linear models and related areas*, 205–230. Springer.

Guillera-Arroita, G., Lahoz-Monfort, J. J., van Rooyen, A. R., Weeks, A. R. and Tingley, R. (2017) Dealing with false-positive and false-negative errors about species occurrence at multiple levels. *Methods in Ecology and Evolution*, **8**, 1081–1091.

Jane, S. F., Wilcox, T. M., McKelvey, K. S., Young, M. K., Schwartz, M. K., Lowe, W. H., Letcher, B. H. and Whiteley, A. R. (2015) Distance, flow and PCR inhibition: eDNA dynamics in two headwater streams. *Molecular Ecology Resources*, **15**, 216–227.

Jerde, C. L., Mahon, A. R., Chadderton, W. L. and Lodge, D. M. (2011) "sight-unseen" detection of rare aquatic species using environmental DNA. *Conservation Letters*, **4**, 150–157.

Ley, E. and Steel, M. F. J. (2009) On the effect of prior assumptions in Bayesian model averaging with applications to growth regression. *Journal of Applied Econometrics*, **24**, 651—-674.

Li, Y. and Clyde, M. (2018) Mixtures of g-priors in generalized linear models. *Journal of the American Statistical Association*, **113**, 1828–1845.

MacKenzie, D. I., Nichols, J. D., Lachman, G. B., Droege, S., Andrew Royle, J. and Langtimm, C. A. (2002) Estimating site occupancy rates when detection probabilities are less than one. *Ecology*, **83**, 2248–2255.

Natural England (2017) Wildlife licensing newsletter March 2017. *Tech. rep.*, Natural England, Peterborough, UK.

O'Hara, R. B. and Sillanpää, M. J. (2009) A review of Bayesian variable selection methods: what, how and which. *Bayesian Analysis*, **4**, 85–117.

Oldham, R., Keeble, J., Swan, M. and Jeffcote, M. (2000) Evaluating the suitability of habitat for the great crested newt (*Triturus cristatus*). *Herpetological Journal*, **10**, 143–156.

Papastamoulis, P. (2016) label.switching: An R package for dealing with the label switching problem in MCMC outputs. *Journal of Statistical Software*, **69**, 1–24.

Polson, N. G., Scott, J. G. and Windle, J. (2013) Bayesian inference for logistic models using Pólya-Gamma latent variables. *Journal of the American Statistical Association*, **108**, 1339–1349.

Rees, H. C., Maddison, B. C., Middleditch, D. J., Patmore, J. R. and Gough, K. C. (2014) REVIEW: The detection of aquatic animal species using environmental DNA–a review of eDNA as a survey tool in ecology. *Journal of Applied Ecology*, **51**, 1450–1459.

Royle, J. A. and Link, W. A. (2006) Generalized site occupancy models allowing for false positive and false negative errors. *Ecology*, **87**, 835–841.

Sigsgaard, E. E., Carl, H., Møller, P. R. and Thomsen, P. F. (2015) Monitoring the near-extinct european weather loach in Denmark based on environmental DNA from water samples. *Biological Conservation*, **183**, 46–52.

Thomsen, P., Kielgast, J., Iversen, L. L., Wiuf, C., Rasmussen, M., Gilbert, M. T. P., Orlando, L. and Willerslev, E. (2012) Monitoring endangered freshwater biodiversity using environmental DNA. *Molecular Ecology*, **21**, 2565–2573.

Tréguier, A., Paillisson, J.-M., Dejean, T., Valentini, A., Schlaepfer, M. A. and Roussel, J.-M. (2014) Environmental DNA surveillance for invertebrate species: advantages and technical limitations to detect invasive crayfish *Procambarus clarkii* in freshwater ponds. *Journal of Applied Ecology*, **51**, 871–879.

Willoughby, J. R., Wijayawardena, B. K., Sundaram, M., Swihart, R. K. and DeWoody, J. A. (2016) The importance of including imperfect detection models in eDNA experimental design. *Molecular Ecology Resources*, **16**, 837–844.

Online Supplementary Material

Royle and Link (2006) mixture model

Each of S sites are visited K times and the data are in vector y, with entry y_s indicating the number of times the species of interest was detected at site s.

The component weights of the mixture model are ψ and $1-\psi$, while the corresponding component-specific parameters are p_{11} and p_{10}, respectively. The likelihood function is given by

$$L(\psi, p|y) = \prod_{s=1}^{S} \left\{ \psi p_{11}^{y_s} (1-p_{11})^{K-y_s} + (1-\psi) p_{10}^{y_s} (1-p_{10})^{K-y_s} \right\}$$

Calculating the prior probability that $p_{11} < p_{10}$

$$\mathbb{P}(p_{11} < p_{10}) = \mathbb{P}(\text{logit}(p_{11}) < \text{logit}(p_{10})) = \mathbb{P}(\text{logit}(p_{11}) - \text{logit}(p_{10}) < 0)$$

Clearly,

$$\text{logit}(p_{11}) - \text{logit}(p_{10}) \sim \text{N}\left(\text{logit}(a) - \text{logit}(b), \frac{(\text{logit}(a) - \text{logit}(b))^2}{\delta^2}\right)$$

and so

$$\mathbb{P}(p_{11} < p_{10}) = \mathbb{P}(\text{logit}(p_{11}) - \text{logit}(p_{10}) < 0) = \Phi(-\delta).$$

Markov chain Monte Carlo algorithm

In this appendix, we give further details of the computational approach discussed in Section 3. Since the model can be represented in terms of five logistic regressions, it is useful to define a consistent notation for the various quantities needed in these logistic regressions. We define vector n of length S and with sth entry given by $n_s = \sum_{m=1}^{M} w_{sm}$. For the generic parameter ξ with $\xi \in \{\psi, \theta_{11}, \theta_{10}, p_{11}, p_{10}\}$, we define X^{ξ} to be the design matrix (including a first column of ones for the intercept), n^{ξ} and y^{ξ} to be the numbers of trials and positive responses for each site, respectively, and Ω^{ξ} to be a diagonal matrix whose entries arise from the Pólya-gamma sampling scheme. The exact forms of all of these quantities are given in Table 4. We define $\kappa^{p_{11}} = (\mu^{p_{11}}, \beta^{p_{11}})$, $\kappa^{p_{10}} = (\mu^{p_{10}}, \beta^{p_{10}})$, $\kappa^{\theta_{11}} = (\mu^{\theta_{11}}, \beta^{\theta_{11}})$, $\kappa^{\theta_{10}} = (\mu^{\theta_{10}}, \beta^{\theta_{10}})$ and $\kappa^{\psi} = (\mu^{\psi}, \beta^{\psi})$. The Gibbs sampler involves updating parameters using the full conditionals for: $(\gamma^{p_{11}}, \kappa^{p_{11}})$, $(\gamma^{p_{10}}, \kappa^{p_{10}})$, $(\gamma^{\theta_{11}}, \kappa^{\theta_{11}})$, $(\gamma^{\psi}, \kappa^{\psi})$, $\omega^{p_{11}}$, $\omega^{p_{10}}$, $\omega^{\theta_{11}}$, $\omega^{\theta_{10}}$, ω^{ψ}, w, z, and π.

Table S1. The forms of various quantities for the five logistic regressions, with $\omega_s^{p_{11}} = \sum_{m=1}^{M} w_{sm}\omega_{sm}^{p}$ and $\omega_s^{p_{10}} = \sum_{m=1}^{M}(1-w_{sm})\omega_{sm}^{p}$.

ξ	X^ξ	n^ξ	y^ξ	Ω^ξ
ψ	X	$\mathbf{1}_S$	z	$\text{diag}(\omega^\psi)$
θ_{11}	$X_{s,\cdot}\|z_s=1$	$M\mathbf{1}_{\sum z_s}$	$n_s\|z_s=1$	$\text{diag}(\omega_s^\theta\|z_s=1)$
θ_{10}	$X_{s,\cdot}\|z_s=0$	$M\mathbf{1}_{S-\sum z_s}$	$n_s\|z_s=0$	$\text{diag}(\omega_s^\theta\|z_s=0)$
p_{11}	$X_{s,\cdot}\|n_s>0$	$n_s\|n_s>0$	$y_s^{p_{11}}=\sum_{m=1}^{M} y_{sm}w_{sm}$	$\text{diag}(\omega_s^{p_{11}}\|n_s>0)$
p_{10}	$X_{s,\cdot}\|n_s<M$	$(M\mathbf{1}_S-n_s)\|n_s<M$	$y_s^{p_{10}}=\sum_{m=1}^{M} y_{sm}(1-w_{sm})$	$\text{diag}(\omega_s^{p_{10}}\|n_s<M)$

Updating $(\gamma^{p_{11}},\kappa^{p_{11}})$, $(\gamma^{p_{10}},\kappa^{p_{10}})$, $(\gamma^{\theta_{11}},\kappa^{\theta_{11}})$, $(\gamma^{\theta_{10}},\kappa^{\theta_{10}})$, $(\gamma^{\psi},\kappa^{\psi})$

We describe the method for updating the model γ^ξ and parameters κ^ξ for the generic regression for parameter ξ for $\xi \in \{\psi,\theta_{11},\theta_{10},p_{11},p_{10}\}$ with a vector of responses y^ξ denoting the number of successes from n^ξ trials with design matrix X^ξ. The parameter γ^ξ is updated integrating over κ^ξ using a standard Add-Delete-Swap Metropolis-Hastings sampler. In this sampler, a proposed value c^ξ is sampled by either: an *Add* move, where j such that $\gamma_j^\xi = 0$ is chosen at random and $c_j^\xi = 1$ and $c_k^\xi = \gamma_k^\xi$ for $k \neq j$, a *Delete* move, where j such that $\gamma_j^\xi = 1$ is chosen at random and $c_j^p = 0$ and $c_k^p = \gamma_k^\xi$ for $k \neq j$, or a *Swap* move, where j such that $\gamma_j^\xi = 0$ is chosen at random and m such that $\gamma_m^\xi = 1$ is chosen at random then $c_j^\xi = 1$, $c_m^\xi = 0$ and $c_k^\xi = \gamma_k^\xi$ for $k \neq j,m$. The proposed value is accepted with the following probabilities

$$\begin{cases} \min\left\{1, \frac{L(c^\xi)}{L(\gamma^\xi)}\frac{D^\xi - d^\xi}{D^\xi - d^\xi - 1 + \frac{D^\xi - \bar{d}^\xi}{\bar{d}^\xi}}\right\} & \text{Add} \\ \min\left\{1, \frac{L(c^\xi)}{L(\gamma_\xi)}\frac{D^\xi - d^\xi + \frac{D^\xi - \bar{d}^\xi}{\bar{d}^\xi}}{D^\xi - d^\xi + 1}\right\} & \text{Delete} \\ \min\left\{1, \frac{L(c^\xi)}{L(\gamma^\xi)}\right\} & \text{Swap} \end{cases}$$

where

$$L(\gamma^\xi) = \frac{|B^\xi|^{1/2}}{|X^{\xi T}\Omega^\xi X^\xi + B^\xi|^{1/2}} \exp\left\{-\frac{1}{2}\left[b^{\xi T}B^\xi b^\xi - \left(X^{\xi T}\kappa^\xi + B^\xi b^\xi\right)^T\left(X^{\xi T}\Omega^\xi X^\xi + B^\xi\right)^{-1}\left(X^{\xi T}\kappa^\xi + B^\xi b^\xi\right)\right]\right\},$$

where b^ξ and B^ξ are the prior mean and precision matrix for ξ The parameters θ^ξ are sampled from their conditional distribution

$$\theta^\xi \sim \mathrm{N}\left(\left(X^{\xi T}\Omega^\xi X^\xi + B^\xi\right)^{-1}\left(X^{\xi T}\kappa^\xi + B^\xi b^\xi\right), \left(X^{\xi T}\Omega^\xi X^\xi + B^\xi\right)^{-1}\right).$$

Updating ω^ψ, ω^θ and ω^p

The full conditional distributions are $\omega_s^\psi \sim \mathrm{PG}(1, |\mu^\psi + X_s^\psi \beta^\psi|)$, $\omega_s^{\theta_{11}} \sim \mathrm{PG}(M, |\mu^{\theta_{11}} + X_s^{\theta_{11}}\beta^{\theta_{11}}|)$ if $z_s = 1$, $\omega_s^{\theta_{10}} \sim \mathrm{PG}(M, |\mu^{\theta_{10}} + X_s^{\theta_{10}}\beta^{\theta_{10}}|)$ if $z_s = 0$, $\omega_{sm}^{p_{11}} \sim \mathrm{PG}(K, |\mu^{p_{11}} + X_s^{p_{11}}\beta^{p_{11}}|)$ if $w_{sm} = 1$, and $\omega_{sm}^{p_{10}} \sim \mathrm{PG}(K, |\mu^{p_{10}} + X_s^{p_{10}}\beta^{p_{10}}|)$ if $w_{sm} = 0$.

Efficient algorithms for simulating Pólya-Gamma random variables are provided in Polson et al. (2013).

Updating z

If $k_s = 1$, then $z_s = 1$. If $k_s = 0$, the full conditional distribution of z_s is

$$p(z_i = 1) = \frac{(1-\pi)\psi_s\theta_{11s}^{w_s}(1-\theta_{11s})^{M-w_s}}{(1-\pi)\psi_s\theta_{11s}^{w_s}(1-\theta_{11s})^{M-w_s} + (1-\psi_s)\theta_{10s}^{w_s}(1-\theta_{10s})^{M-w_s}}.$$

where $w_s = \sum_{m=1}^M w_{s,m}$

Update w

The full conditional distribution of $w_{s,m}$ is

$$p(w_{s,m} = 1) = \frac{\theta_{11s}^{z_s}\theta_{10s}^{1-z_s}p_{11s}^{y_{s,m}}(1-p_{11s})^{K-y_{s,m}}}{\theta_{11s}^{z_s}\theta_{10s}^{1-z_s}p_{11s}^{y_{s,m}}(1-p_{11s})^{K-y_{s,m}} + (1-\theta_{11s})^{z_s}(1-\theta_{10s})^{1-z_s}p_{10s}^{y_{s,m}}(1-p_{10s})^{K-y_{s,m}}}$$

Update π

The full conditional distribution of π is $\mathrm{Be}\left(1 + \sum_{s=1}^S z_s k_s, 1 + \sum_{s=1}^S z_s(1-k_s)\right)$.

Simulated data

The data for the simulated examples were generated from the model in (1). We generated ten data sets with $K = 12$, $M = 5$, $S = 500$ and different values of π. Data sets with fewer sites, fewer Stage 1 replicates or no known species presences can be generated from

these larger data sets. The covariates were generated using the distributions shown in Table 6. The scaling of variables 4 and 5 guarantees that all continuous variables have

Table S2. Distributions of covariates in the simulated data

Variable	Range	Distribution	Variable	Range	Distribution
1	$\{0,1\}$	Be(0.3)	6	$\{0,1\}$	Be(0.8)
2	$\{1,2,3\}$	Mu(0.3, 0.3, 0.4)	7	$(-\infty,\infty)$	N(0, 1)
3	$\{1,2,3,4\}$	Mu(0.3, 0.1, 0.2, 0.4)	8	$\{1,2,3,4\}$	Mu(0.1, 0.1, 0.4, 0.4)
4	$(-2,2)$	$\frac{\sqrt{3}}{2}$U$(-2,2)$	9	$\{1,2,3\}$	Mu(0.3, 0.3, 0.4)
5	$(-3,3)$	$\frac{2}{3\sqrt{3}}$U$(-3,3)$	10	$\{0,1\}$	Be(0.1)

mean 0 and variance 1. We performed two sets of simulations: set 1 has approximately 50% of sites being occupied while set 2 has 75% of sites being occupied.

In set 1,

$$\eta^{\psi} = -2\,\mathrm{I}(X_2 = 2) + \mathrm{I}(X_2 = 3) - X_4 + X_5 + 0.5\,\mathrm{I}(X_8 = 2) - 0.7\,\mathrm{I}(X_8 = 3) + \mathrm{I}(X_8 = 4) - X_{10}$$

and, in set 2,

$$\eta^{\psi} = 2\,\mathrm{I}(X_2 = 2) + \mathrm{I}(X_2 = 3) - X_4 + X_5 + 0.5\,\mathrm{I}(X_8 = 2) + 0.7\,\mathrm{I}(X_8 = 3) + \mathrm{I}(X_8 = 4) - X_{10}.$$

In both data sets, the other parameters had the same linear predictors for the other parameters:

$$\begin{aligned}\eta^{\theta_{11}} = {} & 2.197 + 0.747\,X_1 - 0.374\,\mathrm{I}(X_2 = 2) - 0.747\,\mathrm{I}(X_2 = 3) + 0.075\,\mathrm{I}(X_3 = 2) \\ & + 0.187\,\mathrm{I}(X_3 = 3) + 0.374\,\mathrm{I}(X_3 = 4) + 0.374\,X_4 + 0.374\,X_6 - 0.374\,\mathrm{I}(X_9 = 2) \\ & - 0.747\,\mathrm{I}(X_9 = 3)\end{aligned}$$

$$\begin{aligned}\eta^{\theta_{10}} = {} & -2.197 - 0.792\,X_1 + 0.198\,\mathrm{I}(X_2 = 2) + 0.396\,\mathrm{I}(X_2 = 3) + 0.198\,\mathrm{I}(X_3 = 2) \\ & + 0.396\,\mathrm{I}(X_3 = 3) + 0.792\,\mathrm{I}(X_3 = 4) - 0.198\,X_5 - 0.040\,\mathrm{I}(X_8 = 2) \\ & - 0.158\,\mathrm{I}(X_8 = 3) - 0.277\,\mathrm{I}(X_8 = 4) + 0.792\,X_{10}\end{aligned}$$

$$\begin{aligned}\eta^{p_{11}} = {} & 2.197 + 0.057\,\mathrm{I}(X_3 = 2) + 0.287\,\mathrm{I}(X_3 = 3) + 0.574\,\mathrm{I}(X_3 = 4) - 0.574\,X_4 \\ & + 0.287\,X_5 - 0.287\,X_6 + 1.149\,X_7 + 0.115\,\mathrm{I}(X_9 = 2) + 0.230\,\mathrm{I}(X_9 = 3) + 0.345\,X_{10}\end{aligned}$$

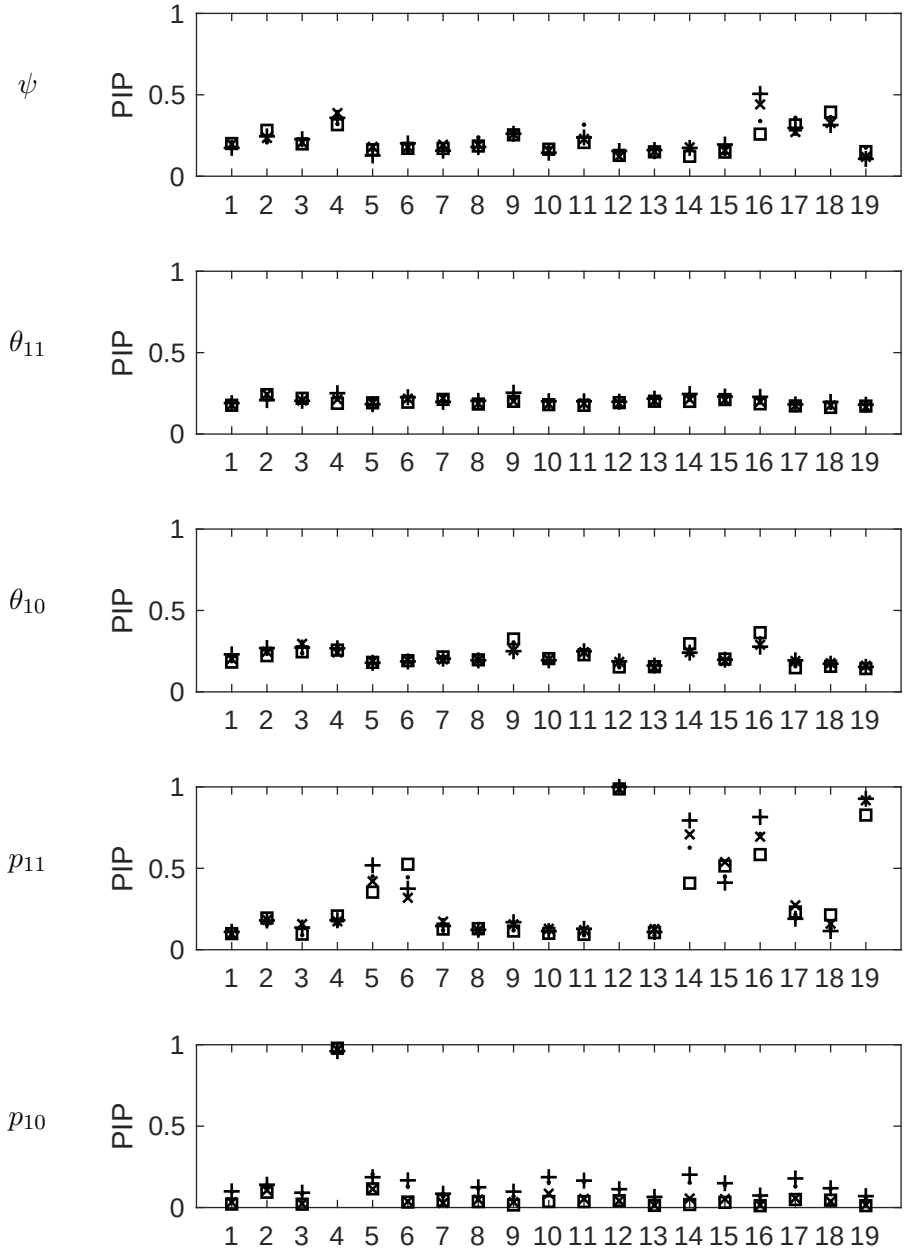
$$\eta^{p_{10}} = -2.197 - 0.044\,\mathrm{I}(X_3 = 2) - 0.218\,\mathrm{I}(X_3 = 3) - 0.435\,\mathrm{I}(X_3 = 4) + 0.870\,X_4 + 0.218\,X_5 - 0.218\,X_6 + 0.435\,X_7 + 0.435\,\mathrm{I}(X_9 = 2) + 0.870\,\mathrm{I}(X_9 = 3) - 0.435\,X_{10}$$

GCN data

Table S3. List and description of pond-specific covariates.

No.	Covariate	Type
1	Permanence	Discrete (Never Dries, (R)arely Dries, (S)ometimes Dries, Dries (A)nnually)
2	Water Quality	Discrete (Bad, (P)oor, (M)oderate, (G)ood)
3	Water Fowl	Discrete (Absent, (Mi)nor, (Ma)jor)
4	Fish	Discrete (Absent, (P)ossible, (Mi)nor, (Ma)jor)
5	Woodland	Discrete (None, (S)ome, (I)mportant)
6	Rough Grass	Discrete (None, (S)ome, (I)mportant)
7	Scrub Hedge	Discrete (None, (S)ome, (I)mportant)
8	Ruderals	Discrete (None, (S)ome, (I)mportant)
9	Inflow	Discrete (Absent, (P)resent)
10	Outflow	Discrete (Absent, (P)resent)
11	Pollution	Discrete (Absent, (P)resent)
12	Max Depth	Continuous
13	Width	Continuous
14	Length	Continuous
15	Area	Continuous
16	Macrophytes	Continuous
17	Overhang	Continuous
18	Shade	Continuous
19	Pond Density	Continuous

Fig. S1. Posterior inclusion probabilities of each covariate for ψ, θ_{11}, θ_{10}, p_{11}, p_{10} using the four hyperparameter settings in Table S4.



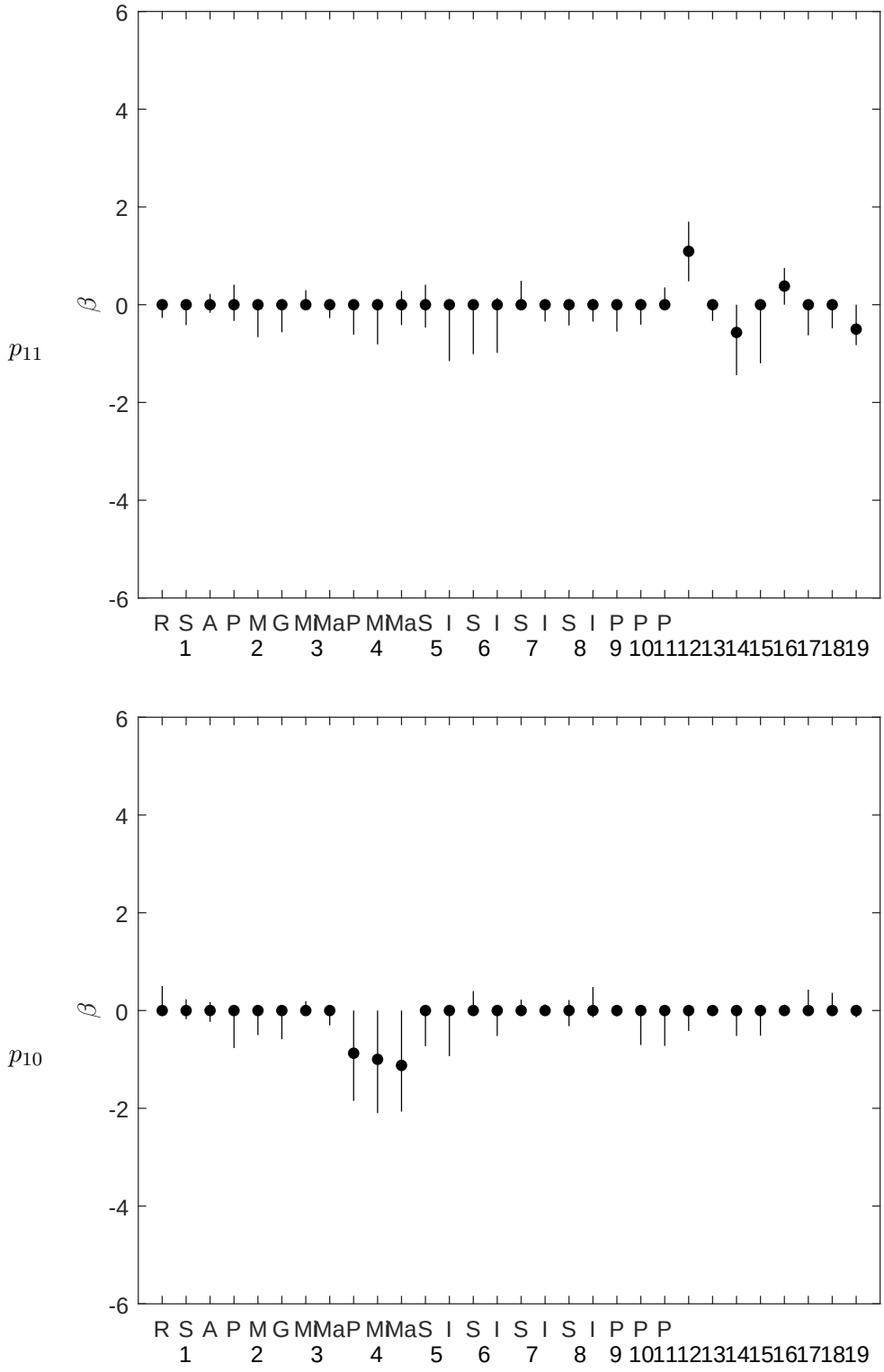


Fig. S2. Inference about the regression coefficients (shown as posterior median and 95% highest probability density region) for p_{11} (top row) and p_{10} (bottom row) with the label of the x-axis showing the covariate number underneath the levels of each covariate as indicated in Table 2.

Table S4. Hyperparameter settings used in the senitivity analysis

Symbol	$\mu_0^{\theta_{11}}$	$\mu_0^{\theta_{10}}$	ϵ (Stage 1)	$\mu_0^{p_{11}}$	$\mu_0^{p_{10}}$	ϵ (Stage 2)
.	$\text{logit}(0.8)$	$\text{logit}(0.2)$	0.025	$\text{logit}(0.9)$	$\text{logit}(0.1)$	0.001
□	$\text{logit}(0.9)$	$\text{logit}(0.1)$	0.001	$\text{logit}(0.9)$	$\text{logit}(0.1)$	0.001
+	$\text{logit}(0.8)$	$\text{logit}(0.2)$	0.025	$\text{logit}(0.8)$	$\text{logit}(0.2)$	0.025
x	$\text{logit}(0.9)$	$\text{logit}(0.1)$	0.001	$\text{logit}(0.8)$	$\text{logit}(0.2)$	0.025

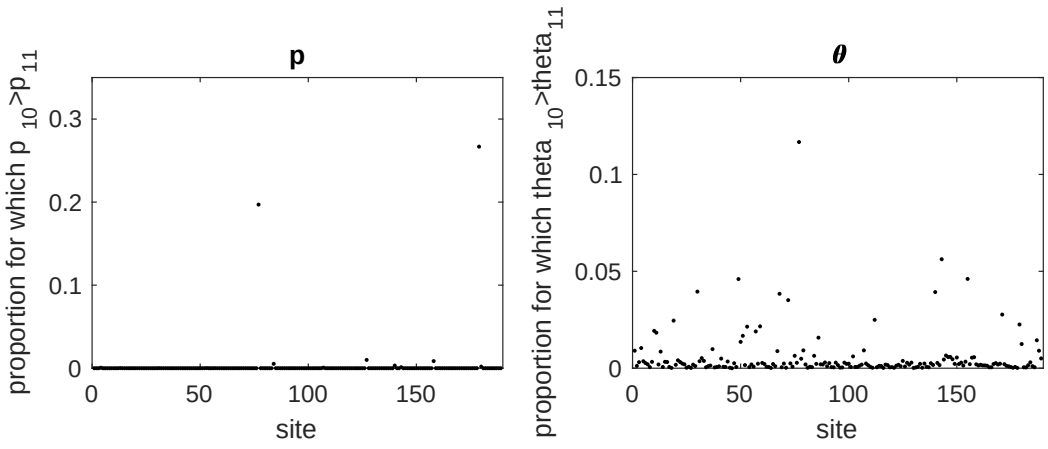


Fig. S3. Posterior probabilities that $p_{10} > p_{11}$ or $\theta_{10} > \theta_{11}$ at each site.