



Kent Academic Repository

Diana, Alex, Matechou, Eleni, Griffin, Jim E., Yu, Douglas W., Luo, Mingjie, Tosa, Marie, Bush, Alex and Griffiths, Richard A. (2024) *eDNAPlus: A unifying modelling framework for DNA-based biodiversity monitoring*. *Journal of the American Statistical Association* . pp. 1-33. ISSN 0162-1459.

Downloaded from

<https://kar.kent.ac.uk/107114/> The University of Kent's Academic Repository KAR

The version of record is available from

<https://doi.org/10.1080/01621459.2024.2412362>

This document version

Author's Accepted Manuscript

DOI for this version

Licence for this version

CC BY (Attribution)

Additional information

Versions of research works

Versions of Record

If this version is the version of record, it is the same as the published version available on the publisher's web site. Cite as the published version.

Author Accepted Manuscripts

If this document is identified as the Author Accepted Manuscript it is the version after peer review but before type setting, copy editing or publisher branding. Cite as Surname, Initial. (Year) 'Title of article'. To be published in **Title of Journal** , Volume and issue numbers [peer-reviewed accepted version]. Available at: DOI or URL (Accessed: date).

Enquiries

If you have questions about this document contact ResearchSupport@kent.ac.uk. Please include the URL of the record in KAR. If you believe that your, or a third party's rights have been compromised through this document please see our [Take Down policy](https://www.kent.ac.uk/guides/kar-the-kent-academic-repository#policies) (available from <https://www.kent.ac.uk/guides/kar-the-kent-academic-repository#policies>).

eDNAPlus: A unifying modelling framework for DNA-based biodiversity monitoring

Alex Diana¹, Eleni Matechou², Jim Griffin³, Douglas W. Yu⁴, Mingjie Luo⁵, Marie Tosa⁶, Alex Bush⁷, and Richard A. Griffiths⁸

¹School of Mathematics, Statistics and Actuarial Science, University of Essex, UK

²School of Mathematics, Statistics and Actuarial Science, University of Kent, UK

³Department of Statistical Science, University College London, UK

⁴School of Biological Sciences, University of East Anglia, UK

⁵Kunming College of Life Sciences, University of Chinese Academy of Sciences, China

⁶Department of Fisheries, Wildlife, Conservation Sciences, Oregon, State University, USA

⁷Lancaster Environment Centre, University of Lancaster, UK

⁸Durrell Institute of Conservation and Ecology, University of Kent, UK

Abstract

DNA-based biodiversity surveys, which involve collecting physical samples from survey sites and assaying them in the laboratory to detect species via their diagnostic DNA sequences, are increasingly being adopted for biodiversity monitoring and decision-making. The most commonly employed method, metabarcoding, combines PCR with high-throughput DNA sequencing to amplify and read ‘DNA barcode’ sequences, generating count data indicating the number of times each DNA barcode was read. However, DNA-based data are noisy and error-prone, with several sources of variation, and cannot alone estimate the species-specific amount of DNA present at a surveyed site (*DNA biomass*). In this paper, we present a unifying modelling framework for DNA-based survey data that allows estimation of *changes in DNA biomass within species, across sites* and their links to environmental covariates, whilst for the first time simultaneously accounting for key sources of variation, error and noise in the data-generating process, and for between-species and between-sites correlation. Bayesian inference is performed using MCMC with Laplace approximations. We describe a re-parameterisation scheme for crossed-effects models designed to improve mixing, and an adaptive approach for updating latent variables, which reduces computation time. Theoretical and simulation results are used to guide study design, including the level of replication at different survey stages and the use of quality control methods. Finally, we demonstrate our new framework on a dataset of Malaise-trap samples, quantifying the effects of elevation and distance-to-road on each species, and produce maps identifying areas of high biodiversity and species DNA biomass.

Keywords: crossed-effects model, environmental DNA, joint species distribution modelling, observation error, occupancy modelling

1 Introduction

Ecology is undergoing a technology revolution that is making it possible to rapidly generate species inventories via automated and high-throughput DNA sequencers and via electronic sensors, such as drones, satellites, camera traps, and acoustic recorders. These techniques can, if coupled with appropriate algorithms and databases, simultaneously identify large numbers of target species, including those that are cryptic, difficult-to-access, tiny, and low-abundance (Bush et al., 2017; Besson et al., 2022; Piper et al., 2019; Ley, 2022). So far, the most efficient method for generating species-resolution inventories is DNA-based surveys, which rely on reading DNA barcodes: short, standardized sections of the genome that can be compared to a reference library to enable taxonomic identifications without the need to examine organism morphologies (Ratnasingham and Hebert, 2007).

DNA barcoding refers to the identification of single species (Hebert et al., 2003), and DNA *metabarcoding* refers to the detection of large numbers of species from environmental DNA (eDNA), which is the collective name for DNA isolated from environmental samples (Taberlet et al., 2018). These environmental samples include water (Thomsen and Willerslev, 2015), soil (Frøslev et al., 2019), air (Clare et al., 2022), and bulk tissue (i.e. mass-trapped organisms) (Ji et al., 2013). For instance, Thomsen and Sigsgaard (2019) demonstrated that traces of eDNA on flower petals could be analysed to describe the diversity of arthropods that visit wildflowers, including pollinators, parasitoids, predators, and herbivores. Ji et al. (2022) used the trace amounts of residual vertebrate blood left in 30,468 blood-sucking leeches to map vertebrate wildlife across a 677 km² nature reserve in China. Finally, Abrego et al. (2021) sequenced 542 mixed-species, bulk-tissue samples of arctic arthropods captured over 14 years and showed that species richness in the study site had declined by 50% during a time period in which local mean temperature had increased by 2C.

The potential of DNA-based surveys for monitoring and managing biodiversity comes with a number of statistical challenges. Firstly, species-specific absolute abundances can-

65 not be estimated using DNA data alone. Secondly, DNA-based surveys yield data that
66 are overdispersed (including zero-inflation) relative to a Poisson distribution due to several
67 types of error and noise (see Section 1.1), some of which are species-specific. The framework
68 presented in this paper addresses these challenges by developing a novel model and corre-
69 sponding efficient inferential tools. Using our framework, we model *within-species change*
70 *in DNA biomass across sites* (described in Section 1.1), which under certain conditions can
71 be considered as a proxy for change in abundance, hence addressing the first challenge. To
72 address the second challenge, we propose a hierarchical crossed-effects model that expresses
73 key sources of variation, error and noise in the data collection and analysis pipeline, whilst
74 accounting for correlation across species and across sites, and for covariate effects on DNA
75 biomass. We also model frequently employed controls at the PCR stage and evaluate their
76 effect on inference.

77 **1.1 DNA-based surveys and associated challenges**

78 Each individual of a species sheds tissue and waste products, and thus its DNA, into
79 the environment. We will refer to this as *DNA biomass*. As we explain in Section 2,
80 the estimates of species DNA biomass obtained from DNA-based surveys alone are only
81 meaningful in comparison between sites, and for that reason, in this paper we focus on
82 modelling *changes in DNA biomass within species, across sites*, referred to as changes in
83 DNA biomass throughout. We achieve this by assuming that the processes are standardised
84 across sites, samples, and PCR replicates and that any differences in the efficiencies of the
85 processes are explained by covariates that can be included in the model. We highlight
86 that, theoretically, the overall amount of DNA biomass for each species is proportional to
87 the species' abundance at that site, but the rate at which each species sheds DNA into
88 the environment is unknown and not estimable using eDNA data alone. Additionally, the
89 relationship between DNA biomass and abundance can vary between species and sites due
90 to environmental conditions, such as DNA degradation rates, and we return to this point
91 in Section 6. Under the assumption that this relationship does not vary with sites then we

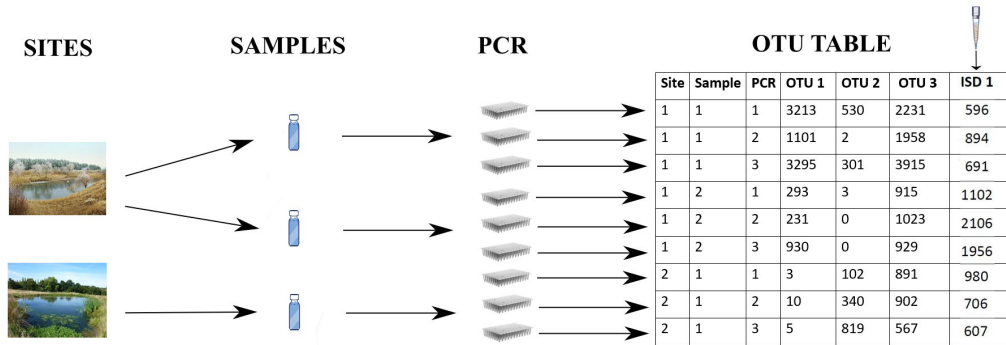


Figure 1: Representation of the DNA biomass collection stage (Stage 1, Sites to Samples) and the DNA biomass analysis stage (Stage 2, Samples to PCR to OTU table). Each of the selected sites to be surveyed hosts a community of species, and hence a certain amount of DNA biomass for each species. One or more physical samples are collected from each surveyed site, and a ‘spike-in’ or ‘internal standard’ ISD, can be added to each sample (last column). Each sample is PCR’d one or more times and then sequenced. This process gives rise to the OTU table.

92 can interpret changes in species DNA biomass as corresponding changes in abundance.

93 DNA-based surveys comprise two stages (Figure 1): the sample collection stage (Stage
 94 1), taking place in the field, and the sample analysis stage (Stage 2), taking place in the
 95 lab.

96 In Stage 1, physical samples are collected from each surveyed site. However, the amount
 97 of DNA biomass of each species collected in each sample is the result of a noisy and error-
 98 prone process (see Table 1). Specifically, the sampling method inevitably favours some
 99 species over others, and as a result, DNA biomass collection rates, conditional on the
 100 available DNA biomasses, are species-specific (*Stage 1 species effect*). The amount of DNA
 101 biomass collected for each species also varies between samples collected at the same site
 102 (*Stage 1 noise*). Finally, there are non-negligible probabilities that (a) no DNA biomass
 103 is collected for a species even if there was DNA biomass of that species at the site (false
 104 negative error) and (b) the DNA biomass in the sample is not the result of species presence,
 105 but instead reflects contamination or deposition from elsewhere (false positive error) (Stage
 106 1 false negative and false positive errors are jointly referred to as *Stage 1 error*).

107 In Stage 2, the physical samples are assayed in the lab. The most frequently used method
 108 for reading DNA barcodes from eDNA samples is ‘amplicon sequencing’ (see Lindahl et al.,

109 2013, for an excellent review). In short, from each sample, all DNA is extracted and purified.
110 After extraction, a small aliquot of DNA from each sample is subjected to Polymerase Chain
111 Reaction (PCR), which selectively amplifies (makes many copies of) just the DNA-barcode
112 sequences. It is common practice in Stage 2 for a sample to be PCR-assayed multiple times,
113 known as technical replicates to distinguish them from sample replicates in Stage 1. The
114 PCR outputs ('amplicons') from all the samples and their technical replicates are pooled
115 and read on a high-throughput DNA sequencer. This procedure ultimately leads to a list
116 of many millions of individual DNA sequences (known as reads), which are processed in a
117 bioinformatic pipeline that removes low-quality reads, groups the remainder into clusters of
118 similar reads that are species hypotheses known as OTUs (Operational Taxonomic Units),
119 and apports each OTU's reads back to its original samples and PCRs. The resulting
120 *OTU table* dataset indicates the number of reads for each OTU in each PCR in each sample
121 in each site (Figure 1), with columns representing the species and rows representing the
122 PCR runs. For simplicity, we hereafter use the terms OTUs and species interchangeably.

123 A real-world complication in DNA-based laboratory pipelines is that samples are typ-
124 ically 'normalised' one or more times. For instance, after the samples are enzymatically
125 digested to break down cells and release their DNA into their 'lysis-buffer' solutions, each
126 sample constitutes a larger volume of liquid than can be used for DNA extraction. The
127 samples are thus normalised by taking a fixed volume from each sample for processing.
128 Another normalisation step happens after PCR, because different PCR replicates can gen-
129 erate different amounts of product. In this case, the PCR products are normalised by
130 taking a certain amount of liquid from each PCR output, either inversely proportional to
131 their concentration, or fixed across PCRs. In the first (lysis buffer) normalisation step,
132 the numerator (amount of lysis buffer taken for extraction) is fixed, while the denominator
133 (total volume of lysis buffer) varies. In the second (PCR product) normalisation step, the
134 numerator (amount of PCR liquid taken for sequencing) varies, while the denominator (to-
135 tal volume of PCR liquid) is fixed. It is standard procedure to record these normalisation

136 fractions, and in Section 2, we show how this information is incorporated into the model.

137 Generally, we should expect a positive relationship between the DNA biomass of a
138 species in a sample and the count of reads obtained for that species in that sample (Luo
139 et al., 2022), but this relationship is imperfect, due to noise and error (see Table 1). First,
140 even given best practice, there are small but non-negligible probabilities (a) that a species’
141 DNA in a sample fails to be amplified or sequenced, leading to false-negative error and (b)
142 that a species’ DNA cross-contaminates other samples and is amplified, leading to false-
143 positive error (Stage 2 false negative and false positive errors are jointly referred to as *Stage*
144 *2 error*). We say that a PCR yields non-negligible reads for a species when the PCR product
145 of that species is successfully read by the DNA sequencer (i.e. the PCR is successful), and
146 otherwise, a PCR yields zero or non-zero but negligible reads, in which case we say that
147 the PCR is not successful for that species. We note that a PCR can be successful, that
148 is, yield non-negligible reads, not only when the biomass is present in the sample but also
149 when it is not, in the latter case because of contamination. Additionally, PCR amplification
150 also inevitably favours some species over others, due to PCR primer mismatch, resulting
151 in species-specific amplification rates (*Stage 2 species effect*, equal within columns of the
152 OTU table), and PCR and sequencing stochasticity results in different total numbers of
153 reads across all species, even for the same sample (*Stage 2 pipeline effect*, equal within rows
154 of the OTU table). Finally, due to the inherent stochasticity of the PCR and sequencing
155 process, there is added noise in the resulting reads *in each cell* of the OTU table (*Stage 2*
156 *noise*).

157 In Stage 2, in addition to recording the normalisation fractions, different approaches
158 are employed to understand and monitor some of the noise and error. One such approach
159 is the so-called internal standard or *spike-in*, during which a known amount of DNA of a
160 synthetic sequence or of a species that is known to be absent from all surveyed sites, is
161 added to each sample. In addition, negative controls, which are samples that are known to
162 not include DNA of any species, can be introduced in Stage 1 and Stage 2 (Ficetola et al.,

163 2015).

Table 1: Description of noise, error, and species/pipeline effects in the two stages of DNA-based surveys.

Stage 1 - DNA biomass collection	
<i>Species effect</i>	Every sample contains a certain amount of DNA biomass of each species, with the amount proportional to the DNA biomass available at the site. However, the proportionality constant is unknown and species-specific, since the DNA of different species can be collected at different rates.
<i>Noise</i>	The amount of DNA biomass collected for each species varies stochastically between samples collected at the same site and time.
<i>Error</i>	It is possible for the DNA of a target species that is present at a site not to be sampled (false negative error), or traces of DNA from one sample to contaminate another sample (false positive error).
Stage 2 - DNA biomass analysis	
<i>Species effect</i>	As a result of differences in gene copy number, DNA extraction efficiency, and PCR amplification efficiency, the correspondence between the source sample DNA biomass and the number of amplicon reads is species-specific (each column of the OTU table).
<i>Pipeline effect</i>	PCR stochasticity and the passing of small aliquots of liquid along the laboratory pipeline affects the total number of reads per technical replicate for all species (each row of the OTU table).
<i>Noise</i>	In addition to the species and pipeline effect, there is added noise in the number of reads per OTU and PCR (each cell of the OTU table).
<i>Error</i>	It is possible for the DNA of a target species that is present in the sample not to be amplified in the lab (false negative error), or traces of DNA of one sample to contaminate and be detected in other samples (false positive error), due to the high species-detection power of amplicon sequencing.

164 1.2 Existing approaches

165 A common approach for modelling metabarcoding data is to convert them to detection/non-
166 detection data by thresholding the number of reads in the OTU table, with user-specified
167 criteria. This allows the use of a generalized linear model (GLM) framework (Saine et al.,
168 2020), which has also been extended to account for species correlation, for example using
169 joint species distribution models (JSDMs) (Ovaskainen and Abrego, 2020). However, this
170 approach does not account for the two stages or the noise and error inherent in DNA-based
171 surveys (Table 1).

172 To that end, several different but related approaches have been proposed. A common
173 approach applies occupancy models that account for false negative observation error to

174 the binary detection/no detection data (Ficetola et al., 2015). More recently, multi-scale
175 extensions of these occupancy models have been proposed to account for false negative
176 error in both stages (Mordecai et al., 2011; Schmidt et al., 2013) and for false positive
177 error (Guillera-Arroita et al., 2017; Griffin et al., 2020) for a single species. However, the
178 occupancy model framework disregards the information in the reads and relies on arbitrary
179 thresholds about what constitutes a detection. Alternatively, the reads have also been
180 modelled within a GLM framework (Takahara et al., 2012; Carraro et al., 2018) but without
181 considering the errors in each stage. A joint model of species occupancy and corresponding
182 reads was developed by Fukaya et al. (2022) but without considering the direct link between
183 species DNA biomass at the site and species reads, or the correlation between species.

184 Finally, we note that an area of research similar to DNA-based biodiversity surveys
185 is microbiome biology, which is the genetic material of all microbial life in an abiotic
186 substrate (e.g. soil) or in a living host (e.g. the human microbiome). When modelling
187 microbiome data, analysis has usually focused on understanding changes in the relative
188 composition of each taxon across different samples. As a result, modelling approaches in
189 this field have revolved around the Dirichlet-Multinomial, which allows inference of the
190 changes, across samples, of the proportions of the species DNA biomasses (Fordyce et al.,
191 2011; Coblenz et al., 2017; McLaren et al., 2019; Clausen and Willis, 2022), although
192 within-species changes in DNA biomass are argued to be informative (Tkacz et al., 2018).
193 A more detailed comparison between the model we introduce in this paper and models for
194 microbiome data is given in Section 2.1.

195 **1.3 Structure of the paper**

196 In this paper, we present a unifying hierarchical modelling framework for OTU reads
197 that considers key sources of variation, noise, and error at both stages of DNA-based
198 biodiversity surveys (Table 1), while also modelling correlation between species and between
199 sites. The model allows us to infer changes in DNA biomass and to link these changes to
200 site-specific covariates.

201 We use state-of-the-art MCMC (Markov chain Monte Carlo) methods that build on
202 recent work for hierarchical and crossed-effects models (Zanella and Roberts, 2021) as well
203 as adaptive MCMC techniques (Andrieu and Thoms, 2008). In particular, we develop a
204 novel sampling technique to improve mixing in the special case of a multivariate crossed-
205 effect model with PCR-specific random effects, and we use adaptive updates of latent
206 variables to focus sampling effort. This allows us to fit our model (with many latent
207 variables across the different stages of DNA surveys) to data from large numbers of sites,
208 samples per site, PCRs per sample, and species.

209 The new model, its properties, and links to existing models are presented in Section
210 2. Details on our approach to inference are given in Section 3. Issues of study design are
211 explored and corresponding simulations are presented in Section 4. A case study of a large
212 Malaise-trap metabarcoding dataset is presented in Section 5, and the paper closes with a
213 discussion in Section 6.

214 2 Model

215 We assume that M_i physical samples are collected from site i , $i = 1, \dots, n$, and K_{im}
216 PCR replicates are performed on the m -th sample from site i . We denote by y_{imk}^s the
217 number of DNA reads of the s -th species, $s = 1, \dots, S$ in the k -th PCR replicate of the
218 m -th sample collected at the i -th site. We have n_z site covariates and X_i^z represents their
219 value at site i and n_w sample covariates, represented as X_{im}^w for the m sample at the i -th
220 site. In what follows, i indexes sites, m samples, k PCR replicates, and s species.

221 Our proposed model (see Figure 2) is hierarchical, with three levels. The first level
222 models the amount of DNA biomass of each species at the surveyed sites, which is a
223 function of environmental and landscape covariates as well as between-species and between-
224 sites correlation (**DNA biomass availability**). The second level models the amount of
225 DNA biomass collected for each species in each physical sample from each site (**DNA**
226 **biomass collection**). Lastly, the third level models the number of reads obtained for
227 each species in each PCR from each physical sample (**DNA biomass analysis**). Data are

228 observed only at the third level, as the result of Stage 2 of the survey, with levels one and
 229 two corresponding to latent states.

DNA biomass availability $L = \{l_i^s\} \sim \text{MN}(B_0 + X_z B, \Sigma, T)$, $T^{-1} \sim \text{GH}$

DNA biomass collection

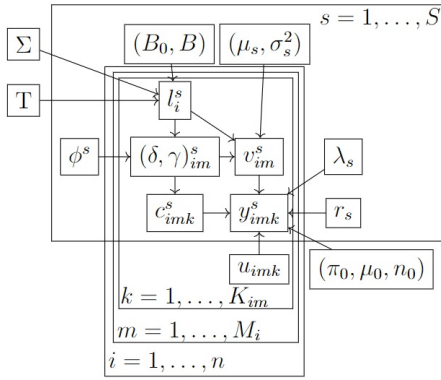
$$\begin{aligned} \text{logit}(\theta_{im}^s) &= \phi_0^s + \phi_1^s l_i^s + X_{im}^w \phi^s \\ \mathbb{P}(\delta_{im}^s = 1) &= \theta_{im}^s, \\ \mathbb{P}(\gamma_{im}^s = 1 \mid \delta_{im}^s = 0) &= \zeta^s, \end{aligned} \quad v_{im}^s \sim \begin{cases} \text{N}(\eta_s + l_i^s + X_{im}^w \beta_s^W, \sigma_s^2) & \text{if } \delta_{im}^s = 1 \\ \text{N}(\mu_s, \nu_s^2) & \text{if } \delta_{im}^s = 0, \gamma_{im}^s = 1 \end{cases}$$

DNA biomass analysis

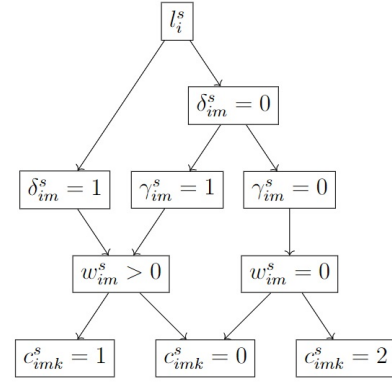
δ_{im}^s	γ_{im}^s	$\mathbb{P}(c_{imk}^s = x \mid \delta_{im}^s, \gamma_{im}^s)$		
		$x = 0$	$x = 1$	$x = 2$
1	—	$1 - p_s$	p_s	0
0	1	$1 - p_s$	p_s	0
0	0	$1 - q_s$	0	q_s

$$y_{imk}^s \sim \begin{cases} \pi \delta_0 + (1 - \pi)(1 + \text{NB}(\mu_0, n_0)) & \text{if } c_{imk}^s = 0 \\ \begin{cases} \text{NB}(\exp(m_{imk}^s), r_s) \\ m_{imk}^s = \lambda_s + v_{im}^s + u_{imk} + o_{imk} \\ u_{imk} \sim \text{N}(0, \sigma_u^2) \end{cases} & \text{if } c_{imk}^s = 1 \\ \text{Pois}(\tilde{\mu}) & \text{if } c_{imk}^s = 2 \end{cases}$$

(a)



(b)



(c)

Figure 2: (a): Model summary, (b): Directed acyclic graph representing the relationships between the variables in the model. (c) Graphical representation of the latent indicator variables in the model.

230 **DNA biomass availability** We denote the logarithm of the amount of DNA biomass of
 231 species s in site i available for collection by l_i^s and denote the $n \times S$ matrix L by $\{L\}_{is} = l_i^s$.
 232 We model DNA biomass correlation between species and spatial correlation between sites
 233 by assuming that L follows a matrix normal distribution, $L \sim \text{MN}(B_0 + X^z B, \Sigma, T)$ (Dawid,
 234 1981), where B_0 is an $n \times S$ matrix with columns $1_n \beta_0^s$, with β_0^s a species-specific intercept,
 235 X^z is a design matrix whose rows are X_i^z , B is an $n_z \times S$ matrix of regression coefficients,
 236 Σ is an $n \times n$ matrix modelling the correlation across sites, and T is an $S \times S$ matrix
 237 modelling the correlation across species. We note that, within this framework, the amount

238 of DNA biomass of a species at the surveyed site cannot be exactly 0, but can be negligible
 239 for modelling purposes as we describe below. We employ a graphical horseshoe (GH) prior
 240 (Li et al., 2019) for the inverse species covariance matrix $Q = T^{-1}$, which is defined by
 241 specifying the following *a priori* independent distributions on each element

$$Q_{ss} \propto \text{Exp}\left(\frac{\lambda}{2}\right), s = 1, \dots, p, \quad Q_{ts} = Q_{st} \sim \text{N}(0, \lambda_{st}^2 \tau^2), \quad \lambda_{st} \sim C^+(0, 1), \quad s < t \leq S$$

242 subject to the constraint $T \in \Omega_S$, where Ω_S is the space of the positive definite $S \times S$
 243 matrices, C^+ represents the half-Cauchy distribution (Gelman, 2006), and $\tau \sim C^+(0, 1)$.
 244 Unlike Li et al. (2019) who specified a flat prior $Q_{ss} \propto 1$, we follow Wang (2012) and
 245 define a proper prior $Q_{ss} \sim \text{Exp}(\frac{\lambda_{GH}}{2})$, ensuring that T , which is latent, has a proper
 246 posterior. We model the spatial correlation matrix Σ using an exponential kernel function,
 247 so that $\Sigma_{i_1 i_2} = \sigma^2 \exp\left\{-\frac{(x_{i_1} - x_{i_2})^2}{l^2}\right\}$, where x_{i_1} and x_{i_2} are the locations of site i_1 and i_2 ,
 248 respectively. We note that we have accounted for species correlations in the DNA biomass
 249 availability stage, but any residual correlations of this type could also be the result of
 250 species correlations in the collection or analysis stages, discussed below. It is not possible,
 251 with metabarcoding data alone, to identify the source of these inferred correlations, and
 252 therefore, species correlations should be interpreted with caution.

253 **DNA biomass collection** We denote by w_{im}^s the amount of DNA biomass of species s
 254 collected in sample m from site i and $v_{im}^s := \log(w_{im}^s)$. To account for *Stage 1 false negative*
 255 *error* at this stage, we introduce the latent variable δ_{im}^s that is equal to 1 if DNA biomass
 256 for species i has been collected in the m -th physical sample from site i , and 0 otherwise.
 257 We assume that $\delta_{im}^s = 1$ with probability θ_{im}^s , which is a function of covariates X_{im}^w , and
 258 of l_i^s , since higher amounts of DNA biomass are expected to lead to a higher probability
 259 of collecting DNA biomass in the sample, leading to $\text{logit}(\theta_{im}^s) = \phi_0^s + \phi_1^s l_i^s + X_{im}^w \phi^s$. We
 260 note that as l_i^s tends to $-\infty$, θ_{im}^s tends to 0, and therefore the species becomes practically
 261 impossible to detect. If the amount of DNA biomass collected is greater than 0 ($\delta_{im}^s = 1$),
 262 we model $v_{im}^s \sim \text{N}(\eta_s + l_i^s + X_{im}^w \beta_s^w, \sigma_s^2)$, where η_s models *Stage 1 species effects* on the
 263 DNA biomass collection rate and σ_s^2 models the species-specific *Stage 1 noise* in the DNA

264 biomass collection rate. To account for *Stage 1 false positive error*, we introduce latent
 265 variable γ_{im}^s , which is equal to 1 with probability ζ^s if the collected DNA biomass is the
 266 result of contamination and 0 otherwise. We assume that γ_{im}^s can be 1 only if $\delta_{im}^s = 0$
 267 and that $v_{im}^s \sim N(\mu_s, \nu_s^2)$ if $\gamma_{im}^s = 1$. In this way, we assume that a sample which already
 268 contains DNA biomass of a species cannot be further contaminated by the DNA of the
 269 same species from another sample or site. We make this assumption as there is not enough
 270 information in the data to partition the collected DNA biomass between that which was
 271 truly collected from the site and that which was contamination from elsewhere.

272 **DNA biomass analysis** As mentioned above, by non-negligible reads we mean that some
 273 of the PCR product is successfully read by the DNA sequencer. We introduce latent variable
 274 c_{imk}^s to model the success of PCR k , sample m , and site i for species s , i.e. *Stage 2 error*.
 275 Firstly, if sample m from site i contains DNA biomass of species s ($w_{im}^s > 0$), PCR run k
 276 can be successful, i.e. non-negligible reads (true positive), $c_{imk}^s = 1$, or not successful, i.e.
 277 negligible reads (false negative), $c_{imk}^s = 0$, and we assume that $c_{imk}^s = 1$ with probability
 278 p_s . We note that we have assumed here that p_s only varies by species and not across sites
 279 or replicates in either stage. However, p_s could depend (negatively) on the total amount
 280 of DNA biomass in the sample, particularly in cases of low DNA concentration for that
 281 species or could vary across primers or between labs. We return to these issues in Section
 282 6. Secondly, if sample m from site i does not contain DNA biomass of species s ($w_{im}^s = 0$),
 283 PCR run k can be successful if it yields non-negligible reads due to lab contamination
 284 (false positive), $c_{imk}^s = 2$, or not successful (again, $c_{imk}^s = 0$, true negative) and assume
 285 that $c_{imk}^s = 2$ with probability q_s .

286 We model the reads conditional on c_{imk}^s as follows. Conditional on $c_{imk}^s = 1$, $y_{imk}^s \sim$
 287 $\text{NB}(\exp(\lambda_s + v_{im}^s + u_{imk} + o_{imk}), r_s)$, where λ_s models the *Stage 2 species effect* on the
 288 amplification rate, u_{imk} is the *Stage 2 pipeline effect*, with $u_{imk} \sim N(0, \sigma_u^2)$, o_{imk} is an
 289 offset modeling the normalisation steps described in Section 1.1, and r_s is a species-specific
 290 variance of the *Stage 2 noise*. If more than one normalisation step is employed, then

291 they can all be incorporated into the same offset as a sum. Conditional on $c_{imk}^s = 0$,
 292 $y_{imk}^s \sim \pi\delta_0 + (1 - \pi)(1 + \text{NB}(\mu_0, n_0))$, that is, there are zero reads with probability π , and
 293 non-zero but negligible reads otherwise. Finally, conditional on $c_{imk}^s = 2$, $y_{imk}^s \sim \text{Pois}(\tilde{\mu}_s)$.
 294 The negative binomial is parameterised in terms of the mean and the number of failures.
 295 A visual representation of the PCR process when $c_{imk}^s = 1$ is shown in Figure 1 of the
 296 Supplementary material.

297 Stage 2 negative control samples (which are known to not contain DNA of any species)
 298 can be easily accounted for in our model by having additional samples for which $\tilde{\delta}_i^s =$
 299 $\tilde{\gamma}_i^s = 0$. Accounting for spike-ins corresponds to having S^* additional species for which
 300 $(v_{im}^{S+1}, \dots, v_{im}^{S+S^*})$ is known. Since the pipeline effect is shared across all species (including
 301 spike-ins), the known values of v_{im}^s for the spike-ins help to better estimate u_{imk} . We further
 302 investigate this effect in Section 4.

303 The model is summarised in Figure 2 (a), the directed acyclic graph of the model is
 304 shown in Figure 2 (b), while a graphical representation of the latent variables introduced
 305 across both stages is shown in Figure 2 (c). The model allows both zero-inflation and
 306 overdispersion (even after accounting for zero-inflation) of the reads. In the case of true
 307 positives (when $c_{imk}^s = 1$), we allow overdispersion through the negative binomial distribu-
 308 tion and the introduction of the offset. The use of negative binomial is a standard choice
 309 for overdispersed data, particularly in Bayesian modelling. Ver Hoef and Boveng (2007)
 310 discuss the merits of negative binomial and quasi-Poisson regression modelling in ecological
 311 data. Datta and Dunson (2016) discuss how a scale mixture of negative-binomial regression
 312 models can be used for so-called quasi-sparse counts, which are often small, not zero.

313 The model presented in Figure 2 is not identifiable in its general form unless certain
 314 constraints are applied, as we discuss below. For example, choosing for simplicity Σ and T
 315 to be diagonal, if we define $\tilde{v}_{im}^s := v_{im}^s - \eta_s - l_i^s$ and $\tilde{l}_i^s := l_i^s - \beta_0^s$, the model for θ_{im}^s and
 316 y_{imk}^s conditional on $c_{imk}^s = 1$ and all offsets o_{imk} set to 0 can be expressed as

$$\begin{cases} \tilde{l}_i^s \sim \text{N}(X_i \beta_s^z, \tau_s^2) \\ \tilde{v}_{im}^s \sim \text{N}(X_{im} \beta_s^w, \sigma_s^2) \\ \theta_{im}^s = \text{logit}(\phi_0^s + \phi_1^s \beta_0^s + \phi_1^s \tilde{l}_i^s + \phi^s X_{im}^s) \\ y_{imk}^s \sim \text{NB}(\exp(\beta_0^s + \tilde{l}_i^s + \eta_s + \tilde{v}_{im}^s + \lambda_s + u_{imk}), r_s) \end{cases} \quad (1)$$

It is evident that the model is invariant to transformations of the form

$$(\beta_0^s)^* = \beta_0^s + c + d, \quad (\lambda_s)^* = \lambda_s - c, \quad (\eta_s)^* = \eta_s - d, \quad (\phi_0^s)^* = \phi_0^s - \phi_1^s(c + d).$$

317 The reason for this unidentifiability is that data are observed only in the third level of
 318 the model, and hence the following sets of species-specific parameters are confounded: the
 319 baseline amount of DNA biomass across all sites (β_0^s) with the baseline collection rate (η_s)
 320 and the baseline amplification rate (λ_s), and the former again with the baseline detection
 321 rate ϕ_0^s . However, by assuming that all these baseline rates are constant across sites,
 322 samples, and PCRs, we are able to infer species-specific *changes* in DNA biomass across
 323 sites and therefore covariate effects.

324 For inferential purposes, we reparameterise the model and set the new baseline (log)
 325 amount of DNA biomass, $(\beta_0^s)^*$, equal to $\beta_0^s + \eta_s$, which means that we can only estimate
 326 the sum of the baseline amount of available DNA biomass and the corresponding baseline
 327 collection rate for the same species. Similarly, we set the new baseline (logit) collection
 328 probability $(\phi_0^s)^*$, equal to $\phi_0^s - \phi_1^s \eta_s$, since the baseline collection probability is also con-
 329 founded with the baseline collection rate (equivalent to setting $\phi_0^s \equiv 0$ and $\eta_s \equiv 0$ in
 330 Equation (1)).

331 As a result, we cannot infer the amount of available DNA biomass separately from the
 332 collection rate, and hence the estimates of log DNA biomass obtained, as mentioned above,
 333 are only meaningful for comparison *within* each species. For the same reason, comparisons
 334 of absolute amount of DNA biomass *across* species are not meaningful. We also note that
 335 depending on the survey design in terms of the number of samples collected per site and
 336 the number of PCR replicates per sample, additional sets of parameters can be confounded
 337 and not estimable. Specifically, the following pairs of parameters are confounded:

- 338 • $S = 1$: pipeline effect u_{imk} and PCR variance r_s ,

- 339 • $K = 1$: PCR variance r_s and sample noise \tilde{v}_{im}^s ,
- 340 • $M = 1$: sample noise \tilde{v}_{im}^s and site noise \tilde{l}_i^s .

341 These are pathological cases that arise when there is no replication at the site/sample/PCR
 342 levels. Replication is vital for being able to account for and to estimate the effects of the
 343 different sources of noise and error (Buxton et al., 2021), an issue to which we return in
 344 Section 4.1. Finally, we note that if the offsets o_{imk} introduced in the model due to the
 345 several normalisations occurring in the pipeline are not recorded, the link between the
 346 amount of DNA biomass in the environment and the reads is broken. However, a potential
 347 way to restore this link is the introduction of spike-ins, which contribute to the estimation
 348 of the “overall” pipeline effects $\tilde{u}_{imk} = u_{imk} + o_{imk}$.

349 2.1 Special cases

350 Two models available in the literature (Section 1.2) arise as special cases of our model.
 351 First, the Dirichlet-Multinomial model (DMM) (Fordyce et al., 2011) is expressed through
 352 the following hierarchy (omitting the indexes m and k to simplify notation):

$$\begin{cases} (y_i^1, \dots, y_i^S) \sim \text{Multi}(N_i, \pi_i^1, \dots, \pi_i^S) \\ (\pi_i^1, \dots, \pi_i^S) \sim \text{Dirichlet}(w\alpha^1, \dots, w\alpha^S) \end{cases} \quad (2)$$

353 where $N_i = \sum_{s=1}^S y_i^s$. The DMM can be seen as a special case of the model described in
 354 Section 2, for the Stage 2 process, conditional on $\delta_i^s = 1$. Specifically, $y_i^s \sim \text{NB}(\exp(\lambda_s +$
 355 $v_i^s + u_i), r_s)$, and therefore, assuming $\lambda_s = u_i = 0$, if $r_s \rightarrow \infty$, the distribution for y_i^s con-
 356 verges to a $\text{Pois}(\exp(v_i^s))$. Conditional on N_i , the model is a $\text{Multi}(N_i, \pi_i^1, \dots, \pi_i^S)$, where
 357 $(\pi_i^1, \dots, \pi_i^S) = \left(\frac{\exp(v_i^1)}{\sum_s \exp(v_i^s)}, \dots, \frac{\exp(v_i^S)}{\sum_s \exp(v_i^s)} \right)$. Next, assuming $\exp(v_i^s) \sim \text{Gamma}(w\alpha_s, \theta)$, we
 358 obtain $(\pi_i^1, \dots, \pi_i^S) \sim \text{Dirichlet}(w\alpha_1, \dots, w\alpha_S)$. Finally, as the DMM does not take errors
 359 into account, the equivalence with our model can be obtained by setting $p_s \equiv 1$.

360 McLaren et al. (2019) propose to account for the Stage 2 species effect in the DMM
 361 framework by modelling the probabilities $(\pi_i^1, \dots, \pi_i^S)$ as $\left(\frac{e^1 \tilde{\pi}_i^1}{\sum_s e^s \tilde{\pi}_i^s}, \dots, \frac{e^S \tilde{\pi}_i^S}{\sum_s e^s \tilde{\pi}_i^s} \right)$, where e_s
 362 models the species-specific efficiencies, which in our model is achieved by using a species-
 363 specific λ_s . The DMM can be extended hierarchically if nested treatments are considered

364 (Coblentz et al., 2017) by defining a nested prior $(\alpha^1, \dots, \alpha^S) \sim \text{Dirichlet}(\alpha_0^1, \dots, \alpha_0^S)$ for
 365 each level. In our model, this is achieved by a hierarchy of normal priors. This highlights
 366 a key difference between the DMM approach and the approach we introduce in this paper,
 367 since we model the propagation of the *absolute* amount of DNA biomass across the different
 368 stages, while the DMM models the propagation of the *relative* amount of DNA biomass.

369 Secondly, the occupancy model of Griffin et al. (2020), in the simple case of no covariates,

$$\begin{cases} z_i \sim \text{Be}(\psi) \\ w_{im} \sim \text{Be}(z_i \xi_1 + (1 - z_i) \xi_0) \\ y_{imk} \sim \text{Be}(w_{im} p + (1 - w_{im}) q) \end{cases} \quad (3)$$

370 designed for (single-species) qPCR, can be seen as a special case of our model when the
 371 information in the counts is not considered. Specifically, letting l_i be binary, with $l_i \in$
 372 $\{-\infty, 0\}$, and defining $z_i = \exp(l_i)$, we obtain $\theta_{im}|(l_i = -\infty) = 0$ and $\theta_{im}|(l_i = 0) =$
 373 $\text{logit}(\phi_0)$. Hence, the model for δ and c becomes

$$\begin{cases} \delta_{im} \sim \text{Be}(z_i(\text{logit}(\phi_0) + (1 - \text{logit}(\phi_0))\zeta) + (1 - z_i)\zeta) \\ c_{imk} \sim \text{Be}(\delta_{im} p + (1 - \delta_{im}) q) \end{cases},$$

374 which is identical to the Griffin et al. (2020) model after defining $\xi_1 = \text{logit}(\phi_0) + (1 -$
 375 $\text{logit}(\phi_0))\zeta$ and $\xi_0 = \zeta$.

376 **3 Inference**

377 Samples can be drawn from the posterior distribution of the parameters using a Gibbs
 378 sampler. Posterior sampling is greatly helped by representing the negative binomial dis-
 379 tribution as a Gamma-Poisson mixture, which allows many parameters to be updated in
 380 closed form from their full conditional distribution.

381 For the parameters σ_s , μ_s , B and B_0 , the full conditional distribution is available in
 382 closed form, and therefore posterior sampling is straightforward. We use simple random
 383 walk Metropolis-Hastings steps for parameters π , μ_0 , n_0 , and r_s and Metropolis-Hastings
 384 steps with a Laplace approximation proposal for the parameters l_i^s , λ_s , v_{im}^s , u_{imk} and r_s .
 385 However, on its own, this naive Gibbs sampler will mix slowly since we have a complex
 386 hierarchical model with crossed-effects and many latent variables. We address this by

387 updating parameters in blocks using re-parameterisation and an adaptive updating scheme
 388 for the discrete latent variables.

389 To illustrate our approach to blocking and re-parameterisation, we consider the error-
 390 free version of our model

$$\begin{cases} l_i^s \sim N(0, \tau_s^2) \\ v_{im}^s \sim N(l_i^s, \sigma_s^2) \\ u_{imk} \sim N(0, \sigma_u^2) \\ y_{imk}^s \sim \text{NB}(\exp(\lambda_s + v_{im}^s + u_{imk}), r_s) \end{cases} \quad (4)$$

391 A naive Gibbs sampler updating each parameter from its full conditional leads to pro-
 392 hibitively slow mixing, due to the form of the likelihood where λ_s , v_{im}^s and u_{imk} appear as
 393 a sum. To address the slow mixing in the nested effects, λ_s and v_{im}^s , the use of a centred
 394 parameterisation (Papaspiliopoulos et al., 2007) has been suggested, which corresponds to
 395 defining $\bar{v}_{im}^s := \lambda_s + v_{im}^s$ and $\bar{l}_i^s := \lambda_s + l_i^s$. However, issues of slow mixing still exist between
 396 \bar{v}_{im}^s and u_{imk} and, as noted by Zanella and Roberts (2021), re-parameterisation does not
 397 improve mixing in the case of crossed-effects models. In a classic crossed-effect model of the
 398 form $y_{jkl} \sim N(\lambda + v_j + u_k, \sigma^2)$, Papaspiliopoulos et al. (2020) propose a collapsed Gibbs sam-
 399 pler by first jointly sampling λ with v_j and then λ jointly with u_k . However, this approach
 400 does not scale well in our setup, since it would involve sampling all the λ_s and u_{imk} jointly,
 401 which have dimensions S and the total number of PCR technical replicates $\sum_{i,m} K_{im}$ re-
 402 spectively. Zanella and Roberts (2021) propose the use of identifiability constraints on
 403 the model, which in Equation (4) correspond to assuming $\sum_s v_{im}^s = \sum_k u_{imk} = 0$. Since
 404 sampling conditionally on constraints can be challenging, we propose a simpler strategy to
 405 improve mixing that is more suited to our framework. We consider re-parameterising to
 406 the *factor averages* $\hat{v}_{im} = \frac{1}{S} \sum_{s=1}^S \bar{v}_{im}^s$ and $\hat{u}_{im} = \frac{1}{K} \sum_{k=1}^K u_{imk}$ and the *factor increments*
 407 $\tilde{v}_{im}^s = \bar{v}_{im}^s - \hat{v}_{im}$ and $\tilde{u}_{imk} = u_{imk} - \hat{u}_{im}$ and performing an update by first sampling jointly
 408 the factor means conditional on the increments, that is, from $(\hat{v}_{im}, \hat{u}_{im} | \tilde{v}_{im}^s, \tilde{u}_{imk})$ and next
 409 using the standard updates $(u_{imk} | v_{im}^1, \dots, v_{im}^S)$ and $(v_{im}^j | u_{im1}, \dots, u_{imK})$. In our simula-
 410 tions, we have found that jointly updating the factor means considerably improves mixing.

411 The sampling scheme for the complete model is presented in the Supplementary material.

412 The indicator variables $(\delta_{im}^s, \gamma_{im}^s, c_{imk}^s)$ can be updated directly from their full condi-
413 tional distributions but, since there are $nMS(\bar{K} + 2)$ (where \bar{K} is the average number of
414 PCR replicates) of these variables and often one value of $(\delta_{imk}, \gamma_{imk}, c_{imk})$ has probability
415 very close to 1, evaluating every full conditional distribution in every iteration can be very
416 time-consuming and computationally wasteful. Therefore, we use a cheap approximation as
417 a proposal in a Metropolis-Hastings step. Specifically, every B iterations, we update the ap-
418 proximation $\hat{p}((\delta_{im}^s, \gamma_{im}^s, c_{imk}^s) = (\epsilon_1, \epsilon_2, \epsilon_3)) = \frac{1}{T} \sum_{t=1}^T \mathbb{I}((\delta_{im}^s)^{(t)}, (\gamma_{im}^s)^{(t)}, (c_{imk}^s)^{(t)}) = (\epsilon_1, \epsilon_2, \epsilon_3))$,
419 where $(\delta_{im}^s)^{(t)}, (\gamma_{im}^s)^{(t)}, (c_{imk}^s)^{(t)}$ is the value of $(\delta_{im}^s, \gamma_{im}^s, c_{imk}^s)$ at the t -th iteration, $\mathbb{I}(A)$ is
420 the indicator function, which takes the value 1 if A is true and 0 otherwise, and T is the
421 number of current iterations. Using this update scheme, we only need to evaluate the
422 likelihood if the state is proposed to change. If the probability of one state is close to
423 one, the adaptive scheme often proposes the current state, which can be accepted without
424 computation. The adaptive scheme does not affect convergence of the MCMC algorithm
425 since the approximation clearly has diminishing adaptation, and the state space of the
426 indicator variables is discrete (see *e.g.* Roberts and Rosenthal, 2009, for more discussion of
427 conditions for convergence of adaptive MCMC schemes).

428 4 Study design

429 In this section, we use a simplified version of the model to investigate the properties
430 of our modelling approach under different study designs in terms of the number of sites,
431 samples per site, and PCRs per sample, as well as the number of spike-ins. In each section,
432 we consider the estimates of the differences in log DNA biomass, when log DNA biomass
433 is not a function of site-specific covariates (no covariate case), and the estimates of the
434 covariate coefficients when log DNA biomass is a function of a single continuous covariate
435 (regression case). In Section 4.1 we present theoretical results using a continuous version
436 of our model that does not account for error in either stage. In Section 4.2 we fit our model
437 as presented in Section 2 under different scenarios for study design by varying the number

438 of sites, number of samples per site, and number of PCRs per sample. Finally, in Section
 439 4.3, we explore the effect of spike-ins for different levels of noise in each stage of the process
 440 and different study designs.

441 4.1 Theoretical results for a simplified version of the model

442 We consider a normal approximation of the model presented in Section 2, which assumes
 443 no species or site correlations, that both stages are error-free by setting $\theta_{im}^s = p_s = 1$, and
 444 that the variances of the distributions of the noise at each stage are the same across species.
 445 As mentioned in Section 2, the use of spike-ins corresponds to the presence of species in
 446 the sample for which $(v_{im}^{S+1}, \dots, v_{im}^{S+S^*})$ is known. We assume, without loss of generality,
 447 that $v_{im}^{S+j} = 0$ for $j = 1, \dots, S^*$. We have the following proposition.

Proposition 4.1. *Consider the model $\lambda_s \sim N(0, \sigma_\lambda^2)$ for $s = 1, \dots, S + S^*$ and, for
 $i = 1, \dots, n$, $k = 1, \dots, K$, $m = 1, \dots, M$,*

$$u_{imk} \sim N(0, \sigma_u^2), \quad v_{im}^s \begin{cases} \sim N(l_i^s, \sigma^2), & s = 1, \dots, S \\ = 0, & s = S + 1, \dots, S + S^* \end{cases},$$

$$y_{imk}^s \sim N(u_{imk} + \lambda_s + v_{im}^s, \sigma_y^2), \quad s = 1, \dots, S + S^*$$

448 where σ^2 , σ_u^2 and σ_y^2 are known.

(a) If we assume $p(l_i^s) \propto 1$ and $\sigma_\lambda^2 \in (0, \infty)$ is known, then

$$\text{Var}(l_1^s - l_2^s | y) = \frac{1}{M} \left(\sigma^2 + \frac{\sigma_y^2}{K} \left(1 + \frac{\frac{\sigma_y^2}{\sigma_u^2}}{\frac{\sigma_y^2}{\sigma_u^2} S^* + 1} \right) \right). \quad (5)$$

449 (b) If we observe a single covariate $X_i \stackrel{i.i.d.}{\sim} N(0, 1)$ for the i -th site and assume $l_i^s \sim$
 $N(X_i \beta_s, \tau^2)$ with $\sigma_\lambda^2 = \infty$ (i.e. $p(\lambda_s) \propto 1$) and $p(\beta_s) \propto 1$, then

$$\text{Var}(\beta_s | y) = \frac{1}{n-1} \left(\tau^2 + \frac{1}{M} \left(\sigma^2 + \frac{\sigma_y^2}{K} \right) \right) \times (1 + C) \quad (6)$$

450 where $C = \frac{\sigma_u^2}{\sigma_y^2 + (M\tau^2 + \sigma^2)K(1 + S^* \frac{\sigma_y^2}{\sigma_u^2}) + \sigma_u^2(S + S^* - 1)}$.

451 Here σ_y^2 models the variance of the noise in Stage 2, as was the case for r_s in the original
 452 model. Equations (5) and (6) show the contributions of the variances at each stage to the
 453 posterior variance of the corresponding estimates (changes in biomass between sites, on the
 454 log scale, and covariate coefficients, respectively) in this special case.

455 The results for this special case suggest that, for both $\text{Var}(l_1^s - l_2^s|y)$ and $\text{Var}(\beta|y)$,
 456 increasing replication at a given stage decreases the contribution of the error variance at
 457 that stage and all downstream stages. For example, increasing the number of samples M
 458 per site reduces the contribution of the noise variance σ^2 at Stage 1 and at all downstream
 459 stages, i.e. σ_u^2 and σ_v^2 in Stage 2. Whereas, increasing the number of PCR replicates, K ,
 460 only reduces the contribution of the Stage 2 variances (σ_u^2 and σ_v^2). Additionally, the benefit
 461 of the spike-in is greater as the ratio of variances $\frac{\sigma_u^2}{\sigma_v^2}$ increases. Moreover, in the case of
 462 $\text{Var}(\beta|y)$, if σ^2 is much greater than σ_y^2 , the benefit of the spike-in is negligible, as the noise
 463 induced by σ^2 greatly outweighs the noise that can be mitigated via the use of spike-ins.

464 4.2 Simulated results for the full model; varying n , M , and K

465 We turn our attention to the full model in Fig. 2 and again consider two cases: no
 466 covariates and a single covariate, $X_i \sim N(0, 1)$. In the no covariate case, we consider the
 467 model's ability to estimate the correct sign of the difference of species log DNA biomasses
 468 at two sites. We use the Brier score $b(i_1, i_2, s) := (\bar{p}(l_{i_1}^s > l_{i_2}^s) - \delta_{i_1, i_2})^2$, where $\bar{p}(l_{i_1}^s > l_{i_2}^s)$ is
 469 the posterior probability of $l_{i_1}^s > l_{i_2}^s$ and δ_{i_1, i_2} is 1 if the true value of $l_{i_1}^s$ is greater than the
 470 true value of $l_{i_2}^s$ and 0 otherwise. We generate $l_i^s \sim \begin{cases} N(1, \tau_s^2) & i \text{ odd} \\ N(0, \tau_s^2) & i \text{ even} \end{cases}$ which separates
 471 the sites between those with "high" DNA biomass and those with "low" DNA biomass. We
 472 use $S = 40$ species, $n = 300$ sites, $M \in \{1, 2, 3, 4, 5\}$ samples per site and $K \in \{1, 2, 3, 4\}$
 473 PCR replicates. The values of the other parameters are reported in the Supplementary
 474 Material. We have performed 50 replications for each combination of values of the design
 475 parameters, M and K . We report the average $b(i_1, i_2, j)$ spanning i_1 across the sites with
 476 low DNA biomass, i_2 across the sites with high DNA biomass, and s across all species

477 and across the replicates. As expected, the Brier score decreases, and hence the ability to
 478 distinguish between sites with low and high DNA biomass increases, as M and K increase
 479 (Figure 3). However, the benefit of increasing K decreases with M , which highlights the
 480 greater importance of multiple sample replicates per site in Stage 1.

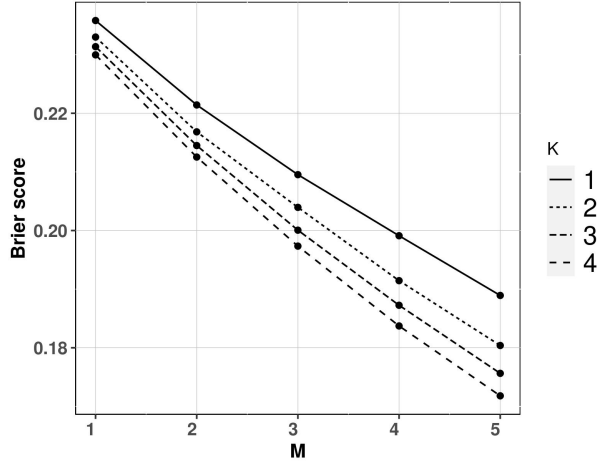


Figure 3: Brier score for distinguishing high and low DNA biomass sites, as a function of the number of samples (M) and number of PCR replicates (K). We have only considered $M \leq 5$, since greater M is unrealistic, and set $n = 300$.

481 In the regression case, we consider the absolute error and posterior standard deviation
 482 of the covariate coefficient β_s . We use $n \in \{50, 100, 200\}$ sites, $M \in \{1, 2, 3\}$ samples per
 483 site and $K \in \{1, 2, 3\}$ PCR replicates per sample and $S = 40$ species. The values of the
 484 other parameters are reported in the Supplementary Material. We performed 50 replicates
 485 for each combination of values of the design parameters and averaged results across all
 486 replicates and species. Results are shown in Figure 4.

487 As expected, absolute error and posterior standard error both decrease with more sites
 488 n , samples per site M , and PCRs per sample K . Doubling the number of sites from 50
 489 to 100 has a bigger effect than doubling them again from 100 to 200, suggesting that the
 490 benefit of increasing the number of sampled sites decreases as the number of sites gets large.
 491 Collecting two samples per site instead of one drastically decreases both absolute error and
 492 posterior standard deviation, whereas the effect is less pronounced when the number of
 493 samples is further increased to three compared to two, and the same can be said about the
 494 number of PCRs.

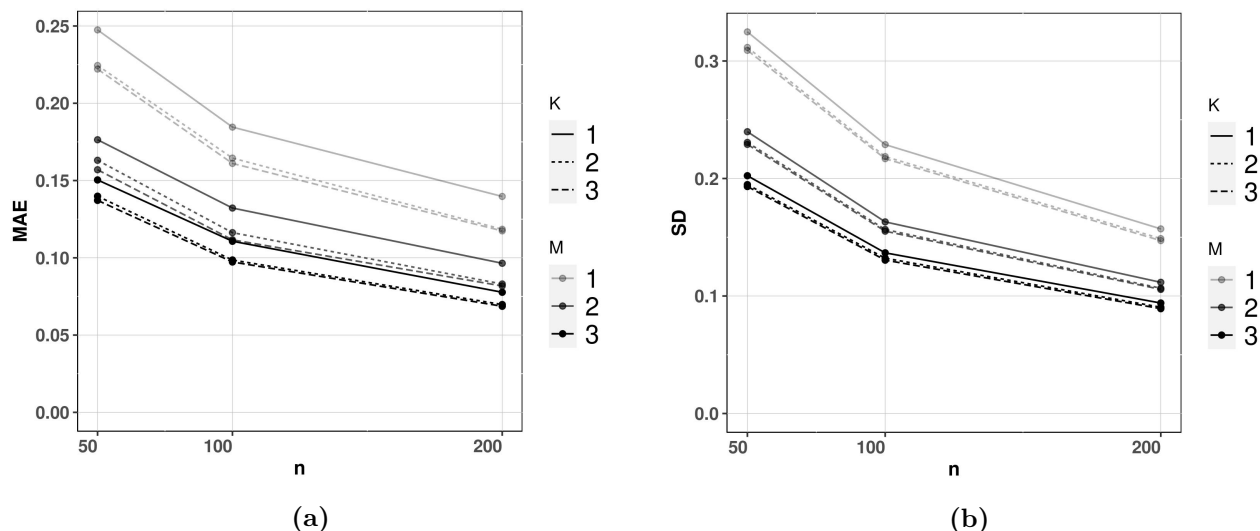


Figure 4: Mean absolute error, (a), and posterior standard deviation, (b), averaged across all species and all simulations, of the covariate coefficient β^s for varying numbers of sites (n), samples per site (M), and numbers of PCR replicates per sample (K).

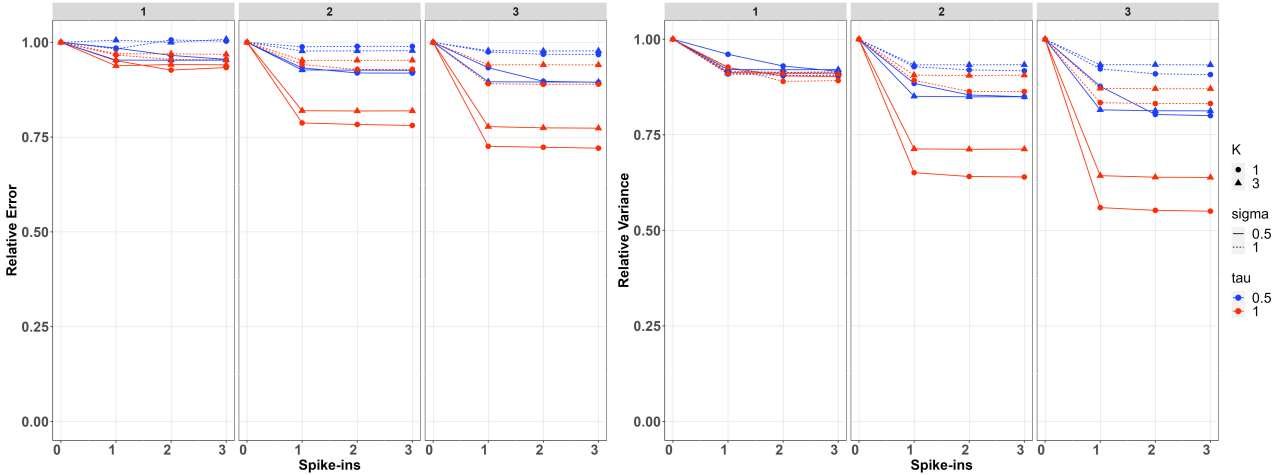
4.3 Spike-ins

In this section, we consider the improvement in inference when S^* spike-ins are employed in Stage 2. The effect of the spike-ins is maximised in the case of no false negative/positive errors, otherwise the benefit of the spike-ins is lower, and dependent upon the level of error. Therefore, in this section we consider data and corresponding model with no false positive/negative errors.

We simulated data on $n = 300$ sites, $M \in \{1, 2, 3\}$ samples per site, and $K \in \{1, 3\}$ PCR replicates per sample on $S = 10$ species. For each setting of M and K , we have fitted the model when $S^* \in \{0, 1, 2, 3\}$ and report in each case the posterior relative error and posterior relative variance of the estimates, which are calculated by dividing the posterior error/variance by the corresponding error/variance when using $S^* = 0$ (which is the case with the greatest error/variance).

Results of the simulation study are presented in Figure 5. In both cases, improvements diminish for $S^* \geq 2$, and in most cases $S^* = 1$ already provides most of the improvement, suggesting that the benefit of more than one spike-ins is minimal. The no covariate case is shown in the first row of Figure 5. Spike-ins contribute more to reducing biomass-change estimation error and variance with $M > 1$, with $M = 1$ resulting in virtually no

No covariate



Regression

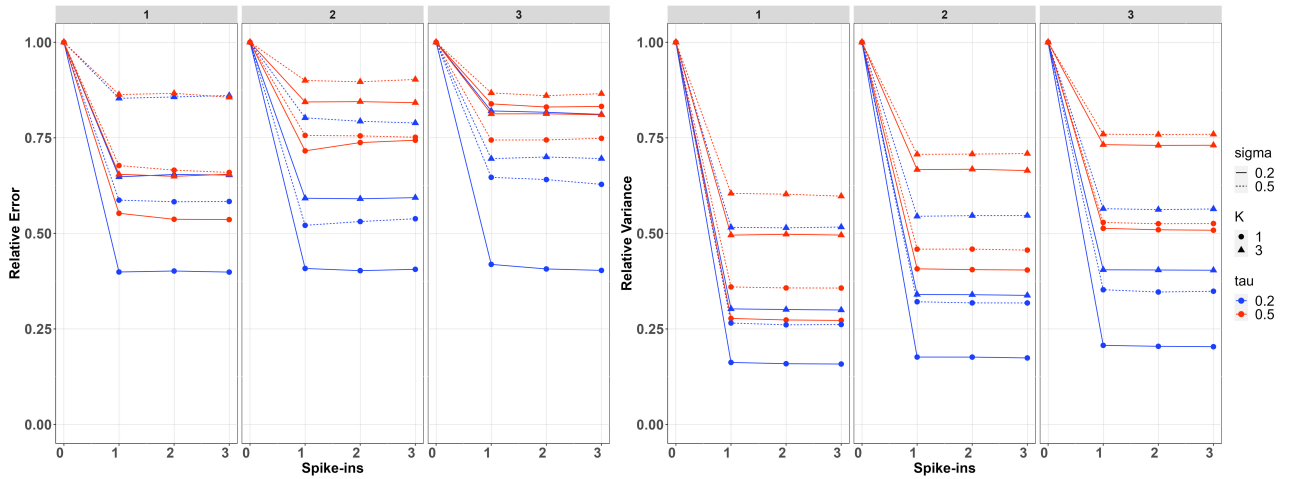


Figure 5: Effect of spike-ins on inference. The three facets per figure represent simulations with $M = 1/2/3$ samples per site. The between-samples standard deviation, σ , is represented by the line type, the between-sites standard deviation, τ , is represented by the color, the number of PCR replicates, K , is represented by the symbols. The first column represents the posterior relative error of the estimates and the second column represents the posterior relative variance.

512 improvements for any setting considered in the simulation. When $M > 1$, improvement
 513 is more pronounced when $K = 1$ instead of $K = 3$, because in the latter case, thanks
 514 to this replication at Stage 2, there is already increased information for estimating the
 515 pipeline effect. This is particularly true when τ is 1 instead of 0.5, because in this case,
 516 the differences between sites are more pronounced. For both values of τ , improvements are
 517 bigger when the between-samples standard deviation (σ) is smaller, since otherwise, Stage
 518 1 noise dominates the process and understanding noise in Stage 2 decreases the overall

519 variance proportionally less.

520 The second row of Figure 5 shows the regression case. We have chosen smaller values
521 for σ and τ (.2 and .5), since the relative contribution of the spike-ins is negligible with
522 larger values. Spike-ins contribute more to reducing error and variance when the between-
523 samples standard deviation (σ) and the between-sites standard deviation (τ) is smaller
524 because, similar to before, the noise at early stages dominates the process, and therefore
525 the relative contribution of the spike-ins is smaller. Also similar to the no covariate case,
526 the contribution of the spike-ins is higher for $K = 1$ PCR replicates compared to $K = 3$.
527 However, unlike that case, the contribution does not appear to increase as the number of
528 samples per site M increases.

529 5 Case study

530 We apply our model to an unpublished amplicon sequencing dataset of arthropod inver-
531 tebrates collected using 121 Malaise-trap samples from 89 sample sites in the H.J. Andrews
532 Experimental Forest (HJA), Oregon, USA (225 km²) in July 2018 (site details are provided
533 in Li et al. 2024). Each trap was left to collect for seven days, and samples were transferred
534 to fresh 100% ethanol to store at room temperature until extraction. The management ob-
535 jective that motivated the collection of this dataset is to interpolate continuous species
536 distributions among the 89 sample points so that areas of higher and lower conservation
537 value at the HJA can be identified.

538 For each sample, the collected invertebrate samples were combined with a lysis buffer,
539 in an amount proportional to the starting sample mass, to digest the tissue, and a fixed
540 aliquot was then taken from the overall mixture (and recorded) for DNA extraction and
541 subsequent three PCRs. This normalization, as described in Section 2, was accounted for
542 in the model by setting the offset o_{imk} equal to the log ratio between the aliquot and the
543 overall amount of liquid mixture in each case. We included 50 species in the study by
544 selecting the species that have the most non-zero counts across all PCR replicates. Log
545 DNA biomass is modelled as a function of two environmental covariates: log elevation and

546 log distance-to-road.

547 Figure 6 presents the 95% posterior credible intervals (PCIs) for the species-specific
548 coefficients of log elevation and log distance-to-road in the model for log DNA biomass. The
549 effects of the covariates on species DNA biomass are not consistent within each taxonomic
550 order, which suggests low phylogenetic inertia at this rank for response to these landscape
551 characteristics. Elevation is a stronger predictor for species DNA biomass than distance-to-
552 road for this ecosystem. This makes ecological sense, since distance-to-road is only expected
553 to exert an effect over about 100 meters, via canopy openness, whereas elevation exerts a
554 pervasive effect via its effects on temperature, precipitation, and vegetation.

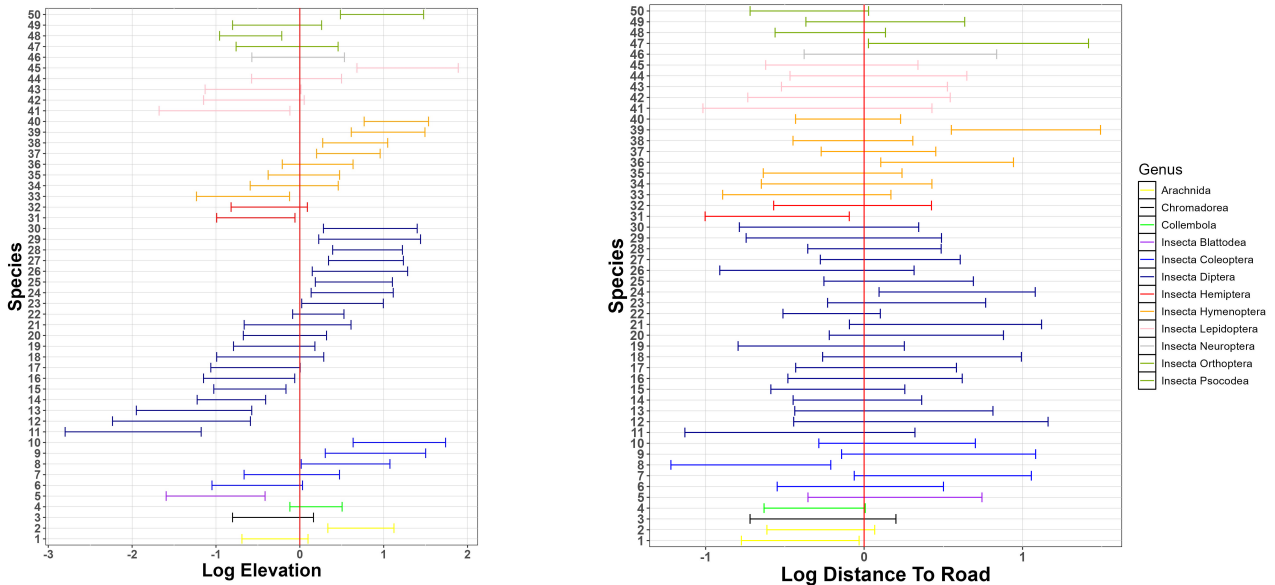


Figure 6: Case study: 95% PCI of the species-specific coefficients of log elevation (left) and distance to road (right) in the model for log DNA biomass. Species are grouped taxonomically.

555 Figure 7 (a) presents the posterior mean of the between-species residual correlations.
556 We set $\lambda_{GH} = 1$ in the GH prior and we emphasize that the GH prior assumes no prior
557 structure imposed on the taxa. Species in the Diptera (flies, spp. 14-30) exhibit higher pos-
558 itive correlations with each other, as well as with several species in the Hymenoptera (ants,
559 bees, and wasps) and Lepidoptera (butterflies and months). We conservatively interpret
560 these positive residual correlations as indicative of unmeasured environmental covariates,
561 such as canopy openness, rather than of biotic interactions. We also note that two species in

562 the Lepidoptera, (spp. 41, 43), one in the Hymenoptera (sp. 33), and one in the Psocodea
 563 (barklice, sp. 50) are among the few species showing strong negative residual correlation
 564 with many of the other species, and again, we conservatively interpret these correlations
 565 as indicative of unmeasured environmental covariates. There is a strongly positive, pair-
 566 wise correlation between two tabanid fly species *Hybomitra liorhina* and *Hybomitra* sp.
 567 (spp. 12, 13), which might indicate the oversplitting of one biological species into two
 568 OTUs during the bioinformatic pipeline. Finally, there is also a strongly positive, pairwise
 569 correlation between the moth species *Ceratodelia gueneata* (sp. 44) and the predatory fly
 570 (Scathophagidae, *Microprosopa* sp., (sp. 20), which might indeed indicate a specialised
 571 predator-prey relationship. All that said, we highlight that these inferred correlations have
 572 been accounted for in the DNA availability stage of the model, but, as we discuss in Section
 573 2, they can also be the result of the DNA biomass collection or analysis stages, so should
 574 be interpreted with caution.

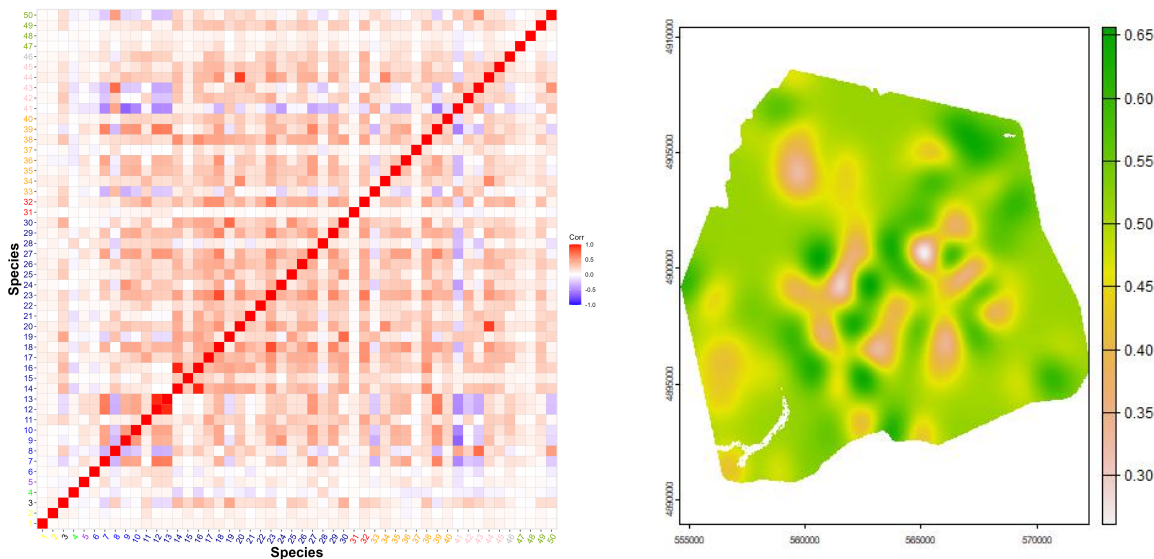


Figure 7: Case study. Left: Correlation plot of all species. Red represents positive correlations while blue represents negative correlations. Species are grouped taxonomically. Right: Posterior mean of biomass-weighted species richness across the study area. For each species, we rescale the log-biomass amount across all study sites into the range $[0, 1]$ and next we compute the species richness as the sum of all the rescaled biomasses across all species.

575 In Figure 7 (b), we show the biodiversity map for the area, which is useful for identifying
 576 areas of higher species richness and compositional distinctiveness, which together can be

577 used to identify areas of higher conservation value (i.e. higher ‘site irreplaceability’ *sensu*
578 Baisero et al., 2022). The predicted mean log DNA biomasses on a continuous map over
579 the HJA for all individual species are presented in the Supplementary Material. These can
580 be used to identify species with a wide spatial range, such as the click beetle (*Megapenthes*
581 *caprella*), or with a restricted range, such as the leafhopper (*Osbornellus borealis*).

582 Finally, Figure 8 (a) suggests that generally, there is a similar amount of variation
583 between sites and between samples for these species. As suggested by Figure 8 (b), the
584 species that we have considered have similar collection probabilities across the several sites,
585 possibly due to the fact that the most frequently detected species across PCRs have been
586 selected. Figure 8 (c) demonstrates, as expected, that the Stage 2 true positive probability
587 is close to 1 for all species. We highlight here that this probability is modelled as species-
588 specific but assumed constant across all replicates. Similarly, the figure also suggests that
589 the probability of a Stage 2 false negative error is very close to 0 for all but three species.
590 One of these three (sp. 14) is in the fly family Tachinidae, which are parasitoids of other
591 insects and thus might have been collected not only as adults but also occasionally as eggs
592 attached to the adults of other (insect) host species, with the latter case being classified as
593 false positives in Stage 2, given that an egg would contribute very low amounts of starting
594 DNA biomass.

595 elfig:cov`results

596 **6 Discussion**

597

598 Over the last decade, DNA-based biodiversity studies, primarily using metabarcoding,
599 have rapidly increased in popularity, and multivariate statistical models are now starting
600 to be deployed to analyse metabarcoding data (e.g. Lin et al., 2021; Pichler and Hartig,
601 2021; Abrego et al., 2021; Fukaya et al., 2022; Ji et al., 2022). Our paper provides the
602 first unifying modelling framework that considers and quantifies key sources of variation,
603 error and noise in metabarcoding surveys (Table 1). As a result, our modelling framework
604 allows more reliable and more powerful biodiversity monitoring and inference on species

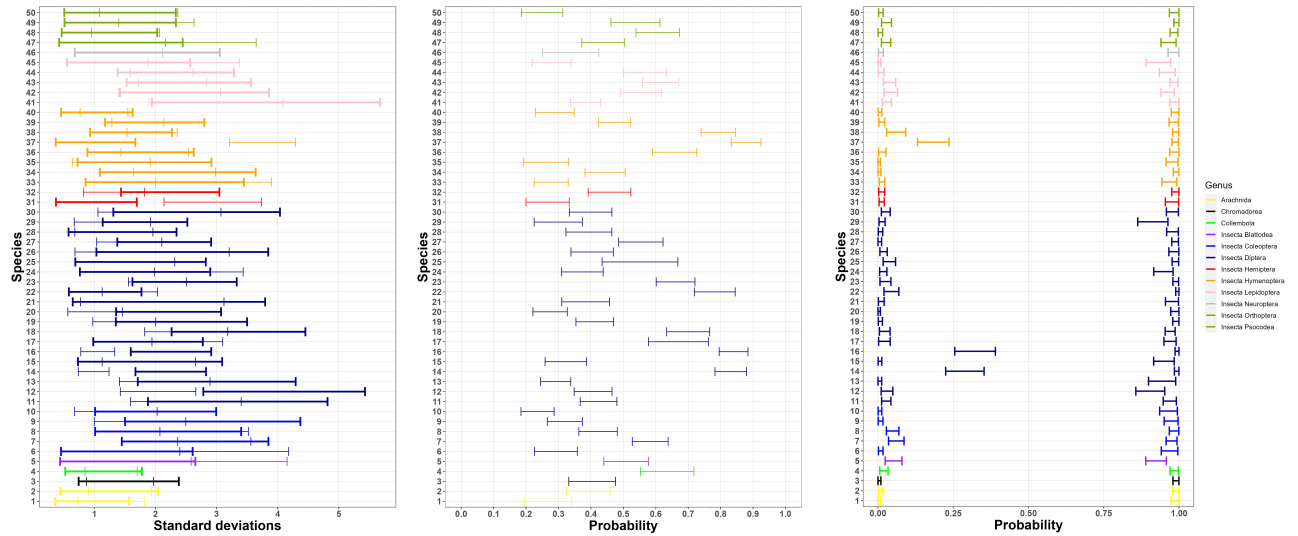


Figure 8: Case study: (left) 95% PCI of the species-specific between-samples standard deviation σ_s and between-sites standard deviation $\sqrt{T_{ss}}$ (in bold). (center) 95% PCI of the species-specific average collection probability θ_{im}^s across all sites. (right) 95% PCI of the species-specific Stage 2 false-positive probabilities q_s (on the left of the plot) and true-positive probabilities p_s (on the right of the plot). Species are grouped taxonomically.

605 responses to landscape characteristics than has been possible before. We have employed,
 606 extended, and developed a number of inferential tools to deal with the complexity of the
 607 proposed hierarchical model, which involves two latent stages and a large number of latent
 608 variables. Finally, this is the first modelling approach that accounts for spike-ins and
 609 negative controls (empty tubes), which are widely used quality-control methods in DNA-
 610 based biodiversity surveys but rarely explicitly considered within a modelling framework.
 611 We explored the benefits of spike-ins on inference and provided analytical and simulation
 612 results of the effects of study design choices on parameter estimates. As is the case in all
 613 models, we make certain assumptions about the data-generating process and if (any of)
 614 these assumptions are violated, then inference can be biased. Below, we discuss the key
 615 assumptions and corresponding model extensions, when appropriate.

616 Our new framework allows us to infer and map species DNA biomass change across
 617 surveyed sites (Figure 7 (b)), and to link these to landscape characteristics (Figure 6).
 618 The resulting maps can be used to identify areas of high conservation value, as well as
 619 areas where particular species or groups of species are more or less prevalent, and to
 620 detect species-specific shifts, expansions, and shrinkage. We are also able to study pairwise

621 correlations across large numbers of species (Figure 7 (a)), which is considerably more
622 scalable using metabarcoding data than using standard observational data. We note that,
623 as discussed in the corresponding sections of the model and the case study, we cannot
624 unambiguously identify the sources of the estimated correlations using the available data
625 alone, as factors other than the affinity between species, such as competition for primers,
626 could affect the inferred species correlations. We have shown that using spike-ins can
627 substantially increase inference accuracy for parameters of interest (Figure 5). Our results
628 also demonstrate that the current practice of collecting a single sample from each surveyed
629 site considerably reduces our ability to infer changes in species DNA biomass and that
630 replication at both stages as well as the use of normalisation-ratio offsets or spike-ins is the
631 optimal approach to designing metabarcoding studies (Figure 3).

632 In metabarcoding data, the baseline DNA biomass of each species is confounded with its
633 species-specific collection and amplification rates. Hence, we cannot infer absolute values
634 of species-specific DNA biomass across sites using metabarcoding data alone. However,
635 by assuming that baseline species-specific collection and amplification rates are the same
636 across sites, samples, and PCR replicates, we can infer species-specific DNA biomass *change*
637 across sites, species-specific covariate effects, and pairwise species correlations. Finally, we
638 model species amplification rates as independent random effects, but competition between
639 species for primers, polymerases and nucleotides during PCR amplification might violate
640 this independence assumption, and future experimental work, alongside model extensions,
641 should explore this issue further.

642 We note that we have not allowed the probability of Stage 2 species detection, p_s , to vary
643 between samples or PCR replicates, and hence we have assumed that it does not depend
644 on the DNA biomass of other species in the sample/PCR replicate. However, because
645 of the PCR product normalisation step, described in Section 1.1, in PCR replicates with
646 relatively high resulting overall DNA biomass, relatively low-DNA-biomass species might
647 be less likely to be drawn in high enough concentration to be detected, an issue that is

648 often referred to as PCR dropout. Empirically, it is known that such PCR competition can
649 be mitigated by using a lower number of PCR cycles (Yang et al., 2021) and by sequencing
650 each sample replicate more deeply. When extending the model of this paper, Stage 2 species
651 detection can be modelled as a function of DNA biomass, so that $\text{logit}(p_{imk}^s) = \beta_0^p + \beta_p(v_{im}^s +$
652 $o_{imk})$. Model extensions of this type are important but are expected to introduce further
653 identifiability issues and computational challenges and hence require careful investigation.

654 Generally, modelling changes in (proxies of) abundance, such as changes in DNA
655 biomass, is a more powerful monitoring tool than modelling changes in species presence
656 across survey sites (Joseph et al., 2006). Metabarcoding studies yield count data without
657 any consequence on associated cost, and hence overcome the time and cost implications
658 associated with collecting count data for multiple species. Our model uses the raw count
659 data, and does not rely on ad-hoc rules about what constitutes a practically zero count for
660 converting them to binary data, which has been the standard practice thus far (Ovaskainen
661 et al., 2017; Bush et al., 2020). To model changes in (log)biomass for each species across
662 sites, we rely on the investigator being able to record any normalisation steps (or to include
663 a spike-in), otherwise the relationship between change in read counts and change in the
664 amount of biomass in the environment cannot be inferred, and instead the counts can only
665 be used to infer composition, as is standard practice in metabarcoding studies. We have
666 allowed for over-dispersion in the count data using a negative binomial distribution, but
667 future work could consider alternative parametrisations, such as the discrete Weibull distri-
668 bution. The model can also be extended to account for multiple primers or for differences
669 between labs, if samples are processed by more than one lab, by introducing regression
670 models for corresponding parameters.

671 Metabarcoding studies, particularly when applied to microbiomes and meiofauna (e.g.
672 nematodes, micro-eukaryotes), can detect 1000s of species, which leads to large numbers of
673 latent variables and coefficients in the model. There are several ways that the inferential
674 tools presented here could be further extended to scale to these cases. Firstly, the posterior

675 distribution conditional on the u_{imk} is independent across species. If u_{imk} could be esti-
676 mated at a first stage then inference across species could be easily parallelized. Secondly,
677 variational Bayes methods could be applied to avoid the use of sampling methods. The
678 choice of variational distribution will be important and can exploit the conditional normal-
679 ity of much of the model. Alternatively, the model could be adapted by assuming that the
680 coefficient matrices such as $\beta^z = (\beta_1^z, \dots, \beta_S^z)$, have a low-dimensional representation. We
681 highlight that in its current format, the model assumes species-specific parameters, and
682 hence there is potentially a large number of parameters to be estimated for each species.
683 Therefore, if a species only has a few non-zero PCR reads from potentially only a few sites,
684 estimating all of these species-specific parameters is difficult. Future work should explore
685 sharing parameters between species, making inference for rarely-observed species possible.

686 We are not modelling species presence/absence and instead we have focused on mod-
687 elling biomass on a continuous scale. As a result, we cannot infer whether a species is
688 absent from a particular study site, but instead only if its DNA biomass at a given site is
689 practically zero. We have assumed that a sample which already contains DNA biomass of
690 a species cannot be further contaminated by the DNA of the same species from another
691 sample or site in Stage 1. This is a reasonable but also necessary assumption, because of
692 model identifiability issues otherwise. It is possible that there exists contamination from
693 other sites if their samples are all processed in the same laboratory, especially at the same
694 time, or that there is contamination during the collection or transfer of samples. However,
695 with only metabarcoding data to hand, it is not possible to identify the source of contami-
696 nation, or to model the possibility that a sample that contains DNA of a species has been
697 further contaminated by the DNA of the same species from another sample or site in Stage
698 1. This is yet another reason to take measures that minimise contamination risk.

699 eDNA metabarcoding has revolutionised the cost-effectiveness, precision, and scale at
700 which biodiversity assessment can be performed. Nevertheless, the multiple stages at which
701 imperfect detection of DNA biomass can occur during the workflow are not insignificant. By

702 facilitating estimates of within-species changes in DNA biomass as a function of covariates,
703 while accounting for workflow uncertainties, our modelling framework provides a substantial
704 improvement in the design and analysis of eDNA metabarcoding data.

705 **Data Availability**

706 The sequence data, bioinformatic scripts, and the three sample by species tables and
707 environmental covariates are archived on DataDryad at doi.org/10.5061/dryad.4f4qrfjjb.

708 **7 Acknowledgments**

709 The work was funded by NERC project NE/T010045/1 “Integrating new statistical
710 frameworks into eDNA survey and analysis at the landscape scale” and benefited from the
711 sCom Working Group at iDiv.de. DWY and MJL were supported by the Strategic Priority
712 Research Program of Chinese Academy of Sciences, Grant No. XDA20050202, the Key
713 Research Program of Frontier Sciences, CAS (QYZDY-SSW-SMC024), the State Key Lab-
714 oratory of Genetic Resources and Evolution (GREKF19-01, GREKF20-01, GREKF21-01)
715 at the Kunming Institute of Zoology, and the University of Chinese Academy of Sciences.

716 **References**

- 717 Abrego, N., Roslin, T., Huotari, T., et al. (2021). Accounting for species interactions is
718 necessary for predicting how arctic arthropod communities respond to climate change.
719 *Ecography*, 44(6):885–896.
- 720 Andrieu, C. and Thoms, J. (2008). A tutorial on adaptive MCMC. *Statistics and Comput-*
721 *ing*, 18(4):343–373.
- 722 Baisero, D., Schuster, R., and Plumptre, A. J. (2022). Redefining and mapping global
723 irreplaceability. *Conservation Biology*, 36(2):e13806.
- 724 Besson, M., Alison, J., Bjerge, K., et al. (2022). Towards the fully automated monitoring
725 of ecological communities. *Ecological Letters*, to appear.
- 726 Bush, A., Monk, W. A., Compson, Z. G., et al. (2020). DNA metabarcoding reveals
727 metacommunity dynamics in a threatened boreal wetland wilderness. *Proceedings of the*
728 *National Academy of Sciences*, 117(15):8539–8545.
- 729 Bush, A., Sollmann, R., Wilting, A., et al. (2017). Connecting Earth observation to high-

throughput biodiversity data. *Nature Ecology & Evolution*, 1(7):0176.

Buxton, A., Matechou, E., Griffin, J., et al. (2021). Optimising sampling and analysis protocols in environmental DNA studies. *Scientific Reports*, 11(1):11637.

Carraro, L., Hartikainen, H., Jokela, J., et al. (2018). Estimating species distribution and abundance in river networks using environmental DNA. *Proceedings of the National Academy of Sciences*, 115(46):11724–11729.

Clare, E. L., Economou, C. K., Bennett, F. J., and others. (2022). Measuring biodiversity from DNA in the air. *Current Biology*, 32(3):693–700.e5.

Clausen, D. S. and Willis, A. D. (2022). Modeling complex measurement error in microbiome experiments. *arXiv preprint arXiv:2204.12733*.

Coblentz, K. E., Rosenblatt, A. E., and Novak, M. (2017). The application of Bayesian hierarchical models to quantify individual diet specialization. *Ecology*, 98(6):1535–1547.

Datta, J. and Dunson, D. B. (2016). Bayesian inference on quasi-sparse count data. *Biometrika*, 103:971—983.

Dawid, A. P. (1981). Some matrix-variate distribution theory: notational considerations and a Bayesian application. *Biometrika*, 68(1):265–274.

Ficetola, G. F., Pansu, J., Bonin, A., et al. (2015). Replication levels, false presences and the estimation of the presence/absence from eDNA metabarcoding data. *Molecular Ecology Resources*, 15(3):543–556.

Fordyce, J. A., Gompert, Z., Forister, M. L., and Nice, C. C. (2011). A hierarchical Bayesian approach to ecological count data: a flexible tool for ecologists. *PloS One*, 6(11):e26785.

Frøslev, T. G., Kjølner, R., Bruun, H. H., et al. (2019). Man against machine: Do fungal fruitbodies and eDNA give similar biodiversity assessments across broad environmental gradients? *Biological Conservation*, 233:201–212.

Fukaya, K., Kondo, N. I., Matsuzaki, S.-i. S., and Kadoya, T. (2022). Multispecies site occupancy modelling and study design for spatially replicated environmental DNA metabarcoding. *Methods in Ecology and Evolution*, 13(1):183–193.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, 1(3):515–533.

Griffin, J. E., Matechou, E., Buxton, A. S., et al. (2020). Modelling environmental DNA data; Bayesian variable selection accounting for false positive and false negative errors. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 69(2):377–392.

Guillera-Arroita, G., Lahoz-Monfort, J., van Rooyen, A., Weeks, A., and Tingley, R. (2017).

763 Dealing with false-positive and false-negative errors about species occurrence at multiple
764 levels. *Methods in Ecology and Evolution*, 8(9):1081–1091.

765 Hebert, P. D., Cywinska, A., Ball, S. L., and DeWaard, J. R. (2003). Biological identifi-
766 cations through DNA barcodes. *Proceedings of the Royal Society of London. Series B:
767 Biological Sciences*, 270(1512):313–321.

768 Ji, Y., Ashton, L., Pedley, S. M., et al. (2013). Reliable, verifiable and efficient monitoring
769 of biodiversity via metabarcoding. *Ecology Letters*, 16(10):1245–1257.

770 Ji, Y., Baker, C. C. M., Popescu, V. D., et al. (2022). Measuring protected-area effectiveness
771 using vertebrate distributions from leech iDNA. *Nature Communications*, 13(1):1555.

772 Joseph, L. N., Field, S. A., Wilcox, C., and Possingham, H. P. (2006). Presence–absence ver-
773 sus abundance data for monitoring threatened species. *Conservation Biology*, 20(6):1679–
774 1687.

775 Ley, R. (2022). The human microbiome: there is much left to do. *Nature*,
776 606(7914):435–435.

777 Li, Y., Craig, B. A., and Bhadra, A. (2019). The graphical horseshoe estimator for inverse
778 covariance matrices. *Journal of Computational and Graphical Statistics*, 28(3):747–757.

779 Li, Y., Devenish, C., Tosa, et al. (2024). Combining environmental dna and remote sensing
780 for efficient, fine-scale mapping of arthropod biodiversity. *Philosophical Transactions of
781 the Royal Society B: Biological Sciences*, 379(1904):20230123.

782 Lin, M., Simons, A. L., et al. (2021). Landscape analyses using edna metabarcoding and
783 earth observation predict community biodiversity in california. *Ecological Applications*,
784 31(6):e02379.

785 Lindahl, B. D., Nilsson, R. H., Tedersoo, L., et al. (2013). Fungal community analysis
786 by high-throughput sequencing of amplified markers—a user’s guide. *New Phytologist*,
787 199(1):288–299.

788 Luo, M., Ji, Y., Warton, D., and Yu, D. W. (2022). Extracting abundance information
789 from DNA-based data. *Molecular Ecology Resources*, to appear.

790 McLaren, M. R., Willis, A. D., and Callahan, B. J. (2019). Consistent and correctable bias
791 in metagenomic sequencing experiments. *Elife*, 8:e46923.

792 Mordecai, R. S., Mattsson, B. J., Tzilkowski, C. J., and Cooper, R. J. (2011). Addressing
793 challenges when studying mobile or episodic species: hierarchical Bayes estimation of
794 occupancy and use. *Journal of Applied Ecology*, 48(1):56–66.

795 Ovaskainen, O. and Abrego, N. (2020). *Joint Species Distribution Modelling: With Appli-*

796 *cations in R*. Cambridge University Press.

797 Ovaskainen, O., Tikhonov, G., Dunson, D., et al. (2017). How are species interactions
798 structured in species-rich communities? A new method for analysing time-series data.
799 *Proceedings of the Royal Society B: Biological Sciences*, 284(1855):20170768.

800 Papaspiliopoulos, O., Roberts, G. O., and Sköld, M. (2007). A general framework for the
801 parametrization of hierarchical models. *Statistical Science*, 22(1):59–73.

802 Papaspiliopoulos, O., Roberts, G. O., and Zanella, G. (2020). Scalable inference for crossed
803 random effects models. *Biometrika*, 107(1):25–40.

804 Pichler, M. and Hartig, F. (2021). A new joint species distribution model for faster and
805 more accurate inference of species associations from big community data. *Methods in*
806 *Ecology and Evolution*.

807 Piper, A. M., Batovska, J., Cogan, N. O. I., et al. (2019). Prospects and challenges of
808 implementing DNA metabarcoding for high-throughput insect surveillance. *GigaScience*,
809 8(8):giz092.

810 Ratnasingham, S. and Hebert, P. D. (2007). Bold: The barcode of life data system
811 (<http://www.barcodinglife.org>). *Molecular Ecology Notes*, 7(3):355–364.

812 Roberts, G. O. and Rosenthal, J. S. (2009). Examples of adaptive MCMC. *Journal of*
813 *Computational and Graphical Statistics*, 18(2):349–367.

814 Saine, S., Ovaskainen, O., Somervuo, P., and Abrego, N. (2020). Data collected by fruit
815 body-and DNA-based survey methods yield consistent species-to-species association net-
816 works in wood-inhabiting fungal communities. *Oikos*, 129(12):1833–1843.

817 Schmidt, B. R., Kery, M., Ursenbacher, S., et al. (2013). Site occupancy models in the
818 analysis of environmental DNA presence/absence surveys: a case study of an emerging
819 amphibian pathogen. *Methods in Ecology and Evolution*, 4(7):646–653.

820 Taberlet, P., Bonin, A., Zinger, L., and Coissac, E. (2018). *Environmental DNA: for*
821 *biodiversity research and monitoring*. Oxford University Press, Oxford, United Kingdom.

822 Takahara, T., Minamoto, T., Yamanaka, H., et al. (2012). Estimation of fish biomass using
823 environmental DNA. *PloS one*, 7(4):e35868.

824 Thomsen, P. F. and Sigsgaard, E. E. (2019). Environmental DNA metabarcoding of wild
825 flowers reveals diverse communities of terrestrial arthropods. *Ecology and Evolution*,
826 9(4):1665–1679.

827 Thomsen, P. F. and Willerslev, E. (2015). Environmental DNA – An emerging tool in con-
828 servation for monitoring past and present biodiversity. *Biological Conservation*, 183:4–18.

- 829 Tkacz, A., Hortala, M., and Poole, P. S. (2018). Absolute quantitation of microbiota
830 abundance in environmental samples. *Microbiome*, 6(1):110.
- 831 Ver Hoef, J. M. and Boveng, P. L. (2007). Quasi-Poisson vs. Negative Binomial Regression:
832 How Should We Model Overdispersed Count Data? *Ecology*, 88:2766–2772.
- 833 Wang, H. (2012). Bayesian graphical lasso models and efficient posterior computation.
834 *Bayesian Analysis*, 7(4):867–886.
- 835 Yang, C., Bohmann, K., Wang, X., and others. (2021). Biodiversity Soup II: A bulk-sample
836 metabarcoding pipeline emphasizing error reduction. *Methods in Ecology and Evolution*,
837 12(7):1252–1264.
- 838 Zanella, G. and Roberts, G. (2021). Multilevel linear models, Gibbs samplers and multigrid
839 decompositions (with discussion). *Bayesian Analysis*, 16(4):1309–1391.