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

Alex Diana¹, Eleni Matechou¹, Jim Griffin², Douglas W. Yu³,⁴,⁵, Mingjie Luo⁴, Marie Tosa⁵, Alex Bush⁶, Richard Griffiths⁷

¹ School of Mathematics, Statistics and Actuarial Science, University of Kent, UK,
² Department of Statistical Science, University College London, UK,
³ School of Biological Sciences, University of East Anglia, UK,
⁴ Kunming College of Life Sciences, University of Chinese Academy of Sciences, China,
⁵ Center for Excellence in Animal Evolution and Genetics &
State Key Laboratory of Genetic Resources and Evolution &
Yunnan Key Laboratory of Biodiversity and Ecological Security of Gaoligong Mountain &
Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming, China,
⁶ Department of Fisheries, Wildlife, & Conservation Sciences, Oregon, State University, USA,
⁷ Lancaster Environment Centre, University of Lancaster, UK,
⁸ Durrell Institute of Conservation and Ecology, University of Kent, UK

Abstract

DNA-based biodiversity surveys involve collecting physical samples from survey sites and assaying the contents in the laboratory to detect species via their diagnostic DNA sequences. DNA-based surveys are increasingly being adopted for biodiversity monitoring and decision-making. The most commonly employed method is metabarcoding, which combines PCR with high-throughput DNA sequencing to amplify and then read ‘DNA barcode’ sequences. This process generates 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. In this paper, we present a unifying modelling framework for DNA-based survey data, for the first time simultaneously allowing for all key sources of variation, error and noise in the data-generating process. The model can be used to estimate within-species biomass changes across sites and to link those changes to environmental covariates, while accounting for between-species and between-sites correlation. Inference is performed using MCMC, where we employ Gibbs or Metropolis-Hastings updates with Laplace approximations. We further implement a re-parameterisation scheme, appropriate for crossed-effects models, leading to improved mixing, and an adaptive approach for updating latent variables, which reduces computation time. We discuss study design and present theoretical and simulation results to guide decisions on replication at different survey stages and on the use of quality control methods. Finally, we demonstrate the new framework on a dataset of Malaise-trap samples. Specifically, we quantify the effects of elevation and distance-to-road on each species, infer species correlations, and produce maps identifying areas of high biodiversity and species biomass, which can be used to rank areas by conservation value. We also estimate the level of noise between sites and within sample replicates, and the probabilities of error at the PCR stage, which
are found to be close to zero for most species considered, validating the employed laboratory processing.

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 (Tosa et al., 2021; van Klink et al., 2022; 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 (Bohmann et al., 2014; Pawlowski et al., 2020; Taberlet et al., 2018). These environmental samples include water (Thomsen and Willerslev, 2015), soil (Frosliev et al., 2019), air (Clare et al., 2022; Lynggaard 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 cannot be estimated using DNA data alone. Secondly, DNA-based surveys yield data that have been subjected to several types of error and noise (see Section 1.1), some of which are species-specific. The framework presented in this paper addresses these challenges by developing a novel model and corresponding efficient inferential tools. Using our framework, we model within-species change in DNA biomasses (described in Section 1.1), which under certain conditions can be considered as a proxy for change in abundance, hence addressing the first challenge. To address the second challenge, we propose a hierarchical crossed-effects model that expresses all major sources of variation, error and noise in the data collection and analysis pipeline, whilst accounting for correlation across species and across sites, and for covariate effects on DNA biomass. We also model frequently employed controls at the PCR stage and evaluate their effect on inference.

1.1 DNA-based surveys and associated challenges

Each individual of a species sheds tissue and waste products, and thus its DNA, into the environment. We will refer to this as DNA biomass or simply biomass. Theoretically, the overall amount of biomass for each species is proportional to the species’ abundance at that site, but the rate at which each species sheds DNA into the environment is unknown and not estimable using eDNA data alone. Since DNA-based surveys target the biomass of each species in the environment, they cannot measure population abundances, and so in this paper we do not refer to species abundance. However, if the relationship between abundance and corresponding biomass is the same across surveyed sites, then changes in a species’ biomass across sites can be interpreted as corresponding changes in species abundances, and we can use the former to monitor the latter. In general, the species-specific relationship between abundance and biomass can vary across sites as a function
of site variation in environmental conditions, such as DNA degradation rates, an issue that we return to in Section 6. Additionally, as we explain in Section 2, the estimates of species biomass obtained from DNA-based surveys alone are not meaningful, unless in comparison between sites, and for that reason, in this paper we focus on modelling changes in within-species biomass across sites. We achieve that by assuming that the processes are standardised across sites, samples, and PCR replicates, and that any differences in the efficiencies of the processes are explained by covariates that can be included in the model.

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

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

In Stage 2, the physical samples are assayed in the lab. The most frequently used method for reading DNA barcodes from eDNA samples is ‘amplicon sequencing’ (see Lindahl et al., 2013 for an excellent review). In short, from each sample, all DNA is extracted and purified. After extraction, a small aliquot of DNA from each sample is subjected to Polymerase Chain Reaction (PCR), which selectively amplifies (makes many copies of) just the DNA-barcode sequences. It is common practice in Stage 2 for a sample to be PCR-assayed multiple
Figure 1: Representation of the biomass collection stage (Stage 1, Sites to Samples) and the biomass analysis stage (Stage 2, Samples to PCR to OTU table). One or more samples is 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.

The PCR outputs (‘amplicons’) from all the samples and their technical replicates are pooled and read on a high-throughput DNA sequencer. This procedure ultimately leads to a list of many millions of individual DNA sequences (known as reads), which are processed in a bioinformatic pipeline that removes low-quality reads, then groups the remainder into clusters of similar reads that are species hypotheses known as OTUs (Operational Taxonomic Units), and apportions each OTU’s reads back to its original samples and PCRs. The resulting OTU table dataset indicates the number of reads for each OTU in each PCR in each sample in each site (Figure 1). For simplicity, we hereafter use the terms OTUs and species interchangeably.

A real-world complication in DNA-based laboratory pipelines is that samples are typically ‘normalised’ one or more times. For instance, after the samples are enzymatically digested to break down cells and release their DNA into their ‘lysis-buffer’ solutions, each sample constitutes a larger volume of liquid than can be used for DNA extraction. The samples are thus normalised by taking a fixed volume from each sample for processing. Another normalisation step happens after PCR, because different PCR replicates can generate different amounts of product. In this case, the PCR products are normalised, by taking
a certain amount of liquid from each PCR output, either inversely proportional to their concentration, or fixed across PCRs. In the first example, the numerator (amount of lysis buffer taken for extraction) is fixed, while the denominator (total volume of lysis buffer) varies. In the second, the numerator (amount of PCR liquid taken for sequencing) varies, while the denominator (total volume of PCR liquid) is fixed. It is standard procedure to record these normalisation fractions, and in Section 2, we show how this information is incorporated into the model.

Generally, we should expect a positive relationship between the biomass of a species in a sample and the count of reads obtained for that species in that sample (Luo et al., 2022), but this relationship is imperfect, due to noise and error (see Table 1). First, even given best practice, there are small but non-negligible probabilities (a) that a species’ DNA in a sample fails to be amplified or sequenced, leading to false-negative error and (b) that a species’ DNA cross-contaminates other samples and is amplified, leading to false-positive error (Stage 2 false negative and false positive error, jointly referred to as Stage 2 error). We say that a PCR yields non-negligible reads for a species when the PCR product of that species is successfully read by the DNA sequencer (i.e. the PCR is successful), and otherwise, a PCR yields non-zero but negligible reads, in which case we say that the PCR is not successful for that species. We note that a PCR can be successful, that is, yield non-negligible reads, not only when the biomass is present in the sample but also when it is not, in the latter case because of contamination. Additionally, PCR amplification also inevitably favours some species over others, due to PCR primer mismatch, resulting in species-specific amplification rates (Stage 2 species effect, equal across rows of the OTU table), and PCR and sequencing stochasticity results in different total numbers of reads across all species, even for the same sample (Stage 2 pipeline effect, equal across columns of the OTU table). Finally, also due to the inherent stochasticity of the PCR and sequencing process, in addition to the species and pipeline effects, there is added noise in the resulting reads in each cell of the OTU table (Stage 2 noise).
In Stage 2, different approaches are employed to understand and monitor some of the noise and error. One such approach is the so-called internal standard or spike-in, during which a known amount of DNA of a synthetic sequence or of a species that is known to be absent from all surveyed sites, is added to each sample. In addition, negative controls, which are samples that are known to not include DNA of any species, can be introduced in Stage 1 and Stage 2 (Ficetola et al., 2015; Goldberg et al., 2016).

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

| **Stage 1 - biomass collection** |
|----------------------------------|
| **Species effect**               | Every sample contains a certain amount of DNA biomass of each species, with the amount proportional to the 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. |
| **Noise**                        | The amount of biomass collected for each species varies stochastically between samples collected at the same site. |
| **Error**                        | 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 - biomass analysis** |
|----------------------------------|
| **Species effect**               | As a result of differences in gene copy number, DNA extraction efficiency, and PCR amplification efficiency, the correspondence between the source sample biomass and the number of amplicon reads is species-specific (column of the OTU table). |
| **Pipeline effect**              | 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 (row of the OTU table). |
| **Noise**                        | 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). |
| **Error**                        | 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. |
1.2 Existing approaches

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

To that end, several different but related approaches have been proposed. A common approach applies occupancy models that account for false negative observation error to the binary presence/absence data (Ficetola et al., 2015). More recently, multi-scale extensions of these occupancy models have been proposed to account for false negative error in both stages (Mordecai et al., 2011; Schmidt et al., 2013) and for false positive error (Guillera-Arroita et al., 2017; Griffin et al., 2020) for a single species. However, the occupancy model framework disregards the information in the reads and relies on arbitrary thresholds about what constitutes a zero read. Alternatively, the reads have also been modelled within a GLM framework (Takahara et al., 2012; Carraro et al., 2018) but without considering the errors in each stage. A joint model of species occupancy and corresponding reads was developed by Fukaya et al. (2022) but without considering the direct link between species biomass at the site and species reads, or the correlation between species.

Finally, we note that an area of research similar to DNA-based biodiversity surveys is microbiome biology, which is the genetic material of all microbial life in an abiotic substrate (e.g. soil) or in a living host (e.g. the human microbiome). When modelling microbiome data, interest usually lies in understanding changes in the relative composition of each taxon across different samples. As a result, modelling approaches in this field have revolved around the Dirichlet-Multinomial, which allows inference of the changes, across samples, of the proportions of the species biomasses, but not changes in absolute biomass.
within each species across samples (Fordyce et al., 2011; Coblentz et al., 2017; McLaren et al., 2019; Clausen and Willis, 2022). A more detailed comparison between the model we introduce in this paper and models for microbiome data is given in section 2.1.

1.3 Structure of the paper

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

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

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

2 Model

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

Our proposed model (see Figure 2) is hierarchical, with three levels. The first level models the amount of DNA biomass of each species at the surveyed sites, which is a function of environmental and landscape covariates as well as between-species and between-sites correlation (Biomass availability). The second level models the amount of biomass collected for each species in each physical sample from each site (Biomass collection). Lastly, the third level models the number of reads obtained for each species in each PCR from each physical sample (Biomass analysis). Data are observed only at the third level, as the result of Stage 2 of the survey, with levels one and two corresponding to latent states.

**Biomass availability** We denote the logarithm of the amount of biomass of species \( s \) in site \( i \) available for collection by \( l_i^s \) and denote the \( n \times S \) matrix \( L \) by \( \{L\}_{is} = l_i^s \). We model biomass correlation between species and spatial correlation between sites by assuming that \( L \) follows a matrix normal distribution, \( L \sim MN(B_0 + X^z B, \Sigma, T) \) (Dawid, 1981), where \( B_0 \) is an \( n \times S \) matrix with columns \( 1_n \beta_0^s \), with \( \beta_0^s \) a species-specific intercept, \( X^z \) is a design matrix whose rows are \( X_i^z \), \( B \) is an \( n_z \times S \) matrix of regression coefficients, \( \Sigma \) is an \( n \times n \) matrix modelling the correlation across sites, and \( T \) is an \( S \times S \) matrix modelling the correlation across species. We note that, within this framework, the amount of biomass of a species at the surveyed site cannot be exactly 0, but can be negligible for modelling purposes as we describe below. We employ a graphical horseshoe (GH) prior (Li et al., 2019) for the inverse species covariance matrix \( Q = T^{-1} \), which has the following form

\[
Q_{ss} \propto \exp \left( \frac{\lambda}{2} \right), \quad s = 1, \ldots, p, \quad Q_{ts} = Q_{st} \sim N(0, \lambda_{st}^2 \tau^2), \quad \lambda_{st} \sim C^+(0, 1), \quad s < t \leq p
\]

where \( C^+ \) represents the half-Cauchy distribution (Gelman, 2006) and \( \tau \sim C^+(0, 1) \). Unlike Li et al. (2019) who specified a flat prior \( T_{ss} \propto 1 \), we follow Wang (2012) and define a proper prior, ensuring that \( T \), which is latent, has a proper posterior. The spatial correlation
Biomass availability \( L = \{l^s_i\} \sim MN(B_0 + XZ, \Sigma, T), \quad T^{-1} \sim GH \)

Biomass collection
\[
\begin{align*}
\text{logit}(\theta^{s}_{im}) &= \phi^{s}_0 + \phi^{s}_1 l^s_i + X^w_{im} \phi^{s}_W \\
\mathbb{P}(\delta^{s}_{im} = 1) &= \theta^{s}_{im}, \\
\mathbb{P}(\gamma^{s}_{im} = 1 | \delta^{s}_{im} = 0) &= \zeta^{s}, \\
\end{align*}
\]

Biomass analysis
\[
\begin{align*}
\delta^{s}_{im} &\sim \left\{ \begin{array}{ll}
N(\eta^{s}_i + l^{s}_i + X^w_{im} \beta^{s}_W, \sigma^{2}_s) & \text{if } \delta^{s}_{im} = 1 \\
N(\mu^{s}_s, \nu^{2}_s) & \text{if } \delta^{s}_{im} = 0, \gamma^{s}_{im} = 1 
\end{array} \right.
\]

\[
\begin{array}{c|c|c|c|c|c|c|c}
\delta^{s}_{im} & \gamma^{s}_{im} & \mathbb{P}(c^{s}_{imk} = x | \delta^{s}_{im}, \gamma^{s}_{im}) \\
\hline
1 & 0 & 1 - p_s & p_s & 0 \\
0 & 1 & 1 - p_s & p_s & 0 \\
0 & 0 & 1 - q_s & 0 & q_s \\
\end{array}
\]

\[
y^{s}_{imk} \sim \left\{ \begin{array}{ll}
\pi \delta_0 + (1 - \pi)(1 + NB(\mu_0, n_0)) & \text{if } c^{s}_{imk} = 0 \\
\left(\text{NB} (\exp(m^{s}_{imk}), r_v)\right) \\
m^{s}_{imk} = \lambda_s + v^{s}_{imk} + u^{s}_{imk} + o^{s}_{imk} & \text{if } c^{s}_{imk} = 1 \\
u^{s}_{imk} \sim N(0, \sigma^{2}_u) & \text{if } c^{s}_{imk} = 2 \\
Pois(\tilde{\mu}) & \text{if } c^{s}_{imk} = 2
\end{array} \right.
\]

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.
matrix $\Sigma$ can be modelled from a kernel function, $k$, such as an exponential kernel, so that $\Sigma_{ij} = k(x_i, x_j)$, where $x_i$ and $x_j$ are the locations of site $i$ and $j$, respectively.

**Biomass collection** We denote by $w_{im}^s$ the amount of biomass of species $s$ collected in sample $m$ from site $i$ and $v_{im}^s := \log(w_{im}^s)$. To account for Stage 1 false negative error at this stage, we introduce latent variable $\delta_{im}^s$ that is equal to 1 if biomass for species $i$ has been collected in the $m$-th physical sample from site $i$, and 0 otherwise. We assume that $\delta_{im}^s = 1$ with probability $\theta_{im}^s$, which is a function of covariates $X_{im}^w$, and of $l_{im}^s$, since higher amounts of biomass are expected to lead to a higher probability of collecting biomass in the sample, leading to $\logit(\theta_{im}^s) = \phi_0^s + \phi_1^s l_{im}^s + X_{im}^w \phi^w$. We note that as $l_{im}^s$ tends to $-\infty$, $\theta_{im}^s$ tends to 0, and therefore the species becomes practically impossible to detect. If the amount of biomass collected is greater than 0 ($\delta_{im}^s = 1$), we model $v_{im}^s \sim N(\eta_s + l_{im}^s + X_{im}^w \beta^w, \sigma^2_s)$, where $\eta_s$ models Stage 1 species effects on the biomass collection rate and $\sigma^2_s$ models the species-specific Stage 1 noise in the biomass collection rate. To account for Stage 1 false positive error, we introduce latent variable $\gamma_{im}^s$, which is equal to 1 with probability $\zeta^s$ if the collected biomass is the result of contamination and 0 otherwise. We assume that $\gamma_{im}^s$ can be 1 only if $\delta_{im}^s = 0$ and that $v_{im}^s \sim N(\mu_s, \nu^2_s)$ if $\gamma_{im}^s = 1$. In this way, we assume that a sample which already contains biomass of a species cannot be further contaminated by the DNA of the same species from another sample or site. We make this assumptions as there is not enough information in the data to partition the collected biomass between that which was truly collected from the site and that which was contamination from elsewhere.

**Biomass analysis** As mentioned above, by non-negligible reads we mean that some of the PCR product is successfully read by the DNA sequencer. We introduce latent variable $c_{imk}^s$ to model the success of PCR $k$, sample $m$ and site $i$ for species $s$, i.e. Stage 2 error. Firstly, if sample $m$ from site $i$ contains biomass of species $s$ ($w_{im}^s > 0$), PCR run $k$ can be successful (true positive) , $c_{imk}^s = 1$, or not successful (false negative), $c_{imk}^s = 0$, and we assume that $c_{imk}^s = 1$ with probability $p_s$. Secondly, if sample $m$ from site $i$ does not contain biomass of species $s$ ($w_{im}^s = 0$), PCR run $k$ can be successful if it yields non-negligible reads.

12
due to lab contamination (false positive), $c^{s}_{imk} = 2$, or not successful (again, $c^{s}_{imk} = 0$, true negative) and assume that $c^{s}_{imk} = 2$ with probability $q_s$.

We model the reads conditional on $c^{s}_{imk}$ as follows. Conditional on $c^{s}_{imk} = 1$, $y^{s}_{imk} \sim \text{NB}(\exp(\lambda_s + v^{s}_{im} + u^{s}_{imk} + o^{s}_{imk}), r_s)$, where $\lambda_s$ models the Stage 2 species effect on the amplification rate, $u^{s}_{imk}$ is the Stage 2 pipeline effect, with $u^{s}_{imk} \sim \mathcal{N}(0, \sigma^2_u)$, $o^{s}_{imk}$ is an offset modeling the normalisation steps described in Section 1.1 and $r_s$ is a species-specific variance of the Stage 2 noise. If more than one normalisation steps are employed, then they can all be incorporated into the same offset i.e. they can be added up. Conditional on $c^{s}_{imk} = 0$, $y^{s}_{imk} \sim \pi\delta_0 + (1 - \pi)(1 + \text{NB}(\mu_0, n_0))$, that is, there are zero reads with probability $\pi$, and non-zero but negligible reads otherwise. Finally, conditional on $c^{s}_{imk} = 2$, $y^{s}_{imk} \sim \text{Pois}(\tilde{\mu}_s)$. The negative binomial is parameterised in terms of the mean and the number of failures. A visual representation of the PCR process when $c^{s}_{imk} = 1$ is shown in Figure 1 of the Supplementary material.

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

The model is summarised in Figure 2 (a), the directed acyclic graph of the model is shown in Figure 2 (b), while a graphical representation of the latent variables introduced across both stages is shown in Figure 2 (c).

The model presented in Figure 2 is not identifiable in its general form unless certain constraints are applied, as we discuss below. For example, if we define $\tilde{v}^{s}_{im} := v^{s}_{im} - \eta_s - l^s_i$ and $\tilde{l}^{s}_{i} := l^s_i - \beta^s_0$ the model for $\theta^{s}_{im}$ and $y^{s}_{imk}$ conditional on $c^{s}_{imk} = 1$ and all offsets $o^{s}_{imk}$ set to 0 can be expressed as
\[
\begin{align*}
\tilde{l}_i^s & \sim N(X_i\beta^s, \tau^2_s) \\
\tilde{v}_{im}^s & \sim N(X_{im}\beta^w_s, \sigma^2_s) \\
\theta_{im}^s & = \text{logit}(\phi^s_0 + \phi^s_1\beta^s_0 + \phi^s_2\tilde{l}_i^s + \phi^s_{im}X^s_{im}) \\
y_{imk}^s & \sim \text{NB}\left(\exp(\beta^s_0 + \tilde{l}_i^s + \eta_s + \tilde{v}_{im}^s + \lambda_s + u_{imk}), r_s\right)
\end{align*}
\]

(1)

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

\[
(\beta^s_0)^* = \beta^s_0 + c + d, \quad (\lambda_s)^* = \lambda_s - c, \quad (\eta_s)^* = \eta_s - d, \quad (\phi^s_0)^* = \phi^s_0 - \phi^s_1(c + d).
\]

The reason for this unidentifiability is that data are observed only in the third level of the model, and hence the following sets of species-specific parameters are confounded: the baseline amount of biomass across all sites (\(\beta^s_0\)) with the baseline collection rate (\(\eta_s\)) and the baseline amplification rate (\(\lambda_s\)), and the former again with the baseline detection rate \(\phi^s_0\). However, by assuming that all these baseline rates are constant across sites, samples, and PCRs, we are able to infer species-specific changes in biomass across sites and therefore covariate effects and correlations between species and between sites.

For inferential purposes, we set \(\phi^s_0 \equiv 0\) and \(\eta_s \equiv 0\). Using these constraints, the new baseline (log) amount of biomass, \((\beta^s_0)^*\), is equal to \(\beta^s_0 + \eta_s\), which means that we can only estimate the sum of the baseline amount of available biomass and corresponding baseline collection. Similarly, the new baseline (logit) collection probability \((\phi^s_0)^*\) is equal to \(\phi^s_0 - \phi^s_1\eta_s\), and therefore the baseline collection probability is also confounded with the baseline collection rate. As a result, we cannot infer the amount of available biomass separately from the collection rate, and hence the estimates of log biomass obtained, as mentioned above, are only meaningful for comparison within each species. For the same reason, comparisons of absolute amount of biomass across species are not meaningful.

We also note that depending on the survey design in terms of the number of samples collected per site and the number of PCR replicates per sample, additional sets of parameters can be confounded and not estimable. Specifically, the following pairs of parameters are confounded:
• $S = 1$: pipeline effect $u_{imk}$ and PCR variance $r_s$,
• $K = 1$: PCR variance $r_s$ and the sample noise $\tilde{v}_{im}$,
• $M = 1$: sample noise $\tilde{v}_{im}$ and site noise $\tilde{l}_i$.

These are pathological cases that do not involve replication at the site/sample/PCR levels. Replication is vital for being able to account for and to estimate the effects of the different sources of noise and error (Buxton et al., 2021), an issue to which we return in Section 4.1.

Finally, we note that if the offsets $o_{imk}$ introduced in the model due to the several normalizations occurring in the pipeline are not recorded, estimates from the model would be biased. However, a potential way to mitigate this bias is the introduction of spike-ins, which contribute to the estimation of the “overall” pipeline effects $\tilde{u}_{imk} = u_{imk} + o_{imk}$.

### 2.1 Special cases

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

\[
\begin{align*}
(y_1^i, \ldots, y_S^i) &\sim \text{Multi}(N_i, \pi_1^i, \ldots, \pi_S^i) \\
(\pi_1^i, \ldots, \pi_S^i) &\sim \text{Dirichlet}(w\alpha^1, \ldots, w\alpha^S)
\end{align*}
\]

(2)

where $N_i = \sum_{s=1}^S y_s^i$. The DMM can be seen as a special case of the model described in Section 2 for the Stage 2 process, conditional on $\delta_s^i = 1$. Specifically, $y_s^i \sim \text{NB}(\exp(\lambda_s + v_s^i + u_i), r_s)$, and therefore, assuming $\lambda_s = u_i = 0$, if $r_s \to \infty$, the distribution for $y_s^i$ converges to a $\text{Pois}(\exp(v_s^i))$. Conditional on $N_i$, the model is a $\text{Multi}(N_i, \pi_1^i, \ldots, \pi_S^i)$, where $(\pi_1^i, \ldots, \pi_S^i) = \left(\frac{\exp(v_1^i)}{\sum \exp(v_s^i)}, \ldots, \frac{\exp(v_S^i)}{\sum \exp(v_s^i)}\right)$. Next, assuming $\exp(v_s^i) \sim \text{Gamma}(w\alpha_s, \theta)$, we obtain $(\pi_1^i, \ldots, \pi_S^i) \sim \text{Dirichlet}(w\alpha_1, \ldots, w\alpha_S)$. Finally, as the DMM does not take errors into account, the equivalence with our model can be obtained by setting $p_s \equiv 1$.

McLaren et al. (2019) propose to account for the Stage 2 species effect in the DMM framework by modelling the probabilities $(\pi_1^i, \ldots, \pi_S^i)$ as $(\frac{e_{1s}^1 \pi_1^i}{\sum e_{ss}^1 \pi_s}, \ldots, \frac{e_{Ss}^S \pi_S^i}{\sum e_{ss}^S \pi_s})$, where $e_s$ models the species-specific efficiencies, which in our model is achieved by using a species-specific $\lambda_s$. The DMM can be extended hierarchically if nested treatments are considered.
(Coblentz et al., 2017) by defining a nested prior \((\alpha^1, \ldots, \alpha^S) \sim \text{Dirichlet}(\alpha^1_0, \ldots, \alpha^S_0)\) for each level. In our model, this is achieved by a hierarchy of normal priors. This highlights a key difference between the DMM approach and the approach we introduce in this paper, since we model the propagation of the absolute amount of DNA across the different stages, while the DMM models the propagation of the relative amount of DNA.

Secondly, the occupancy model of [Griffin et al., 2020], in the simple case of no covariates,

\[
\begin{align*}
  z_i &\sim \text{Be}(\psi) \\
  w_{im} &\sim \text{Be}(z_i\xi_1 + (1-z_i)\xi_0) \\
  y_{imk} &\sim \text{Be}(w_{im}p + (1-w_{im})q)
\end{align*}
\]

(3)
designed for (single-species) qPCR, can be seen as a special case of the model in Section 2 when the information in the counts is not considered. Specifically, letting \(l_i\) be binary, with \(l_i \in \{-\infty, 0\}\), and defining \(z_i = \exp(l_i)\), we obtain \(\theta_{im}|(l_i = -\infty) = 0\) and \(\theta_{im}|(l_i = 0) = \text{logit}(\phi_0)\). Hence, the model for \(\delta\) and \(c\) becomes

\[
\begin{align*}
  \delta_{im} &\sim \text{Be}(z_i(\text{logit}(\phi_0) + (1 - \text{logit}(\phi_0))\zeta) + (1-z_i)\zeta) \\
  c_{imk} &\sim \text{Be}(\delta_{im}p + (1-\delta_{im})q)
\end{align*}
\]

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

3 Inference

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

For the parameters \(\sigma_s, \mu_s, B\) and \(B_0\), the full conditional distribution is available in closed form, and therefore posterior sampling is straightforward. We use simple random walk Metropolis Hastings steps for parameters \(\pi, \mu_0, n_0\), and \(r_s\) and Metropolis-Hastings steps with a Laplace approximation proposal for the parameters \(l^s_i, \lambda_s, v^s_{im}, u_{imk}\) and \(r_s\).
However, on its own, this naive Gibbs sampler will mix slowly since we have a complex hierarchical model with cross effects and many latent variables. We address this by updating parameters in blocks using re-parameterisation and an adaptive updating scheme for the discrete latent variables.

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

\[
\begin{align*}
    l_i^s &\sim N(0, \tau_s^2) \\
    v_{im}^s &\sim N(l_i^s, \sigma_s^2) \\
    u_{imk} &\sim N(0, \sigma_u^2) \\
    y_{imk}^s &\sim NB(\exp(\lambda_s + v_{im}^s + u_{imk}), r_s)
\end{align*}
\]  

(4)

A naive Gibbs sampler updating each parameter from its full conditional leads to prohibitively slow mixing, due to the form of the likelihood where \(\lambda_s\), \(v_{im}^s\) and \(u_{imk}\) appear as a sum. To address the slow mixing in the nested effects, \(\lambda_s\) and \(v_{im}^s\), the use of a centred parameterisation (Papaspiliopoulos et al., 2007) has been suggested, which corresponds to 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 \(\bar{v}_{im}^s\) and \(u_{imk}\) and, as noted by Zanella and Roberts (2021), re-parameterisation does not improve mixing in the case of crossed-effects models. In a classic crossed-effect model of the form \(y_{jkl} \sim N(\lambda + v_j + u_k, \sigma^2)\), Papaspiliopoulos et al. (2020) propose a collapsed Gibbs sampler by first jointly sampling \(\lambda\) with \(v_j\) and then \(\lambda\) jointly with \(u_k\). However, this approach does not scale well in our setup, since it would involve sampling all the \(\lambda_s\) and \(u_{imk}\) jointly, which have dimensions \(S\) and \(\sum_{i,m} K_{im}\) respectively. Zanella and Roberts (2021) propose the use of identifiability constraints on the model, which in Equation (4) correspond to assuming \(\sum_s v_{im}^s = \sum_k u_{imk} = 0\). Since sampling conditionally on constraints can be challenging, we propose a simpler strategy to improve mixing that is more suited to our framework. We consider re-parameterising to the factor averages \(\hat{v}_{im} = \frac{1}{S} \sum_{s=1}^S v_{im}^s\) and \(\hat{u}_{im} = \frac{1}{K} \sum_{k=1}^K u_{imk}\) and the factor increments \(\tilde{v}_{im}^s = v_{im}^s - \hat{v}_{im}\) and \(\tilde{u}_{imk} = u_{imk} - \hat{u}_{im}\) and performing an update by first sampling jointly 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 using the stan-
standard updates \((u_{imk}|v_{im1}, \ldots, v_{im}^S)\) and \((v_{imj}^j|u_{im1}, \ldots, u_{imK})\). In our simulations, we have found that jointly updating the factor means considerably improves mixing. The sampling scheme for the complete model is presented in the Supplementary material.

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

## 4 Study design

In this section, we investigate the properties of the model under different study designs in terms of the number of sites, samples per site, and PCRs per sample, as well as the number of spike-ins. In each section, we consider the estimates of the differences in log-biomass within species, when that is not a function of site-specific covariates (no covariate case), and the estimates of the covariate coefficients when log-biomass is a function of a single continuous covariate (regression case).

In Section 4.1, we present theoretical results using a continuous version of our model that
does not account for error in either stage. In Section 4.2 we fit our model as presented in Section 2 under different scenarios for study design by varying the number of sites, number of samples per site, and number of PCRs per sample. Finally, in Section 4.3 we explore the effect of spike-ins for different levels of noise in each stage of the process and different study designs.

4.1 Theoretical results

We consider a normal approximation of the model presented in Section 2, which assumes no species or sites correlation, that both stages are error-free by setting $\theta_{im}^s = p_s = 1$, and that the variances of the distributions of the noise at each stage are the same across species.

As mentioned in Section 2, the use of spike-ins corresponds to the presence of species in the sample for which $(v_{im}^{S+1}, \ldots, v_{im}^{S+S^*})$ is known. We assume, without loss of generality, that $v_{im}^{S+j} = 0$ for $j = 1, \ldots, S^*$. We have the following proposition.

**Proposition 4.1.** Consider the model

$$
\begin{align*}
\begin{cases}
    v_{im}^s \sim N(l_{im}^s, \sigma^2) & i = 1, \ldots, n \quad m = 1, \ldots, M \quad s = 1, \ldots, S \\
    v_{im}^s = 0 & i = 1, \ldots, n \quad m = 1, \ldots, M \quad s = S + 1, \ldots, S + S^* \\
    u_{imk} \sim N(0, \sigma_u^2) & i = 1, \ldots, n \quad m = 1, \ldots, M \quad k = 1, \ldots, K \\
    \lambda_s \sim N(0, \sigma_\lambda^2) & s = 1, \ldots, S + S^* \\
    y_{imk}^s \sim N(u_{imk} + \lambda_s + v_{im}^s, \sigma_y^2) & i = 1, \ldots, n \quad k = 1, \ldots, K \quad s = 1, \ldots, S + S^*
\end{cases}
\end{align*}
$$

Then

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

(5)

If we assume that

$$
l_i^s \sim N(X_i \beta, \tau^2) \quad i = 1, \ldots, n \quad s = 1, \ldots, S
$$

and $\sigma_\lambda^2 \gg \max\{\sigma_u^2, \sigma^2, \sigma_y^2\}$, we obtain

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

$$
\left( 1 + \frac{\sigma_u^2}{\sigma_y^2} + (M \tau^2 + \sigma^2) K (1 + S^* \frac{\sigma_u^2}{\sigma_y^2}) + \sigma_y^2 (S + S^* - 1) \right).
$$

(6)

Here $\sigma_y^2$ models the variance of the noise in Stage 2, as was the case for $r_s$ in the original model. In the no covariate case we have assumed a flat prior on $l_i^s$, while in the regression
case, we have assumed a single covariate. Equations (5) and (6) show the contribution of
the variance of each stage to the posterior variance of the corresponding estimates (changes
in species log-biomass between sites and covariate coefficients, respectively) in this special
case.

The results for this special case suggest that, for both $\text{Var}(l_1^s - l_2^s|y)$ and $\text{Var}(\beta|y)$,
increasing replication at a given stage decreases the contribution of the error variance at
that stage and at all downstream stages. For example, increasing the number of samples
per site, $M$, reduces the contribution of the noise variances at Stage 1, $\sigma^2$, and of all
downstream stages, i.e. $\sigma_y^2$ and $\sigma_u^2$ in Stage 2. Whereas, increasing the number of PCR
replicates, $K$, only reduces the contribution of the Stage 2 variances ($\sigma_u^2$ and $\sigma_y^2$).

Additionally, the benefit of the spike-in is greater as the ratio of variances $\sigma_u^2 / \sigma_y^2$ increases.
Moreover, in the case of $\text{Var}(\beta|y)$, if $\sigma^2 \gg \sigma_y^2$, the benefit of the spike-in is negligible, as
the noise induced by $\sigma^2$ greatly outweighs the noise that can be mitigated via the use of
spike-ins.

4.2 Varying $n$, $M$, and $K$

We turn our attention to the full model in Section 2 and again consider the no covariate
and regression cases. In the no covariate case, we consider the model’s ability to estimate
the correct sign of the difference of species log-biomasses 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 the posterior probability of
$l_{i_1}^s > l_{i_2}^s$ and $\delta_{i_1,i_2}$ is 1 if the true value of $l_{i_1}^s$ is greater than the true value of $l_{i_2}^s$ and 0
otherwise. We separate the sites between those with “high” biomass and those with “low”
biomass and generate $l_i^s \sim \begin{cases} N(1, \tau_s^2) & i \text{ odd} \\ N(0, \tau_s^2) & i \text{ even} \end{cases}$ We use $S = 40$ species, $n = 300$ sites,
$M \in \{1, 2, 3, 4, 5\}$ samples per site and $K \in \{1, 2, 3, 4\}$ PCR replicates. The values of the
other parameters are reported in the Supplementary Material. We have performed a set
of 10 simulations for each combination of values of the design parameters, $M$ and $K$. We
report the average $b(i_1, i_2, j)$ spanning $i_1$ across the sites with low biomass, $i_2$ across the
sites with high biomass, and $s$ across all species and across the 10 simulations. As expected, the Brier score decreases, and hence the ability to distinguish between sites with low and high biomass increases, as $M$ and $K$ increase (Figure 3). However, when $M = 1$, increasing $K$ yields little benefit, once more highlighting the greater importance of multiple sample replicates per site in Stage 1, whereas the benefit of larger $K$ increases as $M$ gets larger.

Figure 3: Brier score for distinguishing high- and low-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.

In the regression case, we consider the absolute error and posterior standard deviation of the covariate coefficient $\beta_s$. We use $n \in \{50, 100, 200\}$ sites, $M \in \{1, 2, 3\}$ samples per site and $K \in \{1, 2, 3\}$ PCR replicates per sample and $S = 40$ species. The values of the other parameters are reported in the Supplementary Material. We performed a set of 10 simulations for each combination of values of the design parameters and averaged results across all simulations and species, which are shown in Figure 4.

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

4.3 Spike-ins

In this section, we consider the improvement in inference when spike-ins are employed in
Stage 2. We consider a version of the model that does not account for false negative/positive
errors, as the effect of the spike-ins is maximised in this case. If errors are present in either
or both stages, the benefit of the spike-ins is lower, and dependent upon the level of error.

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
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, $\sigma$, is represented by the line type, the between-sites standard deviation, $\tau$, 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 relative posterior variance.

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

The second row of Figure 5 shows the regression case. We have chosen smaller values for \( \sigma \) and \( \tau \) (0.2 and 0.5), since the relative contribution of the spike-ins is negligible with larger values. Spike-ins contribute more to reducing error and variance when the between-samples standard deviation (\( \sigma \)) and the between-sites standard deviation (\( \tau \)) is smaller since, similar to before, the noise at early stages dominates the process, and therefore the relative contribution of the spike-ins is smaller due to the improved estimation of the PCR noise. In a similar way to the no covariate case, the contribution of the spike-ins increases for \( K = 1 \) PCR replicates compared to \( K = 3 \). However, unlike that case, the contribution does not appear to increase as the number of samples per site \( M \) gets larger.

5 Case study

We apply our model to an unpublished dataset of mostly arthropod invertebrates collected using 121 Malaise-trap samples from 89 sample sites in the H. J. Andrews Experimental Forest (HJA), Oregon USA (225 km\(^2\)) in July 2018. Each trap was left to collect for seven days, and samples were transferred to fresh 100% ethanol to store at room temperature until extraction. The management objective that motivated the collection of this dataset is to interpolate continuous species distributions among the 89 sample points so that areas of higher and lower conservation value at the HJA can be identified.

For each sample, the collected invertebrate biomass was combined with a lysis buffer, in an amount proportional to the starting biomass, to digest the tissue, and a fixed aliquot was then taken from the overall mixture for DNA extraction and subsequent PCR. This normalization, as described in Section 2, was accounted for in the model by setting the offset \( o_{imk} \) equal to the log ratio between the aliquot and the overall amount of liquid mixture in each case.

We included 50 species in the study by selecting the species that have the most non-zero counts across all PCR replicates. Log-biomass is modelled as a function of two environ-
mental covariates: log elevation and log distance-to-road.

Figure 6 presents the 95% posterior credible intervals (PCIs) for the species-specific coefficients of log elevation and log distance-to-road in the model for log-biomass. The effects of the covariates on species biomass are not consistent within each taxonomic order, which suggests low phylogenetic inertia at this rank for response to these landscape characteristics. Elevation is a stronger predictor for species biomass than distance-to-road for this ecosystem. This makes ecological sense, since distance-to-road is only expected to exert an effect over a ca. hundred metres, via canopy openness, whereas elevation exerts a 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-biomass. Species are grouped taxonomically.

Figure 7 (a) presents the posterior mean of the between-species residual correlations. Species in the Diptera (flies, spp. 14-30) exhibit higher positive correlations with each other, as well as with several species in the Hymenoptera (ants, bees, and wasps) and Lepidoptera (butterflies and months). We conservatively interpret these positive residual correlations as indicative of unmeasured environmental covariates, such as canopy openness, rather than of biotic interactions. We also note that two species in the Lepidoptera, (spp. 41, 43), one in the Hymenoptera (sp. 33), and one in the Psocodea (barklice, sp. 50) are among the few species showing strong negative residual correlation with many of the other species, and
again, we interpret these correlations as indicative of unmeasured environmental covariates. There is a strongly positive, pairwise correlation between two tabanid fly species *Hybomitra liorhina* and *Hybomitra* sp. indet (spp. 12, 13), which might indicate the oversplitting of one biological species into two OTUs during the bioinformatic pipeline. Finally, there is also a strongly positive, pairwise correlation between the moth species *Ceratodelia gueneata* (sp. 44) and the predatory fly (Scathophagidae, *Microprosopa* sp. indet, sp. 20), which might indeed indicate a specialised predator-prey relationship.

![Correlation plot of all species](image1)

![Posterior mean of biomass-weighted species richness across the study area](image2)

Figure 7: Case study: (a) Correlation plot of all species. Red represents positive correlations while blue represents negative correlations. Species are grouped taxonomically. (b) Posterior mean of biomass-weighted species richness across the study area. The value in each point has been computed by rescaling the log biomass of each species in the [0, 1] range and then summing all the rescaled biomasses.

In Figure 7(b) we show the biodiversity map for the area, which is useful for identifying areas of higher species richness and compositional distinctiveness, which together can be used to identify areas of higher conservation value (i.e. higher ‘site irreplaceability’ sensu Baiser et al., 2022). The predicted mean log-biomasses on a continuous map over the HJA for all individual species are presented in the Supplementary Material. These can be
used to identify species with a wide spatial range, such as the click beetle (*Megapenthes caprella*), or with a restricted range, such as the leafhopper (*Osbornellus borealis*).

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

![Figure 8](image-url)

Figure 8: Case study: (left) 95% PCI of the species-specific between-samples standard deviation $\sigma_s$ and between-sites standard deviation $\sqrt{T_{ss}}$ (in bold). (center) 95% PCI of the species-specific average collection probability $\theta_{im}$ 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.
6 Discussion

Over the last decade, DNA-based biodiversity studies, primarily using metabarcoding, have rapidly increased in popularity, and multivariate statistical models are now starting to be deployed to analyse metabarcoding data (e.g. Lin et al., 2021; Pichler and Hartig, 2020; Abrego et al., 2021; Fukaya et al., 2022; Ji et al., 2022). Our paper provides the first unifying modelling framework that considers and quantifies all main sources of variation, error and noise in metabarcoding surveys (Table 1). As a result, our modelling framework should allow more reliable and more powerful biodiversity monitoring and inference on species responses to landscape characteristics than has been possible before. We have employed, extended, and developed a number of inferential tools to deal with the complexity of the proposed hierarchical model, which involves two latent stages and a large number of latent variables. Finally, this is the first modelling approach that accounts for spike-ins and negative controls (empty tubes), which are widely used quality-control methods in DNA-based biodiversity surveys but rarely explicitly considered within a modelling framework. We explored the benefits of spike-ins on inference and, for the first time, explored issues of study design in metabarcoding data, providing both theoretical and simulation results.

Our new framework allows us to infer and map species biomass as well as biodiversity across surveyed sites (Figure 7 (b)), and to link these to landscape characteristics (Figure 6). The resulting maps can be used to identify areas of high conservation value, as well as areas where particular species or groups of species are more or less prevalent, and to detect species-specific shifts, expansions, and shrinkage. We are also able to study pairwise correlations across large numbers of species (Figure 7 (a)), which is considerably more scalable using metabarcoding data than using standard observational data. We have shown that using spike-ins can substantially increase inference accuracy for parameters of interest (Figure 5). Our results also demonstrate that the current practice of collecting a single sample from each surveyed site considerably reduces our ability to infer changes in species biomass and that replication at both stages as well as the use of spike-ins is the optimal
approach to designing metabarcoding studies (Figure 3).

In metabarcoding data, the baseline biomass of each species is confounded with its species-specific collection and amplification rates. Hence, we cannot infer absolute values of species-specific biomass across sites using metabarcoding data alone. However, by assuming that baseline species-specific collection and amplification rates are the same across sites, samples, and PCR replicates, we can infer species-specific biomass change across sites, species-specific covariate effects, and pairwise species correlations. These assumptions are justifiable in the case study, where eDNA release rates cannot differ across sites because the samples are whole organisms, and there is also no reason to believe that Malaise traps vary in trapping efficiency across sites. In other scenarios, eDNA release or degradation rates could vary across the environment, due to large variations in, for instance, food availability and water pH, respectively. In these cases, site covariates could be used to account for this suspected variation.

Generally, modelling changes in (proxies of) abundance, such as changes in biomass, is a more powerful monitoring tool than modelling changes in species presence across surveys sites (Joseph et al., 2006). Metabarcoding studies yield count data without any consequence on associated cost, and hence overcome the time and cost implications associated with collecting count data for multiple species. Our model uses the raw count data, and does not rely on ad-hoc rules about what constitutes a practically zero count for converting them to binary data, which has been the standard practice thus far (Ovaskainen et al., 2017; Bush et al., 2020).

The model can easily be extended to account for multiple primers, which essentially introduce different levels in Stage 2, with each level corresponding to a set of primers. For instance, vertebrate eDNA surveys can profitably use two primers (e.g. the 16S mammal and 12S vertebrate primers used in Ji et al., 2022), since the two primers together detect the largest set of vertebrates, with overlap. Additionally, our approach can be extended and applied to metagenomic data. In metagenomics, no PCR is employed. Instead, all
the DNA in each mixed-species sample is sequenced, and the DNA barcodes (or other taxonomically informative genetic sequences) are discovered bioinformatically.

Metabarcoding studies, particularly when applied to microbiomes and meiofauna (e.g. nematodes, micro-eukaryotes), can detect 10000s of species, which leads to large numbers of latent variables and coefficients in the model. There are several ways that the inferential tools presented here could be further extended to scale to these cases. Firstly, the posterior distribution conditional on the $u_{imk}$ is independent across species. If $u_{imk}$ could be estimated at a first stage then inference across species could be easily parallelized. Secondly, variational Bayes methods could be applied to avoid the use of sampling methods. The choice of variational distribution will be important and can exploit the conditional normality of much of the model. Alternatively, the model could be adapted by assuming that the coefficient matrices such as $\beta^z = (\beta^z_1, \ldots, \beta^z_S)$, have a low-dimensional representation.

The metabarcoding process produces large numbers of different DNA-barcode sequences, which represent not only different species but also within-species genetic diversity and outright errors generated during PCR and sequencing. Currently, these sequences are clustered into species hypotheses as part of the bioinformatic process, following a number of heuristic rules and thresholds (hence, the word ‘operational’ in OTU), but future work could explore building the OTU table during the model-fitting process.

We are not modelling species presence/absence and instead we have focused on modelling biomass on a continuous scale. As a result, we cannot infer whether a species is absent from a particular study site, but instead only if its biomass at given site is practically zero for modelling purposes. We have assumed that a sample which already contains biomass of a species cannot be further contaminated by the DNA of the same species from another sample or site in Stage 1. This is a reasonable but also necessary assumption. Under best practice, contamination in Stage 1 is expected to be relatively rare, and therefore, there is not enough information in the data to partition the collected biomass between that which was truly collected from the site and that which was contamination from elsewhere. If this
assumption is violated, then the amount of collected biomass is no longer proportional to the amount of available biomass, and the amount of available biomass at sites where there has been contamination can be overestimated.

eDNA metabarcoding has revolutionised the cost-effectiveness, precision and scale at which biodiversity assessment can be performed. Nevertheless, the multiple stages at which imperfect detection of biomass can occur during the workflow are not insignificant. By facilitating estimates of within-site changes in biomass and covariates while accounting for workflow uncertainties, our modelling framework provides a step-change improvement in the design and analysis of eDNA metabarcoding data.

References
Abrego, N., Roslin, T., Huotari, T., et al. (2021). Accounting for species interactions is necessary for predicting how arctic arthropod communities respond to climate change. *Ecography*, 44(6):885–896.

Andrieu, C. and Thoms, J. (2008). A tutorial on adaptive MCMC. *Statistics and Computing*, 18(4):343–373.

Baisero, D., Schuster, R., and Plumptre, A. J. (2022). Redefining and mapping global irreplaceability. *Conservation Biology*, 36(2).

Besson, M., Alison, J., Bjerge, K., et al. (2022). Towards the fully automated monitoring of ecological communities. *Ecological Letters*, to appear.

Bohmann, K., Evans, A., Gilbert, M. T. P., et al. (2014). Environmental DNA for wildlife biology and biodiversity monitoring. *Trends in Ecology & Evolution*, 29(6):358–367.

Bush, A., Monk, W. A., Compson, Z. G., et al. (2020). DNA metabarcoding reveals metacommunity dynamics in a threatened boreal wetland wilderness. *Proceedings of the National Academy of Sciences*, 117(15):8539–8545.

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.

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., Kjøller, 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.
Goldberg, C. S., Turner, C. R., Deiner, K., et al. (2016). Critical considerations for the application of environmental DNA methods to detect aquatic species. *Methods in Ecology and Evolution*, 7(11):1299–1307.

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

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

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

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

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

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

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

Lin, M., Simons, A. L., Harrigan, R. J., et al. (2021). Landscape analyses using eDNA metabarcoding and Earth observation predict community biodiversity in California. *Ecological Applications*, 31(6).
Lindahl, B. D., Nilsson, R. H., Tedersoo, L., et al. (2013). Fungal community analysis by high-throughput sequencing of amplified markers—a user’s guide. *New Phytologist*, 199(1):288–299.

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

Lynggaard, C., Bertelsen, M. F., Jensen, C. V., et al. (2022). Airborne environmental dna for terrestrial vertebrate community monitoring. *Current Biology*, 32(3):701–707.e5.

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

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

Ovaskainen, O. and Abrego, N. (2020). *Joint Species Distribution Modelling: With Applications in R*. Cambridge University Press.

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

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

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

Pawlowski, J., Apothéloz-Perret-Gentil, L., and Altermatt, F. (2020). Environmental DNA: What’s behind the term? Clarifying the terminology and recommendations for its future use in biomonitoring. *Molecular Ecology*, 29(22):4258–4264.

Pichler, M. and Hartig, F. (2020). sjSDM: Scalable Joint Species Distribution Modeling.
Piper, A. M., Batovska, J., Cogan, N. O. I., et al. (2019). Prospects and challenges of implementing DNA metabarcoding for high-throughput insect surveillance. *GigaScience*, 8(8):giz092.

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

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

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

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

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

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

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

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

Tosa, M. I., Dziedzic, E. H., Appel, C. L., et al. (2021). The Rapid Rise of Next-Generation Natural History. *Frontiers in Ecology and Evolution*, 9:698131.

van Klink, R., August, T., Bas, Y., et al. (2022). Emerging technologies revolutionise insect ecology and monitoring. *Trends in Ecology & Evolution*, 37(10):872–885.
Wang, H. (2012). Bayesian graphical lasso models and efficient posterior computation. *Bayesian Analysis, 7*(4):867–886.

Zanella, G. and Roberts, G. (2021). Multilevel linear models, gibbs samplers and multigrid decompositions (with discussion). *Bayesian Analysis, 16*(4):1309–1391.