Estimating relative species abundance using fossil data identified to different taxonomic levels

Trond Reitan[1], Emanuela Di Martino[1], and Lee Hsiang Liow[1]

[1]University of Oslo

April 1, 2023

Abstract

Site-occupancy modeling is widely used in ecology but its application is still limited in paleoecology, where incomplete detection is routine. Here, we make extensive expansions to an earlier multispecies occupancy model used to estimate the dynamics of relative species abundance in fossil communities. These expansions include incorporating counts of individuals at sites, explicitly allowing for the inclusion of specimens assignable to genus- but not species-level, a situation common in paleontology, and modelling regional presence/absence. We provide simulations to check the performance of this new model, as well as simulations to quantify the benefits of using individual count data versus subsample occupancy data and model estimates versus face-value (raw) estimates, respectively. We also provide an empirical case study using occupancy data from a community of marine benthic colonial animals preserved in the Pleistocene of New Zealand. We find that the new model performs well, especially when it comes to recovering relative abundance dynamics and that it is well worth the effort to both collect individual count data and to include individuals unidentified to species-level in the site-occupancy modelling framework. This extended model can be widely applied in paleoecological settings and is necessary when both the average and uncertainty values of relative abundance dynamics need to be robustly estimated.

Abstract (207/300)

Site-occupancy modeling is widely used in ecology but its application is still limited in paleoecology, where incomplete detection is routine. Here, we make extensive expansions to an earlier multispecies occupancy model used to estimate the dynamics of relative species abundance in fossil communities. These expansions include incorporating counts of individuals at sites, explicitly allowing for the inclusion of specimens assignable to genus- but not species-level, a situation common in paleontology, and modelling regional presence/absence. We provide simulations to check the performance of this new model, as well as simulations to quantify the benefits of using individual count data versus subsample occupancy data and model estimates versus face-value (raw) estimates, respectively. We also provide an empirical case study using occupancy data from a community of marine benthic colonial animals preserved in the Pleistocene of New Zealand. We find that the new model performs well, especially when it comes to recovering relative abundance dynamics and that it is well worth the effort to both collect individual count data and to include individuals unidentified to species-level in the site-occupancy modelling framework. This extended model can be widely applied in paleoecological settings and is necessary when both the average and uncertainty values of relative abundance dynamics need to be robustly estimated.

Keywords: hierarchical modelling, multispecies site-occupancy models, fossil communities, preservation, Bryozoa

Introduction

The abundance of a given species in its community is the consequence of population growth, which in turn is a consequence of survival and reproduction. The latter are influenced by competition, predation, disease, and intraspecific variability and environmental stochasticity. The relative abundance or dominance of different species in natural, contemporary communities are observed to shift on shorter time-scales, where such shifts can be directly attributed to environmental change, invasive species, cyclical behavior, among other factors. On longer time-scales where observations are more challenging, however, the imprint of multiple processes not only obscure underlying mechanisms of such shifting dominance, but may also veil true differences in relative abundances. Yet, it is important to be able to reconstruct population dynamics deeper in time, using genetic evidence, biogeographic and/or paleoecological data to understand the past (Hoban et al. 2019, Dussex et al. 2021) and to use the past as baselines for anthropogenic change (Dillon et al. 2022) .

Site-occupancy modeling uses information from repeated site visits to account for incomplete detection while estimating population and community parameters, including relative abundance. It is widely applied in many branches of ecology but its application is limited in paleoecology, despite detection also being incomplete in the fossil record (Liow 2013, Lawing et al. 2021). Incomplete detection in the fossil record can be in part attributed to non-biological factors, including varying sedimentation rates, storms, bioturbation, lateral transport, erosion and other processes that themselves tend to be temporally varying on longer time scales. A recent study used fossil data to estimate the dynamics of relative species abundance in a Pleistocene benthic community by developing a multispecies occupancy model that takes into consideration the features of fossil preservation (Reitan et al. 2022).

Reitan et al. 2022 were interested in how different species of marine invertebrates encrusting hard substrates change in their relative abundances over 2 million years. More specifically, they wanted to build a hierarchical model to estimate how several co-existing cheilostome bryozoan species waxed and waned over time across several geological formations within the Wanganui basin of New Zealand. In the model they developed, which can also be applied to other paleoecological study systems, detection was in a one-to-one relationship with underlying abundance given site-occupancy.

This previous fossil multispecies occupancy model had features that are particularly suited to data commonly collected or are collectable in paleoecological settings. Like all site-occupancy models, (fossil) sites are re-sampled such that data from the replicate sampling allow us to tease apart site-occupancy and detection. The replicate sampling are subsamples within sites, which in the case of Reitan et al. 2022 were unique shells found within the sites, on which different species of encrusting cheilostome bryozoans were observed.

The current paper extends the Reitan et al. 2022 model by i) using counts of individuals rather than only presence/absence of species on the subsample-level, ii) adding species-level random effects, iii) incorporating specimens assignable to genera but not species, iv) modelling regional presence/absence and v) incorporating information when regional presence is known. Like the original model, these improvements are applicable to many paleocological systems, in addition to the one presented in Reitan et al. 2022. To this end, we extend the dataset presented in Reitan et al. 2022, adding 18 species and 25 sites where observations were made. We provide simulations to explore how well the expanded model recovers parameters of interest, and the performance of model-estimated parameters based on individual counts or subsample-level presence/absence data versus “face-value” information, i.e. raw estimates

(see Methods and Material). We end by discussing why it is important to explicitly model detection and present general recommendations for paleoecological work.

Materials and Methods

Data

The site-occupancy data are collected from a community of fossilized benthic, encrusting cheilostome bryozoans found in the Wanganui Basin of New Zealand (Carter and Naish 1998, Proust et al. 2005, Pillans 2017) previously presented in Reitan et al. 2022. There are now subsamples (= shells, typical substrates for bryozoans) for encrusting cheilostomes in 144 sites in transgressive system track (TST) shell beds from 10 geological formations, spanning about 2 million years. Such shellbeds reflect similar depositional conditions (facies). We tabulated the observed presence of any fossilized individuals of 21 focal cheilostome species on each shell (i.e. subsample) sampled from any given site, including the three previously analyzed in Reitan et al. 2022. With the exception of five species of *Microporella*, two of *Escharoides* and two of *Exochella*, each of these species are, as far as we know, sole representatives of their genera in the Wanganui Basin. This is important for later modeling considerations. As in the previous study, the superspecies represents all other encrusting bryozoan species in the community, excluding the 21 focal species. The observed presence of the superspecies gives information to improve parameter estimates (see Model Description). These observations constitute the occupancy dataset. For additional sources concerning regional occupancy (see Extension (v) below), we draw on data collected for a separate study (Liow et al. 2016) as well as more recently collected material (unpublished but provided in the zip folder “RAMU-MSOM” available via the editor/Ecography office).

Original model: a brief recap

The objective of Reitan et al. 2022, was to estimate the temporal dynamics of relative species abundance. The data in that study had one row per site containing information about the number of subsamples having an observed presence of each species, i.e. subsample counts. A given species, *s*, has the potential of being observed in a given subsample if it is present in a given site, *i*. If a given site is not observed to contain the given species in any of its subsamples, it could mean either that i) the site was truly devoid of that species or ii) that the species was present but not sampled (MacKenzie et al. 2002) .

We denote the site-occupancy probability of a given species as Ψ and detection probability as *p.* More specifically, p is the probability that each subsample has at least one observation of the given species. The probability that a species is found on a given subsample is thus Ψp. The site-occupancy and detection probabilities can be specific to sites *i* belonging to specific time-intervals (i.e. geological formations). Here, formation, $f \in 1, \cdots, N_f$ where N_f is the number of formations, and species, $s \in 1, \cdots, S$, where S is the number of species (and the superspecies is indexed as S). Thus, we write $\Psi_{i,s}(\theta)$ and $p_{i,s}(\theta)$ for the site-occupancy and detection probabilities respectively, where θ is the set of parameters and random variables of the model. Since p is independent for each subsample, the binomial distribution can be used to summarize the chance of observing $y_{i,s}$ out of T_i subsamples in site *i,* with presence of *s*. However, there may be variation in true abundance of a species from site to site, and hence variation in its detection probability, giving rise to overdispersion. Temporal variation within each formation, observational errors and local heterogeneity in preservation can further introduce extra variation, thus, we use a beta-binomial distribution. Since site-occupancy is not guaranteed, this further expands into a zero-inflated beta-binomial distribution. We assume site-occupancy probability and the detection probability are each affected by a

115 random factor ($\delta_{f(i),s}$ and $\varepsilon_{f(i),s}$, respectively) representing individual species dynamics in a
116 given formation. Additional random factors representing dynamics common across species
117 ($v_{f(i)}$ and $u_{f(i)}$ for site-occupancy and detection probabilities, respectively) encompass
118 variation in preservation characteristics and hence detection probabilities in different
119 geological formations.

120 To estimate species relative abundance, we assume that detection probability given
121 occupancy, p, is linked to abundance-given-occupancy such $p = 1 - e^{-\lambda}$ via a Poisson model
122 where λ is the mean number of detections. λ is associated with relative abundance dynamics
123 via a log-link (i.e. the abundance-focused model in Reitan et al. 2022). We use a logistic link
124 between site-occupancy probability and the accompanying random factors. Thus $y_{i,s}$ as a
125 zero-inflated beta-binomial distribution is:

126 $$y_{i,s} \sim z\beta bin\left(T_i, p_{i,s}(\theta) = 1 - exp\left(-exp(\beta_s + u_{f(i)} + \varepsilon_{f(i),s})\right), \kappa_s, \Psi_{i,s}(\theta) = I(s = S) +$$

127 $$I(s < S)\text{logit}^{-1}\left(\alpha_s + v_{f(i)} + \delta_{f(i),s}\right)\Big) \quad (1a)$$

128 $$u_f \sim N(0, \sigma_u^2),\ v_f \sim N(0, \sigma_v^2),\ \delta_{f,s} \sim N\left(0, \sigma_{\delta,s}^2\right),\ \varepsilon_{f,s} \sim N\left(0, \sigma_{\varepsilon,s}^2\right) \quad (1b)$$

129 Here, κ_s is an overdispersion parameter (which we retrospectively found did not need the
130 species-dependency we imposed on it). *I()* is the indicator function which takes value 1 when
131 the statement inside is true and 0 if false. *S* is the total number of species. α_s and β_s give
132 average site-occupancy and detection probabilities for each species on their transformed
133 scales (but see Reitan et al. 2022).

134 Using this, relative abundance is estimated as

135 $$R_{f,s} = \frac{\Psi_{f,s}(\theta)\,\lambda_{f,s}(\theta)}{\sum_{s\prime=1}^{S}\Psi_{f,s\prime}(\theta)\,\lambda_{f,s\prime}(\theta)}. \quad (2)$$

We replaced the site index, *i*, with the formation index *f*, as both site-occupancy probability and abundance-given-occupancy only depend on species and formation here. Site-dependent variation is modelled through overdispersion.

We propose a set of modifications to the above model. Mathematical details of the new model follow after verbal descriptions of the extensions in the following section.

Model extensions

Extension (i): Individual counts versus subsample count data per site

The original modelling was performed on the number of subsamples observed to have at least one individual of a given species (subsample counts). Some subsamples were observed to have tens of individuals of some species, while others just a few or none, reduction of the information to subsample counts constitutes a potentially huge loss of information.

Handling the data on the subsample level for individual counts is likely computationally unfeasible (Reitan et al. 2022), but we can move the analysis up to the site-level (arguments given in SI). Here, we use the negative binomial for an overdispersed version of the Poisson distribution for count data. We assume that the expected number of individuals at a site scales with the number of subsamples in the site, just as for subsample count data.

Extension (ii): Species constants are replaced by random effects

In Reitan et al. 2022, data for only three focal species were available. However, most communities are more species-rich, even when considering common species, as is the community we are considering. Because only three species had to be modelled, they were each given a constant. With more species, we turn these constants into random effects since the data are rich enough for inference on the distribution of species-dependent quantities. By adapting the distribution of these quantities to the data rather than giving each species its own

prior distribution, the model is less sensitive to biases and uncertainty assumptions in the specification of priors.

Extension (iii): Individuals assignable to at least genus but not to species

Cheilostome bryozoans, like some other calcified marine taxa, can be assigned to their species with high confidence based on morphology (Jackson and Cheetham 1990), when preservation is good and post-mortem damage is minimum. However, preservation and damage can reduce the possibility for assigning an individual to a lower taxonomic level (e.g. species or even genus), a situation common in paleoecology. However, if the individual can be identified to genus but not species-level, it still gives information for occupancy modelling. Imagine there are 3 species in a region, species A1, A2 and B, where B belong to a separate genus while A1 and A2 are in the same genus. Then, detecting 100 A1, 100 A2, 200 unidentified individuals belonging to genus A and 100 individuals to B, should suggest there were really 200 A1 and 200 A2 individuals and thus that the abundance of A1 relative to B was 2 to 1 rather than 1 to 1.

We thus need to multiply the estimated abundance-given-occupancy with the probability of non-identification to species-level, in order to get the apparent abundance-given-occupancy for the identified individuals. Note that this is only possible for individual count data, not subsample count data.

Extension (iv): Modelling regional occupancy

In some cases, there were no detections in any of the sites in a given formation for a species that is otherwise quite detectable in other formations. This suggests that it could be absent from the region at that time because that species had not migrated to the region yet; have permanently or temporarily migrated out of the area; not have originated yet; or have gone extinct.

Because site-occupancy is required for site-detections, and regional occupancy (in a formation) is needed for any occupied sites, we now have a deeper hierarchy of explanations:

- Species detected at a site: both site and regional occupancy are required.
- Zero species detections at a given site, but some detection at other sites in the formation (regional occupancy): Either 1) no detection though there is occupancy at the site (at unmeasured or non-preserved subsamples) or 2) absence at the given site (most parsimonious).
- Zero detection in any of the sites in a formation: Either 1) no detections though there is undetected occupancy at some sites and thus regional occupancy, 2) absence in all the sampled sites but presence at unmeasured sites, hence regional occupancy or 3) regional absence (most parsimonious).

Extension (v): External information concerning regional occupancy

In our dataset, and commonly so in other paleoecological datasets, some species that are quite detectable in some formations have no detections in others. Here, we could consider additional data sources (e.g. collected for other purposes or previously documented) external to the occupancy dataset to inform time-interval specific regional occupancy. If external data with certainty tells us that a certain species is in the region at a particular time, we can set regional occupancy to one for that species; where the external does not tell us that the species is present, we can allow for non-zero probability of regional absence.

Likelihood components

As mentioned in Extension (i), we use the negative binomial distribution to calculate the likelihood for the number of individuals of species *s* in a specific site given occupancy, $y_{i,s} \sim negbinom1(\mu_{i,s}, \kappa)$, where $\mu_{i,s}$ is the expected value and κ is the overdispersion parameter. This is not the standard way of parametrizing the negative binomial distribution, so

we designate it "negbinom1" in eq. (3) and (4) (compare with eq. (7)). We assume the same overdispersion for all species and formation as Reitan et al. 2022 suggested that overdispersion could not be distinguished among species. We also separate the expected value per subsample, $\lambda_{f(i),s}$, from T_i. The probability distribution of a single data point in an occupied site is then:

$$P_{negbinom1}(y_{i,s}|T_i\lambda_{f(i),s},\kappa) = \binom{y_{i,s}+1/\kappa-1}{y_{i,s}} \frac{(\lambda_{f(i),s}T_i\kappa)^{y_{i,s}}}{(1+\lambda_{f(i),s}T_i\kappa)^{y_{i,s}+1/\kappa}} \quad (3)$$

The expected value of this distribution is $\mu_{i,s} = \lambda_{f(i),s}T_i$ and the variance is $\lambda_{f(i),s}T_i(1 + \kappa\lambda_{f(i)s}T_i)$. Thus, the closer the overdispersion is to zero, the closer the variance is to the expected value (as for the Poisson distribution).

However, eq. (3) assumes occupancy. If *s* does not occupy the site, the expected value will be zero and the only possible outcome is $y_{i,s} = 0$. Let the independent probability of site-occupancy of each site belonging to a specific species *s* and formation *f(i)* be designated $\Psi_{f(i),s}$. Then, the distribution of $y_{i,s}$ unconditioned on site-occupancy will be zero-inflated:

$$P_{zero,negbinom1}(y_{i,s}|T_i\lambda_{f(i),s},\kappa,\Psi_{f(i),s}) =$$

$$(1-\Psi_{f(i),s})I(y_{i,s}=0) + \Psi_{f(i),s}\binom{y_{i,s}+1/\kappa-1}{y_{i,s}} \frac{(\lambda_{f(i),s}T_i\kappa)^{y_{i,s}}}{(1+\lambda_{f(i),s}T_i\kappa)^{y_{i,s}+1/\kappa}} \quad (4)$$

Here, $I()$, is the indicator function, which is one if the statement inside the parenthesis is true, and zero, if false. We assume the superspecies occupies all sites.

A species can be absent from all sites in a region in the same formation, thus a non-independent lack of occupancy (*Extension (iv)*). We represent the presence/absence of *s* with a continuous variable $\omega_{f,s} \sim N(\mu = \Phi^{-1}(r), \sigma = 1)$, but only for the species+formation combinations where we do not have external information that the species is present in the

region (*Extension (v)*). r represents the probability of regional presence for the set of species+formation combinations and Φ() is the cumulative distribution function of the standard normal distribution. We then define a binary variable,

$$\Omega_{f,s} = I(\omega_{f,s} > 0 \text{ or } A_{f,s} = 1), \quad (5)$$

which indicates whether the region is occupied, where $A_{f,s} \equiv I(\text{external data sources tell that species } s \text{ occupies formation } f)$. Since $\omega_{f,s}$ is centered around $\Phi^{-1}(r)$, $\Omega_{f,s} = 1$ with probability r whenever $A_{f,s} = 0$. Since site-occupancy depends on regional occupancy, the expression $\Omega_{f,s}\Psi_{f,s}$ replaces $\Psi_{f,s}$ in the zero-inflation part of the likelihood component in eq. (4). We then let r determine the distribution of $\omega_{f,s}$ for cases where $A_{f,s} = 0$ and use likelihood r for the cases where $A_{f,s} = 1$. Hence r will represent the probability for regional occupancy in total, rather than just regional occupancy for those cases where $A_{f,s} = 0$. For each species-formation combination, the likelihood picks up a term

$$L_{f,s} \equiv I\big(A_{f,s} = 1\big)r + I\big(A_{f,s} = 0\big)f_N\big(\omega_{f,s}\big|\mu = \Phi^{-1}(r), \sigma = 1\big), \quad (6)$$

where $f_N()$ is the probability density function of the normal distribution.

With unidentified-to-species-level individuals belonging to a genus, given that there are multiple species of that genus, (shortened as “unidentified” and conversely as “identified”), the probability of the combination of identified and unidentified individuals will be the product of the distribution of the identified individuals and the distribution of the unidentified individuals given the identified ones. The identified individuals are described by eq. (4), though when taking into account the possibility of unidentified individuals, the expected value of identified individuals will be modified to $\gamma_{g,f}\lambda_{f,s}$ where $\gamma_{g,f}$ is the identification

probability of an individual. The number of unidentified individuals, $U_{i,g}$, given the identified individuals, $I_{i,g}$, then follows the negative binomial distribution (see SI for details):

$$P\left(U_{i,g}\middle|I_{i,g}\right) = \binom{U_{i,g} + I_{i,g}}{U_{i,g}} \gamma_{g,f(i)}^{I_{i,g}+1} (1 - \gamma_{g,f(i)})^{U_{i,g}} \quad (7)$$

Final likelihood expression

Since informally $\Pr(\text{identified and unidentified}) = \Pr(\text{unidentified}|\text{identified})\ \Pr(\text{identified})$, the likelihood becomes a product of these two contributions:

$$L = \left(\prod_{s=1}^{S} \prod_{f}^{F} L_{f,s} \prod_{i|f(i)=f} P_{zero,negbinom1}\left(y_{i,s}|T_i \gamma_{g(s),f(i)} \lambda_{f(i),s}, \kappa, \Omega_{f(i),s} \Psi_{f(i),s}\right) \right)$$

$$\left(\prod_{g \epsilon UG} \prod_{s \in g} \prod_{i=1}^{\#sites} P\left(U_{i,s}\middle|I_{i,g}\right) \right) \quad (8)$$

where UG is the set of genera that has unidentified individuals. Note that we now let the expected number of identified individuals for each species scale with identifiability probability of the genus it belongs to, $\gamma_{g,f(i)}$. We set $\gamma_{g,f} = 1$ for each genus where there is no possibility for unidentified individuals (see *Data*).

The likelihood depends on the state of the random effects, both the common formation-dependent random effects for site-occupancy and abundance-given-occupancy respectively, v_f and u_f, as well as the species- and formation-dependent random effects for site-occupancy and abundance-given-occupancy respectively, $\delta_{f,s}$ and $\varepsilon_{f,s}$. The site-occupancy and abundance-given-occupancy component in the likelihood express (eq. 9) are thus

$$\lambda_{f,s}(\theta) = exp(\beta_s + u_{f(i)} + \varepsilon_{f(i),s}) \quad (9a)$$

$$\Psi_{i,s}(\theta) = I(s = S) + I(s < S)\text{logit}^{-1}\left(\alpha_s + v_{f(i)} + \delta_{f(i),s}\right) \quad (9b)$$

θ is the parameter set (random variables and top parameters, see Fig. 1). Here, both abundance-given-occupancy and site-occupancy itself are decomposed into a species-dependent, a species+formation-dependent and a purely formation-dependent random variable, parallel to the original model (eq. 1a). The expression for relative abundance (see eq. 2) is also retained.

Random effects

The random effects for species-dependent dynamics and common dynamics (eq. 1b) are likewise retained in the new model.

$$u_f \sim N(0, \sigma_u^2),\ v_f \sim N(0, \sigma_v^2),\ \delta_{f,s} \sim N\left(0, \sigma_{\delta,s}^2\right),\ \varepsilon_{f,s} \sim N\left(0, \sigma_{\varepsilon,s}^2\right) \quad (10)$$

However, we also include new random effects for the species-dependent constants,

$$\alpha_s \sim N(\mu_\alpha, \sigma_\alpha^2), \beta_s \sim N(\mu_\beta, \sigma_\beta^2) \quad (11a)$$

$$\sigma_{\delta,s} \sim logN(\mu_\delta, \sigma_\delta^2),\ \varepsilon_{\delta,s} \sim logN(\mu_\varepsilon, \sigma_\varepsilon^2),\ \text{for } s < S \text{ (superspecies exempted)} \quad (11b)$$

where the original species-dependent constants effects (eq. 11a) and the size of the dynamics (eq. 11b) are now both random factors. Note that the size of the superspecies dynamics for abundance-given-occupancy, $\sigma_{\varepsilon,S}$, is not part of this equation but is instead a top parameter. As the superspecies is an aggregate of many different species, it can be expected to be less dynamic than any single species. The information content of the superspecies is much be greater than for any other species. We hence exclude it in eq. 11 to avoid swamping of random effect parameters for species dynamics.

290 Since we have one identifiability probability for each combination of formation and genera
291 with unidentified colonies, we let it be a random factor, just like the other components in our
292 model that describes dynamics:

293 $$logit(\gamma_{g,f}) \sim N(\mu_\gamma, \sigma_\gamma^2). \quad (12)$$

294 *Top parameters and prior distributions*

295 With our current parametrization, the top parameters are

296 $$\theta_{top} = \{\mu_\alpha, \sigma_\alpha, \mu_\beta, \sigma_\beta, \mu_\delta, \sigma_\delta, \mu_\varepsilon, \sigma_\varepsilon, \sigma_u, \sigma_v, \sigma_{\varepsilon,S}, r, \kappa, \mu_\gamma, \sigma_\gamma\}. \quad (13)$$

297 Note that this parameter set does not increase with an increasing number of species, so the
298 number of top parameters is always 15. For comparison, the Reitan et al. 2022 model had $5 \times S$
299 top parameters, which for our dataset, S=21, would have translated to 105 top parameters.
300 Even so, there was no way of dealing with the genera that has unidentified individuals in that
301 model. For details of our choice of prior distributions and the robustness of our model to our
302 choice of prior, see SI.

303

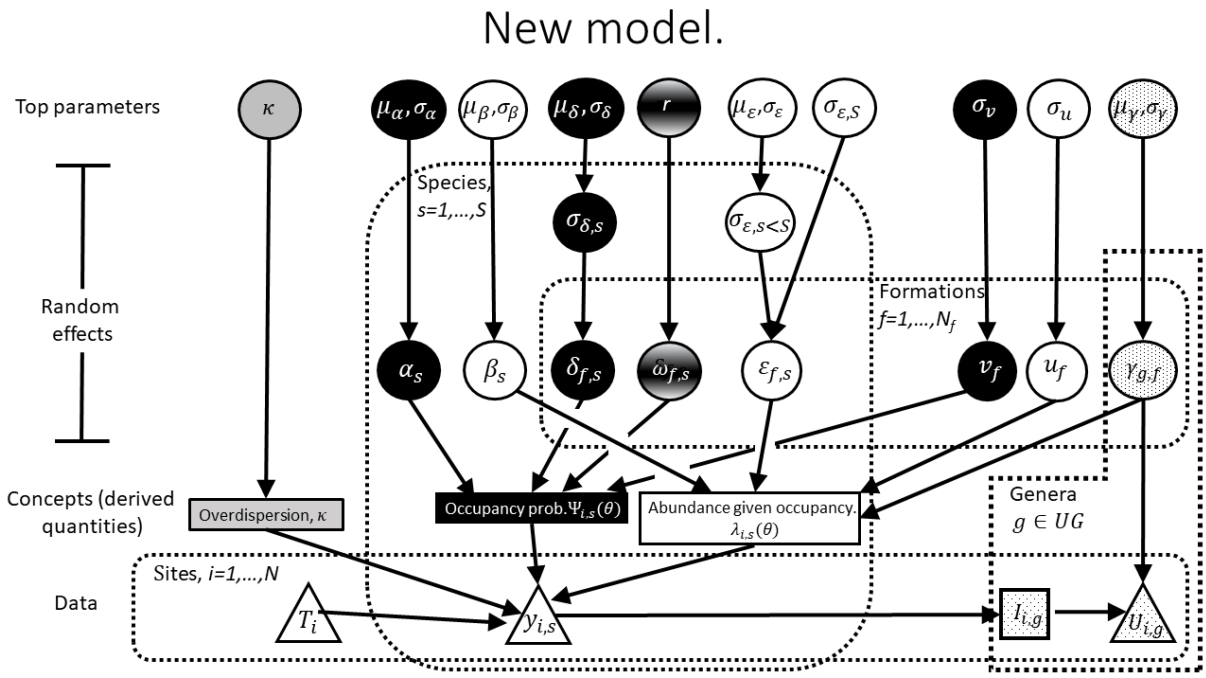


Figure 1: A schematic view of the new model. This overview shows the hierarchical relationships between data, the core components of the occupancy model, random effects and top parameters. The arrows show dependencies. Shapes with white background are associated with abundance-given-occupancy or base data (individual and subsample counts, excluding data associated with taxon identifiability probability). Shapes with black backgrounds are associated with occupancy (solid black for site-dependent occupancy and gradient black for regional occupancy). Shapes with grey backgrounds are associated with overdispersion. Lastly, shapes with dotted backgrounds are associated with taxon identifiability probabilities. Round shapes are parameters/random effects, rectangles are concepts expressed as functions and triangles are data. How the regional occupancy random effects, $\omega_{f,s}$, determines the regional occupancy states are not shown here (but see eqs. 5, 6 and 8). The functions $\lambda_{f,s}(\theta)$ and $\Psi_{i,s}(\theta)$ are expressed in eq. 9. Note also $I_{i,g}$ is a sum of the species data, $y_{i,g}$, for each genus with unidentified colonies, shown as a separate entity because this aggregate is used in a separate part of the likelihood.

Simulation 1: New model performance

To explore the performance of the new model, specifically to examine the accuracy of the inference of not just relative abundance but site-occupancy, regional occupancy and abundance-given-occupancy using individual counts, we set up simulations. We also incorporated all the extensions, namely unidentified individuals, regional occupancy and extra sources pertaining to regional occupancy, in order to test whether the model was able to handle these challenges. See SI for details.

Simulation 2: Are individual counts better than subsample counts?

We use a different set of simulations to test if individual counts perform measurably better than subsample count data (*Extension (i)*). Here, our simulated datasets had a specified site-occupancy probability and abundance-given-occupancy, which gives the relative abundance. We sampled simulated data on the subsample level and then aggregated these to site-level in the form of both individual counts and subsample presence counts. We also wanted to see how well relative abundance estimated from simple ratios worked (i.e. "raw estimates" as opposed to model estimates). We used the occupancy model from Reitan et al. 2022 for the subsample presence counts data and the new model described here for the individual counts. In addition, we used this set of simulations to examine the effect of different levels of observational error (i.e. missing individuals, double counting of individuals and misclassification of species). We judged how well these methods worked using the root-mean-squared-error (RMSE) of the relative abundances. See SI for details. All data and code are supplied in the zip folder "RAMU-MSOM" available via the editor/Ecography office.

Results

Simulation 1

The relative abundance estimates correspond well with the true relative abundance and respond well to regional absence (Fig. 2). The modelled relative abundance estimates had an $RMSE \approx 0.016$. When the existence of unidentified individuals was ignored, $RMSE \approx 0.021$. Thus, the effort to compensate for the unidentified individuals did pay off. Raw estimates had $RMSE \approx 0.024$, both when attempting to compensate for unidentified individuals (by dividing by the ratio of unidentified individuals in each genus) and when not attempting this, suggesting that it is not so easy to do this type of compensation using raw estimates. One cannot expect the latter to converge to true values with increasing data size,

347 though from theory alone we would expect raw estimates corrected for unidentified
348 individuals to converge. However, with our current data volume, identifiability correction in
349 raw estimates do not work better than those without such corrections. Even if the corrected
350 raw estimates do converge, one would need 2.3 times as many data points (sites) to obtain
351 errors as small as the model estimated ones, regardless of absolute data volume (assuming that
352 the squared error is inversely proportional to the dataset size).

353 Site-occupancy dynamics are quite well-estimated for the most abundant species (first in each
354 simulated genus) while the least abundance species (last in each simulated genus) which
355 likewise had a very dynamic true site-occupancy trend, were not (e.g. compare G01_S01 and
356 G01_S04 in Fig. 3). Although the site-occupancy dynamics of species with intermediate
357 abundance (e.g. G01_S02 and G01_S03) are also not too well-captured by the estimates,
358 some of it is absorbed into estimated abundance-given-occupancy (SI Fig. S1). Regional
359 occupancy probability was also sometimes estimated to be low for some species+formation
360 combinations in particular datasets where there were no detections, even though the region
361 was actually occupied. However, when looking at the average score over all datasets, the
362 regional occupancy probabilities are reasonable (Fig. S2).

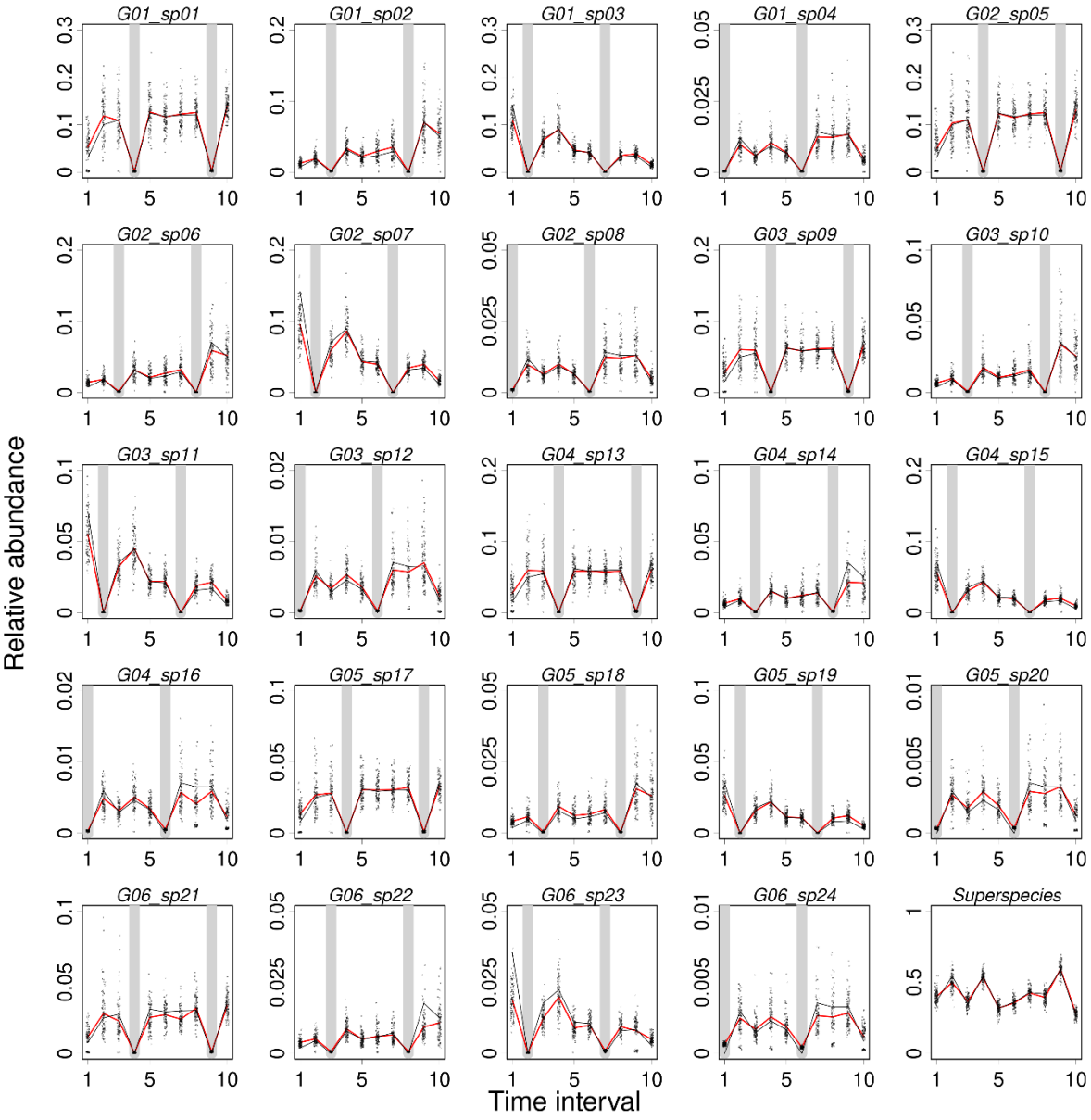


Figure 2: Relative abundance estimates for simulated data. Relative abundance estimates for simulated data (Simulation 1) for individual species are presented in each panel. Solid black lines=true values, red lines=average estimates from 100 simulations, dots=estimates for each simulated dataset, grey vertical bars=true regional absence. Note the different y-axes. The designated species names are shown on top of each panel.

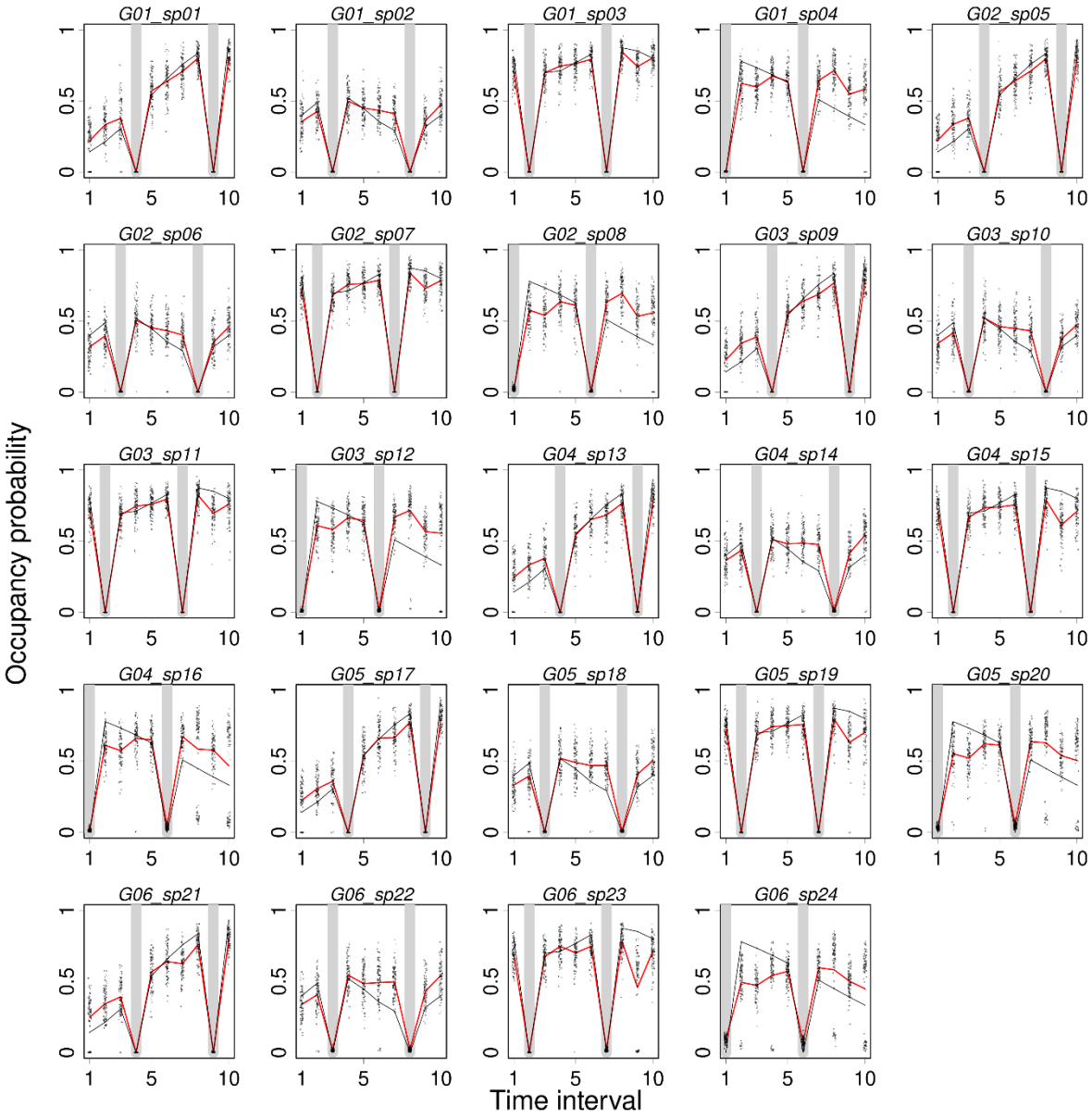


Figure *3*: Occupancy probability estimates for simulated data. Occupancy estimates for simulated data (Simulation 1) for individual species are presented in each panel. Solid black lines=true values, red lines=average estimates from 100 simulations, dots=estimates for each simulated dataset, grey vertical bars=true regional absence. The designated species names are shown on top of each panel.

Simulation 2

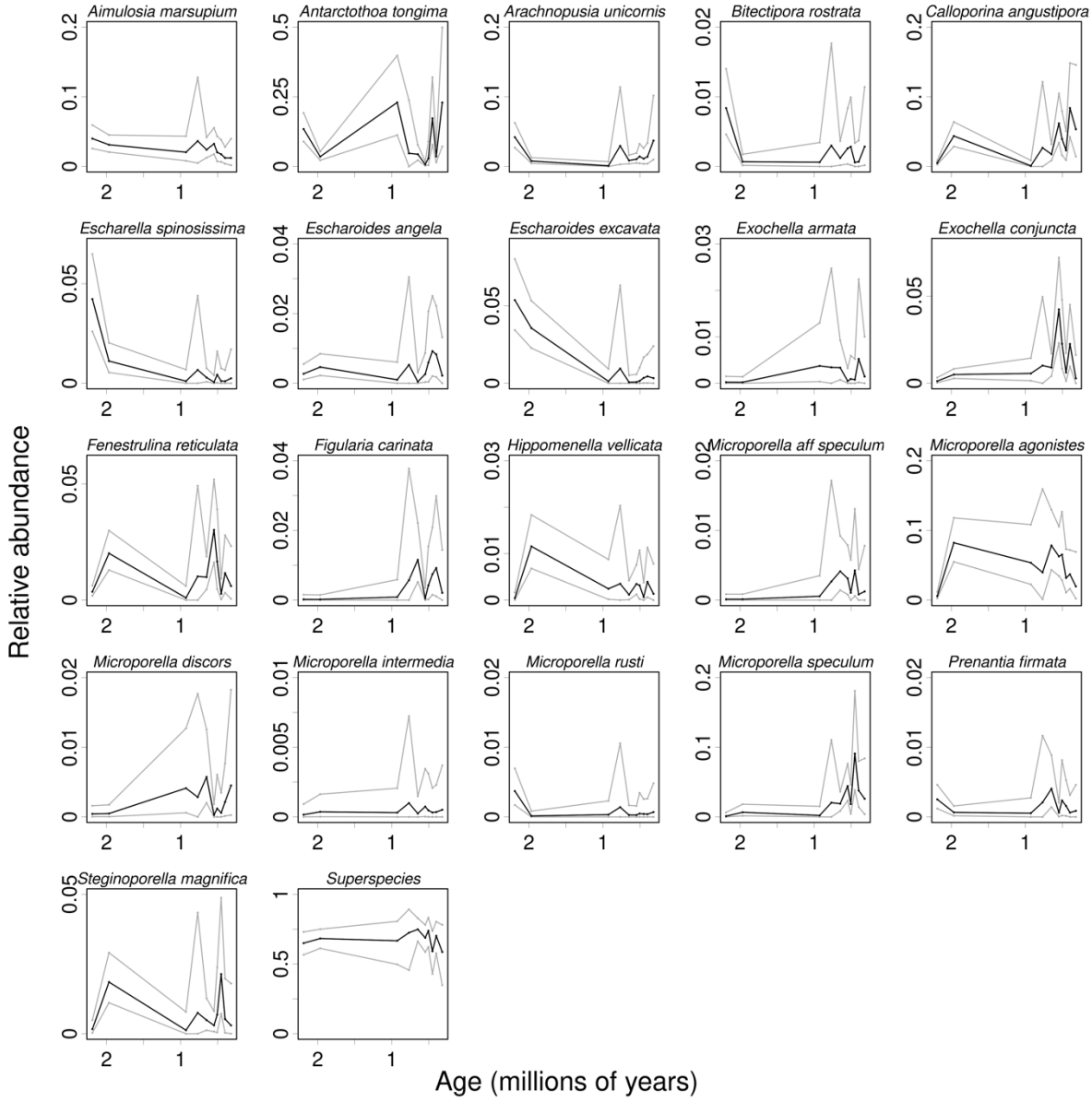
The RMSE of the relative abundance estimates were smallest for model estimates of individual count data ($RMSE \approx 0.023$). Compared to the model estimates for individual count data, the RMSE's for raw estimates for individual count data, for model estimates for subsample count data and the raw estimates for subsample presence count data were 26%,

59% and 285% higher, respectively. We would need 59%, 153% and 1382% more data points for raw estimates on individual count data, model estimates on subsample count data and raw estimates on subsample count data, respectively, to lower the errors to the level of model estimates on individual count data. Here, we assume the standard error to be inversely proportional to the square root of the number of measurements. However, raw estimates on subsample count data cannot be expected to converge towards unbiased results when the number of data increases, as the ratio of subsamples having presence of a given species does not scale linearly with abundance-given-occupancy (Reitan et al. 2022).

The observational error simulations suggested that the relationship between the various RMSEs does not substantially change when the probability of observational errors increased. (SI for details).

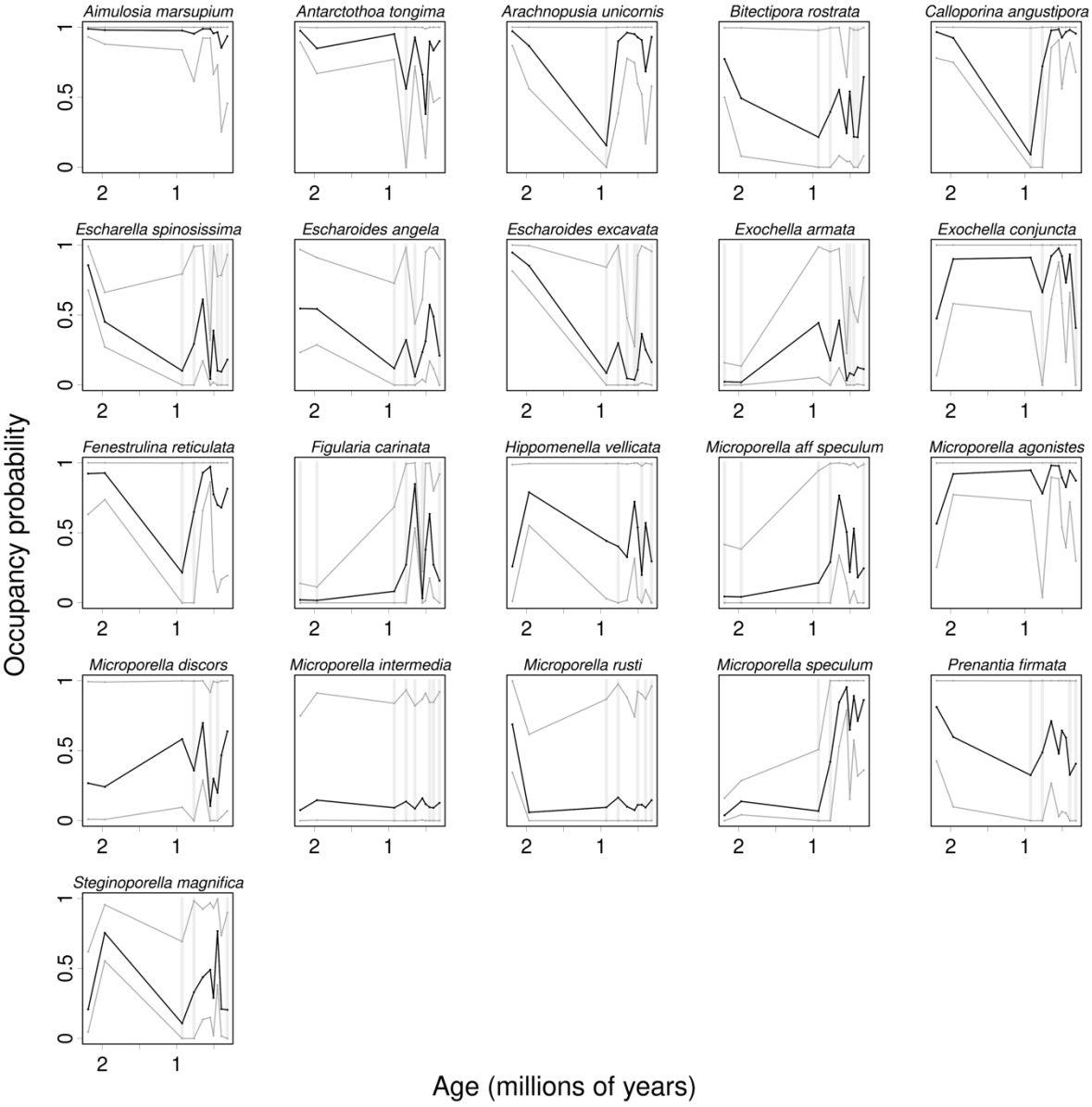
Empirical results

While this work focuses on the details of the new model and simulations for understanding the performance of the model, it was of interest to ensure that the model has empirical relevance. Very briefly, there are clear species-specific temporal dynamics (i.e. non-overlapping credibility bands) in both estimated relative abundance (Fig. 4) and occupancy (Fig. 5) in our empirical dataset. The dynamics of relative abundance and occupancy are appreciably different for species within the same genera (e.g. compare *Microporella speculum, M. agonistes, M. discors*; compare *Escharoides excavata* and *E. angela*). Our model-estimated relative abundances are also robust to different prior widths (see SI).



398

399 **Figure 4: Relative abundance estimates for the empirical dataset.** Each panel shows the point estimates (black lines) and
400 95% credibility bands (grey lines) for the 21 focal species plus the superspecies. Note the different y-axes.



401

402 ***Figure 5:*** **Occupancy estimates for the empirical dataset.** Each panel shows the point estimates (black lines) and 95%
403 credibility bands (grey lines) for the 21 focal species plus the superspecies. Light grey bars indicate no detections in that
404 formation and no indication of presence from external sources, i.e. situations where regional absence is a real possibility.
405 Note the different y-axes.

406 ## *Discussion*

407 Hierarchical site-occupancy modelling is currently still rarely applied to paleoecological
408 datasets, yet prevailing issues of incomplete detection in paleoecology is rampant, just like in
409 ecological studies where occupancy modelling is more commonly applied. Replicate sampling
410 and subsampling within formations is currently not standard practice in paleoecology. We

411 have shown that there are measurable differences in face-value (raw ratios) and model
412 estimates that will impact not just quantitative but also qualitative inferences. However, there
413 is a practical need to strike a balance between the precision and accuracy of parameter
414 estimation and the effort required for data collection. For instance, it is quicker to count
415 subsamples containing focal species, rather than painstakingly counting individuals of those
416 species. However, much is gained in counting individuals rather than just occupied
417 subsamples when estimating relative abundance. In addition, individual counts are crucial
418 when there are individuals unassignable to genera, a situation common in paleoecology. As
419 far as we are aware, ours is the first attempt at explicitly incorporating information on
420 individual unassignable to species while estimating relative abundance and occupancy using
421 paleontological data. Encouragingly, not only do our simulations show that we can recover
422 relative abundance dynamics by explicitly incorporating information on individuals identified
423 to genus- but not species-level, we also recover relative abundance and occupancy dynamics
424 in our empirical data (see Figs. 4 and 5, e.g. species of *Microporella*).

425 There are, of course caveats to the estimates, evident from both simulations and the empirical
426 data analyses. Most notably, dynamics are most recoverable for species that are most
427 commonly observed (i.e. the most prevalent species) in the simulations and hence we have to
428 assume that is the case also for the empirical dataset. That said, less prevalent species still
429 contribute to information important for estimation of more prevalent species through
430 parameters common to all species. How important regional occupancy modelling is depends
431 both on the occupancy data and the "external information" available, which will vary from
432 dataset to dataset. In any case, evidence for regional absence in our empirical system is weak
433 in some cases (Fig. 5), as can be seen from our top parameter posterior distributions and
434 robustness analyses (see SI). Absence is in general more difficult to infer than presence, since

some observed absences are due to detection probability rather than true absence. But absence is not impossible to estimate, as we have shown.

Lest one erroneously concludes that a simpler model can be used for estimating relative abundance in a given area, let us be clear that site-occupancy modelling that teases apart occupancy and detection is a necessary component in estimating abundance. Additionally, one in general does not know whether regional absence is possible before analysing the empirical occupancy data. It is important to replicate sampling in ways that will capture variation in detection since absence of information cannot be proof of absence. In our case, we found clear indications of site absence, but not regional absence. Our model can be applied more widely in paleoecology than is perhaps apparent with our example empirical dataset. For instance, deep-sea cores can be subsampled within time-intervals, as estimated by a combination of depth information and age-models based on sedimentation rates, as can be lake sediment cores. More generally, any regional system where multiple outcrops in which sampling can be replicated will be amendable to this occupancy modelling. We recommend subsampling/replicate-sampling sites within formations/time-intervals for occupancy and abundance estimation for paleoecological systems, even when multiple sites cannot easily be sampled within formations. We also urge detailed documentation of individuals. These data, while requiring a bit more work to collect, can yield vastly better estimates of key ecological parameters.

References cited

Carter, R. M. and Naish, T. R. 1998. A review of Wanganui Basin, New Zealand: global reference section for shallow marine, Plio-Pleistocene (2.5-0 Ma) cyclostratigraphy. - Sedimentary Geology 122: 37–52.

459 Dillon, E. M. et al. 2022. What is conservation paleobiology? Tracking 20 years of research
460 and development. - Frontiers in Ecology and Evolution in press.

461 Dussex, N. et al. 2021. Integrating multi-taxon palaeogenomes and sedimentary ancient DNA
462 to study past ecosystem dynamics. - Proc Biol Sci 288: 20211252.

463 Hoban, S. et al. 2019. Inference of biogeographic history by formally integrating distinct lines
464 of evidence: genetic, environmental niche and fossil. - Ecography 42: 1991–2011.

465 Jackson, J. B. C. and Cheetham, A. H. 1990. Evolutionary significance of morphospecies - a
466 test with cheilostome Bryozoa. - Science 248: 579–583.

467 Lawing, A. M. et al. 2021. Occupancy models reveal regional differences in detectability and
468 improve relative abundance estimations in fossil pollen assemblages. - Quaternary
469 Science Reviews 253: 106747.

470 Liow, L. H. 2013. Simultaneous estimation of occupancy and detection probabilities: an
471 illustration using Cincinnatian brachiopods. - Paleobiology 39: 193–213.

472 Liow, L. H. et al. 2016. Interspecific interactions through 2 million years: are competitive
473 outcomes predictable? - Proceedings of the Royal Society B-Biological Sciences 283:
474 20160981.

475 MacKenzie, D. I. et al. 2002. Estimating site occupancy rates when detection probabilities are
476 less than one. - Ecology 83: 2248–2255.

477 Pillans, B. 2017. Quaternary stratigraphy of Whanganui Basin—a globally significant archive.
478 - In: Shulmeister, J. (ed), Landscape and Quaternary Environmental Change in New
479 Zealand. Atlantis Press, pp. 141–170.

480 Proust, J. N. et al. 2005. Sedimentary architecture of a Plio-Pleistocene proto-back-arc basin:
481 Wanganui Basin, New Zealand. - Sedimentary Geology 181: 107–145.

482 Reitan, T. et al. 2022. Relative species abundance and population densities of the past:
483 developing multispecies occupancy models for fossil data. - Paleobiology 49: 23–38.

484