

In press in: Methods in Ecology & Evolution: NB This is not the final version of record.

## Article Type: APPLICATION

### Extension of the gambin model to multimodal species abundance distributions

Thomas J. Matthews<sup>1,2,3</sup>, Michael K. Borregaard<sup>4</sup>, Colin S. Gillespie<sup>5</sup>, François Rigal<sup>6</sup>, Karl I. Ugland<sup>7</sup>, Rodrigo Ferreira Krüger<sup>8,9</sup>, Roberta Marques<sup>9</sup>, Jon P. Sadler<sup>1</sup>, Paulo A.V. Borges<sup>2</sup>, Yasuhiro Kubota<sup>10,11</sup>, Robert J. Whittaker<sup>4,8</sup>

<sup>1</sup>GEES (School of Geography, Earth and Environmental Sciences), The University of Birmingham, Birmingham, B15 2TT

<sup>2</sup>CE3C – Centre for Ecology, Evolution and Environmental Changes/Azorean Biodiversity Group and Univ. dos Açores – Depto de Ciências e Engenharia do Ambiente, PT-9700-042, Angra do Heroísmo, Açores, Portugal.

<sup>3</sup>Birmingham Institute of Forest Research, University of Birmingham, Edgbaston, Birmingham, B15 2TT, UK

<sup>4</sup>Center for Macroecology, Evolution and Climate, Natural History Museum of Denmark, University of Copenhagen, Universitetsparken 15, DK-2100 Copenhagen Ø, Denmark

<sup>5</sup>School of Mathematics & Statistics, Newcastle University, Newcastle upon Tyne, NE1 7RU UK

<sup>6</sup>CNRS-Université de Pau et des Pays de l'Adour, Institut des Sciences Analytiques et de Physico-Chimie pour l'Environnement et les Matériaux, MIRA, Environment and Microbiology Team, UMR 5254, BP 1155, 64013 Pau Cedex, France

<sup>7</sup>Department of Marine Biology, Institute of Biosciences, University of Oslo, P.O. Box 1066, Blindern, 0316 Oslo, Norway

<sup>8</sup>Conservation Biogeography and Macroecology Group, School of Geography and the Environment, University of Oxford, South Parks Road, Oxford, OX1 3QY, UK

<sup>9</sup>Parasite and Vector Ecology Group, Depto de Microbiologia e Parasitologia, Campus Universitário Capão do Leão, s/nº CEP 96010-900, Pelotas, Rio Grande do Sul, Brasil

<sup>10</sup>Faculty of Science, University of the Ryukyus, Nishihara, Okinawa, Japan

<sup>11</sup>Marine and Terrestrial Field Ecology, Tropical Biosphere Research Center, University of the Ryukyus, Nishihara, Okinawa 903-0213, Japan.

\*Correspondence: Thomas J. Matthews, School of Geography, Earth and Environmental Sciences, University of Birmingham, Birmingham, B15 2TT, UK

Email: txm676@gmail.com

Running header: Multimodal gambin distributions

Word count: abstract = 231 words; main text = 3149 words; 0 Tables; 2 Figures; 4 appendices

## **Keywords**

Compound distributions, gambin, horse flies, multimodal species abundance distributions

## **ABSTRACT**

**1.** Species abundance distributions (SADs) are one of the most widely used tools in macroecology, and it has become increasingly apparent that many empirical SADs can best be described as multimodal. However, only a few SAD models have been extended to incorporate multiple modes and no software packages are available to fit multimodal SAD models. In this study, we present an extension of the gambin SAD model to multimodal SADs.

**2.** We derive the maximum likelihood equations for fitting the bimodal gambin distribution and generalise this approach to fit gambin models with any number of modes. We present these new functions, along with additional functions to aid in the analysis of multimodal SADs, within an updated R package ('gambin'; version 2.4.0) that enables the fitting, plotting and evaluating of gambin models with any number of modes.

**3.** We use a mixture of simulations and empirical datasets to test our new models, including tests of the sensitivity of the model parameters to the number of individuals and the number of species in a sample. We show that the new multimodal gambin models perform well under a variety of circumstances, and that the application of these new models to empirical SAD and other macroecological (e.g. species range size distributions) datasets can provide interesting insights. The updated software package is simple to use and provides straightforward yet flexible statistical analyses of multimodality in SAD-type datasets.

## **INTRODUCTION**

The species abundance distribution (SAD) has been a core focus of macroecology for over eighty years (e.g. Fisher, Corbet & Williams 1943), and is currently the subject of widespread renewed interest (McGill et al., 2007; Alonso, Etienne & Ostling 2008; Arellano et al., 2017). Recently, it has been argued that a gamma-binomial (herein 'gambin') distribution represents a useful SAD model (Ugland et al., 2007). Gambin is a stochastic unimodal model that combines the gamma distribution, in which the scale parameter is fixed at 1, with a binomial sampling method (see Ugland et al., 2007 for a full description of the model). The use of the

gamma distribution as the basis of the model provides gambin with substantial flexibility and tests of the gambin model have found that it generally provides a good fit to a wide range of empirical SAD data, typically out-performing other candidate SAD models (Ugland et al., 2007; Matthews et al., 2014), such as the Poisson lognormal (PLN; Bulmer, 1974) and logseries models (Fisher et al., 1943). The model can also be used with continuous data, and thus extend the analysis of SADs to different measures of abundances (e.g. biomass). The unimodal gambin model has a free parameter ( $\alpha$ ) that determines the shape of the distribution. Low values of  $\alpha$  indicate logseries curve shapes, whilst higher  $\alpha$  values indicate more lognormal-like curve shapes. Thus,  $\alpha$  is an intuitive parameter that has been found to be of use in comparing the SAD of different ecological communities, e.g. disturbed and undisturbed communities, and for testing what variables drive changes in the shape of the SAD along ecological gradients (Dornelas, Soykan & Ugland 2011; Matthews & Whittaker, 2015; Arellano et al., 2017; Matthews, Borges, de Azevedo & Whittaker 2017). Due to the way the statistical model is defined, gambin can only be fitted to data binned into octaves e.g. classes of  $\log_2$  transformed abundance data, with octave 0 containing the number of species with 1 individual, octave 1 the number of species with 2 or 3 individuals, and so on.

It has become increasingly apparent that many empirical SADs can best be described as multimodal (Dornelas & Connolly, 2008; Vergnon, van Nes & Scheffer 2012; Antão, Connolly, Magurran, Soares & Dornelas 2017). For example, Antão et al. (2017) found that between 15% and 22% of the 117 empirical SAD datasets they evaluated showed evidence of multimodality, depending on the model selection tools used. Multimodality may be indicative of particular process regimes (Matthews, Whittaker & Borges 2014) or be due to a combination of different types of species (e.g. trophic groups) in a sample, and its detection may also be relevant to, for example, tests of the theory of emergent neutrality (Vergnon et al., 2012). Hence, describing and testing for multimodality is a priority in SAD research (Antão et al., 2017). To date, few SAD models have been extended to incorporate multiple modes (for the PLN see Dornelas & Connolly, 2008), in part because compound probability distribution models are mathematically and computationally complex. Hence the need for an easy-to-use software package permitting straightforward statistical analysis of multimodality in SAD datasets. We set out to provide a multimodal extension of gambin because the gambin model is relatively simple and it would allow comparison of the fit of unimodal and multimodal models analytically using standard statistical methods.

First, we derive the maximum likelihood equations for fitting gambin models with any number ( $g$ ) of components and incorporate these new functions, along with additional functions to aid in the analysis of multimodal SADs, within an updated version of the R package gambin (version 2.4.0). Second, we use a mixture of simulations and empirical datasets to test the new models, providing examples of the updated package in operation.

## **MULTIMODAL GAMBIN DISTRIBUTIONS AND THE GAMBIN R PACKAGE (VERSION 2.4.0)**

The full derivation of the likelihood functions for multimodal gambin models is provided in Appendix S1 (Supplementary Information). In version 2.4.0 of the gambin R package, the one-component gambin model is taken to have two parameters: the shape parameter ( $\alpha$ ) and the max octave. It should be noted that this differs from previous implementations of the model (e.g. Matthews et al., 2014) that only considered there to be a single parameter ( $\alpha$ ). The two-component gambin model is simply the mixture of two gambin distributions. To allow for the subdivision of all of the observed objects (*species* in the context of SADs) ( $y_{obs}$ ), a parameter ( $w_1$ ) is needed that describes the fraction of objects belonging to the first distribution ( $w_i$  is analogous to the  $p$  parameter in the multimodal PLN context). The fraction of objects belonging to the second component ( $w_2$ ) is  $1 - w_1$ . Thus, the expected number of observed objects is split into two components, consisting of  $w_1 * y_{obs}$  and  $w_2 * y_{obs}$  objects, respectively. Thus,  $y_{obs} = (w_1 * y_{obs}) + (w_2 * y_{obs})$ . With no extra information, we may therefore assume that the number of objects in the  $k$ -th interval ( $k = 1, 2, \dots, i$ ) are  $w_1 * y_k$  and  $w_2 * y_k$ . Thus, the likelihood function for a bimodal gambin model contains five parameters: the shape parameters for the first and second component ( $\alpha_1$  &  $\alpha_2$ ), the max octaves for the first ( $n_{oct1}$ ) and second ( $n_{oct2}$ ) components, and one splitting parameter ( $w_1$ ) representing the fraction of objects in the first component. Note that this is the same number of parameters in the bimodal PLN model; it is simply that the parameters represent different aspects of the distribution in each case. It is relatively straightforward to extend the above approach for fitting the two-component gambin model by maximum likelihood, to fitting gambin models with  $g$  components (where components correspond to the number of modes; see Appendix S1). However, whilst it is possible to use the equations given in Appendix S1 to fit gambin distributions with any number of components, in practice fitting SAD models with more than three (possibly even two depending on sample size) components will likely result in overfitting the data. Sample sizes in ecological studies are generally relatively small, and the number of parameters becomes large with increasing  $g$  (Dornelas & Connolly, 2008). Thus, optimising the likelihood functions becomes increasingly problematic at larger  $g$ ; ecological interpretation of model fits with large numbers of components is also problematic. Accordingly, we do not advise fitting gambin models with more than three components.

In addition to providing functions to fit multimodal gambin distributions (described below), the gambin R package (version 2.4.0; available on CRAN) has been updated to bring it more in line with other distribution functions within the R base ‘stats’ package. For example, the updated gambin package now provides *dgambin* (probability density function), *rgambin* (generate random values from a gambin distribution; the returned values relate to a given octave), *qgambin* (quantile function) and *pgambin* (cumulative distribution function) functions. Likelihood optimisation is undertaken using the Nelder–Mead algorithm. As the likelihood optimisation procedure for multimodal gambin models can be time consuming, the updated package provides the option of using parallel processing to speed up optimisation. The gambin R package documentation and associated vignette provide additional information.

The main function within the package is ‘fit\_abundances’:

```

150 #this fits a gambin distribution with g modes to a vector of abundances,
151 #with the option of subsampling z individuals. If g is set to 1, the
152 #standard unimodal gambin distribution is fitted, g = 2 fits the bimodal
153 #gambin distribution, and so on. When the no_of_components argument is
154 #greater than 1, the 'cores' argument can be used to enable parallel
155 #processing using d cores.

```

```

156 fit_abundances(data, subsample = z, no_of_components = g, cores = d)

```

A primary argument for the prevalence of multimodal SADs in nature is the idea that the different modes represent different categories of species (e.g. native and invasive species, or core and satellite species; Magurran & Henderson 2003; Matthews & Whittaker 2015). A natural next step then is to deconstruct the SAD by visualizing and analysing how different categories of species are distributed across the various modes / modal octaves. This is performed with the new function 'deconstruct\_modes'. If species category information is provided (e.g. native or invasive), the function returns the number and proportion of the various categories in the different modal octaves (a split barplot where the bar for each octave is split according to the number of species in each category can also be returned). Subsequent statistical test (e.g.  $\chi^2$  or G-test) and/or null model tests can then be undertaken to determine whether the number of species representing the different categories significantly differs between octaves. If species category information is not available, the function will simply identify the modal octaves (i.e. the modal octave of each component distribution) in a multimodal gambin model fit (user-specified modal octaves can instead be provided), and also lists the names of the species within each octave (a plot of the model fit with the modal octaves highlighted can also be returned).

```

173 #Fit the bimodal gambin model to SAD data

```

```

174 fit <- fit_abundances(data, no_of_components = 2)

```

```

175 #Deconstruct the model fit and calculate the number of species of
176 #different categories (categ) in each of the modal octaves (peak_val is
177 #set to 'NULL' and thus the modal octaves are calculated from the model
178 #fit). Return a plot of the model fit with the modal octaves highlighted
179  #(plot_modes = TRUE) and run the null model bootstrap sampling with 100 (n
180 #= 100) random draws.

```

```

181 deconstruct_modes(fit, dat = data, peak_val = NULL, categ = "status",
182                   plot_modes = TRUE, n = 100)

```

One of the main applications of the gambin model has been to fit gambin to SADs from different sites (e.g. along a disturbance gradient) and then to compare the resultant alpha values (e.g. Dornelas, Soykan & Ugland 2011; Arellano et al., 2017). Thus, we have also added a function that fits the unimodal gambin model to the SADs from multiple sites and returns the standardised and unstandardised alpha values.

```

188 #Fit the unimodal gambin model to the SADs from multiple sites (mult) and
189 #return the standardised (based on N subsamples of size 'subsample'; NULL
190 # = the number of individuals in the site with the fewest individuals) and
191 #unstandardised alpha values

```

```

192 mult_abundances(mult, N = 100, subsample = NULL)

```

```

193

```

## 194    **EXAMPLES USING EMPIRICAL DATASETS**

### 195    **A Brazilian horse fly dataset**

196    To illustrate the new functionality, we used an empirical dataset comprising abundance  
197    records of horse flies (Diptera, Tabanidae) from a variety of sampling locations in Brazil (see  
198    Appendix S2 in the Supplementary Information). As outlined above, multimodal SADs may  
199    hypothetically arise from the intersection in nature of samples from different habitat types or  
200    of different ecological species groups (Magurran & Henderson, 2003; Antão et al., 2017)  
201    within a dataset. To test this proposition, we first fitted the unimodal, bimodal and trimodal  
202    versions of gambin to the whole Brazilian dataset. We then took a subset of the dataset  
203    relating to one individual locality within Brazil and one type of sampling (see Appendix S2)  
204    and again fitted the three models. In both cases the three models were compared using the  
205    Bayesian information criterion (BIC):

```
206    #load the fly datasets  
207    data(fly)  
208    Brazil <- fly[[1]]#select the data for all of Brazil  
209    site <- fly[[2]]#select the data for a single site within Brazil  
210    #Fit the multimodal gambin models to a given dataset (Brazil or site)  
211    res1 <- lapply(c(1, 2, 3), fit_abundances, abundances = Brazil, subsample  
212    = 0, cores = 3)  
213    #calculate and compare the BIC value of the fitted models  
214    vapply(res1, BIC, FUN.VALUE = numeric(1))  
215    #plot the empirical SADs  
216    barplot.gambin(res1[[1]])  
217    points.gambin (res1[[1]], pch = 17, col = "black") #add the fitted values  
218    points.gambin (res1[[2]], pch = 16, col = "blue")  
219    points.gambin (res1[[3]], pch = 18, col = "green")  
220  
221
```

222    When the three models were fitted to the whole Brazilian horse fly dataset (Fig. 1a), the  
223    bimodal gambin model provided the best fit to the data (BIC = 830.4), followed by the  
224    unimodal model (BIC = 832.5) and the trimodal model (BIC = 837.6). When the three  
225    models were fitted to the subset of data from a single site (Fig. 1b), the unimodal model  
226    provided the best fit (BIC = 236.2), followed by the bimodal model (BIC = 239.9) and the  
227    trimodal model (BIC = 246.5). Thus, whilst the data from a single site are characterised by a  
228    classical unimodal SAD, when pooling records from different localities across Brazil, the  
229    bimodal model was favoured. These findings provide additional support for the claim that  
230    multimodal SADs are more prevalent with increasing taxonomic breadth, sampling variation,  
231    spatial extent (i.e. increasing ecological heterogeneity; Antão et al., 2017), and heterogeneity  
232    in species detectability (Alonso et al. 2008).

### 233    **A set of 275 woody plant SADs**

We took the set of 843 angiosperm woody plant datasets sourced from the literature by Kubota et al. (2018). Each dataset represents an abundance vector of plant species sampled in a forest plot and the datasets have a global distribution. We filtered out datasets with <10 species and <500 individuals. We then fitted the unimodal and bimodal gambin models to the resulting 275 datasets and compared the fits using BIC. The bimodal model was considered as the best fitting model if it had the lowest BIC value and the unimodal model had a  $\Delta$ BIC value of >2.0 (a lower value indicates the models have similar support, in which case the unimodal model should be preferred on grounds of parsimony).

The bimodal model provided the best fit to 51 of the 275 datasets (19%; see Appendix S3 for the full model comparison results).

### **Application to other macroecological phenomena**

Whilst gambin models have so far only been used to analyse SADs, it is possible to fit them to any other type of ecological or general distribution. For example, there is evidence that some species-range size distributions may exhibit multimodality (e.g. see Gaston, 2003, p. 80). As an illustration, we fitted a selection of gambin models to the global range size distribution of 167 marine mammal species, and the occupancy distribution of intestinal helminths in three species of grebe; we observe evidence of multimodality in both distributions. The full methods and resultant model fits are provided in Appendix S3.

### **SIMULATION ANALYSES**

The results of our simulations indicated that in general the  $\alpha$  parameter estimates of the bimodal gambin model were relatively insensitive to the number of species in the sample (Figure S2, Appendix S4).

In contrast, it was found that the  $\alpha$  parameter estimates of the bimodal gambin model were sensitive to the number of individuals in a sample (Figs S3 and S4, Appendix S4). The latter is true of most SAD models (see McGill, 2011) and is worrying given that SAD analyses typically involve small datasets. While this sensitivity is problematic for unimodal gambin, it is less of an issue for applications of the bimodal model. With the unimodal gambin model, the  $\alpha$  value can be used as a type of diversity metric to compare SAD shape across communities (e.g. Arellano et al., 2017). However, for multimodal gambin models the meaning of the  $\alpha$  values is not as clear, and as such, when fitting multimodal gambin models we do not advise using the  $\alpha$  parameter estimates as diversity metrics or as response variables in regression-type comparative analyses. Rather, the benefit of multimodal gambin models is to provide a simple, quick and easy to use test for determining whether empirical SADs are multimodal, and to provide a basis for subsequent deconstruction analysis to examine the identities of species within the octaves.

To test the error rate of our models, we simulated unimodal and bimodal gambin SADs using multiple simulations varying the numbers of individuals and species (Appendix S4), fitting both unimodal and bimodal gambin models to the simulated data. We compared models using BIC and calculated the proportion of times that a bimodal model provides a better fit than a

unimodal model to a unimodal dataset (i.e. false positive) and the proportion of times a unimodal model provides a better fit than a bimodal model to a multimodal dataset (i.e. false negative). When a unimodal gambin distribution was simulated, the error rate (false positive) was roughly 7.0% (see Appendix S4). When a bimodal gambin distribution was simulated, the mean error rate depended on the sample size and the difference between the  $\alpha_1$  and  $\alpha_2$  values in the simulated data (Fig. 2). When the difference between  $\alpha_1$  and  $\alpha_2$  was relatively large, the error rate was very low (e.g. 0%) regardless of the number of species in the sample. In contrast, when the difference between the  $\alpha_1$  and  $\alpha_2$  values was very small, the error rate was high (e.g. 81%) regardless of the number of species. The fact that the error rate increases as the components become closer together (Fig. 2) is to be expected, as the underlying sample distribution starts to resemble a unimodal distribution. As most empirical multimodal SADs have distinct rarer and more common species modes, this is not a substantive issue. The approach can be considered conservative in that the model comparison test is slightly biased towards selecting the unimodal model over the multimodal model.

A full outline of the methodology, results and discussion for each of the simulations, along with a more detailed discussion, is provided in Appendix S4 in the Supplementary Information. All analyses were undertaken in R (version 3.4.3; R Core Team, 2017).

## CONCLUDING REMARKS

In this paper, we have derived the maximum likelihood equations for gambin models with multiple components and integrated these functions into an updated version of the ‘gambin’ R package available on CRAN. Due to the relatively simple underlying mathematics and binning procedure, the models are easy to fit and the maximum likelihood estimation procedure does not require the user to vary the starting parameter values or the optimisation algorithm employed. Hence, multimodal gambin models represent a novel, easily applied test for determining whether SADs or certain other macroecological datasets exhibit multimodality. We have also provided a number of additional functions to aid in the analysis of multimodal SADs.

As Antão et al. (2017, p. 203) state, “multimodality occurs with a prevalence that warrants its systematic consideration when assessing SAD shape and emphasizes the need for macroecological theories to include multimodality in the range of SADs they predict.” The development of multimodal gambin models provides one tool to undertake these types of analyses. Application of these new models to additional datasets will likely be revealing and will help in improving our understanding of multimodality in SADs and possibly in other macroecological data forms.

## ACKNOWLEDGEMENTS

Pedro Cardoso and three anonymous reviewers kindly provided comments on an earlier version of the manuscript. RFK was supported by grants from CNPq (Process numbers 202236/2015-3 and 308908/2016-3).



## AUTHORS' CONTRIBUTIONS

T.J.M designed the study and led the drafting with input from R.J.W. T.J.M designed and ran the analyses with the help of M.K.B and F.R. K.I.U, C.S.G and T.J.M derived the likelihood functions, and C.S.G and T.J.M built the new version of the R package. R.F.K, R.M and Y.K contributed datasets. All authors contributed to the final manuscript.

## DATA ACCESSIBILITY

gambin is freely available from CRAN (<https://CRAN.R-project.org/package=gambin>), whilst the development version is hosted on GitHub (<https://github.com/txm676/gambin>), where feature requests and bug reports can be posted. The Brazilian fly data are freely available with the gambin R package.

## REFERENCES

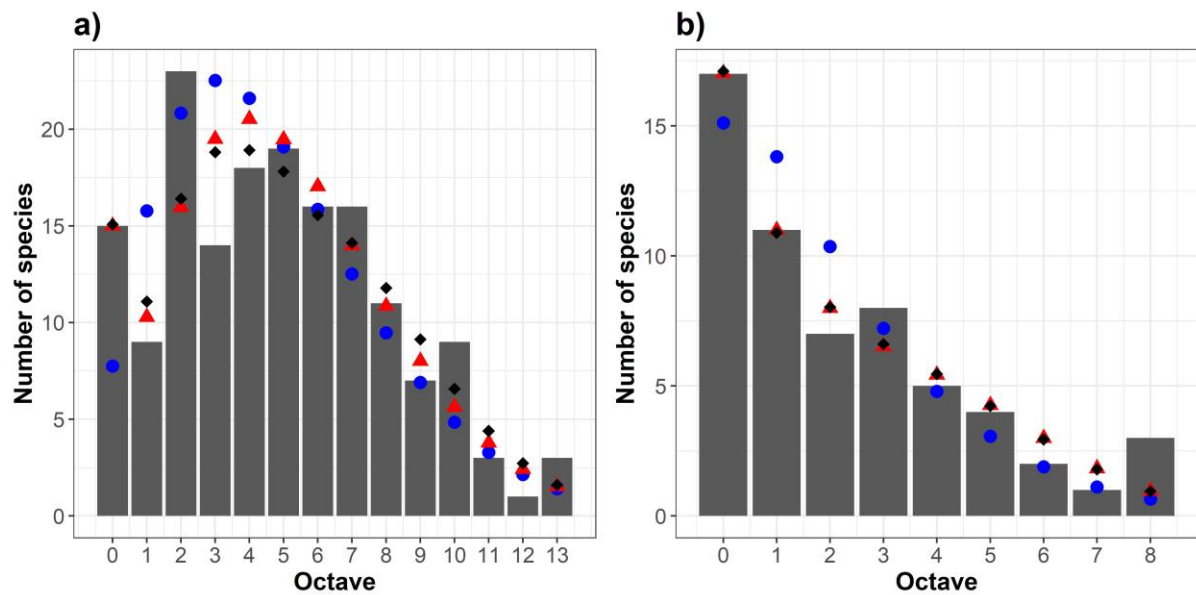
- Alonso, D., Ostling, A. & Etienne, R.S. (2008) The implicit assumption of symmetry and the species abundance distribution. *Ecology Letters*, **11**, 93-105.
- Antão, L.H., Connolly, S.R., Magurran, A.E., Soares, A. & Dornelas, M. (2017) Prevalence of multimodal species abundance distributions is linked to spatial and taxonomic breadth. *Global Ecology and Biogeography*, **26**, 203-215.
- Arellano, G., Umaña, M.N., Macía, M.J., Loza, M.I., Fuentes, A., Cala, V. & Jørgensen, P.M. (2017) The role of niche overlap, environmental heterogeneity, landscape roughness and productivity in shaping species abundance distributions along the Amazon–Andes gradient. *Global Ecology and Biogeography*, **26**, 191-202.
- Bulmer, M.G. (1974) On fitting the Poisson lognormal distribution to species-abundance data. *Biometrics*, **30**, 101-110.
- Burnham, K.P. & Anderson, D.R. (2002) *Model selection and multi-model inference: a practical information-theoretic approach*, 2nd edn. Springer, New-York.
- Dornelas, M. & Connolly, S.R. (2008) Multiple modes in a coral species abundance distribution. *Ecology Letters*, **11**, 1008-1016.
- Dornelas, M., Soykan, C.U. & Ugland, K.I. (2011) Biodiversity and disturbance. *Biological diversity: frontiers in measurement and assessment* (eds A.E. Magurran & B.J. McGill), pp. 237-251. Oxford University Press, Oxford.

- Fisher, R.A., Corbet, A.S. & Williams, C.B. (1943) The relation between the number of species and the number of individuals in a random sample of an animal population. *Journal of Animal Ecology*, **12**, 42-58.
- Gaston, K.J. (2003) *The structure and dynamics of geographic ranges*. Oxford University Press, Oxford.
- Kubota, Y., Kusumoto, B., Shiono, T. & Ulrich, W. (2018) Environmental filters shaping angiosperm trees assembly along climatic and geographical gradients. *Journal of Vegetation Science*. doi:[10.1111/jvs.12648](https://doi.org/10.1111/jvs.12648)
- Magurran, A.E. & Henderson, P.A. (2003) Explaining the excess of rare species in natural species abundance distributions. *Nature*, **422**, 714-716.
- Matthews, T.J., Borges, P.A.V., de Azevedo, E.B. & Whittaker, R.J. (2017) A biogeographical perspective on species abundance distributions: recent advances and opportunities for future research. *Journal of Biogeography*, **44**, 1705–1710.
- Matthews, T.J., Borregaard, M.K., Ugland, K.I., Borges, P.A.V., Rigal, F., Cardoso, P. & Whittaker, R.J. (2014) The gambin model provides a superior fit to species abundance distributions with a single free parameter: evidence, implementation and interpretation. *Ecography*, **37**, 1002-1011.
- Matthews, T.J. & Whittaker, R.J. (2015) On the species abundance distribution in applied ecology and biodiversity management. *Journal of Applied Ecology*, **52**, 443-454.
- McGill, B.J. (2011) Species abundance distributions. *Biological diversity: frontiers in measurement and assessment* (eds A.E. Magurran & B.J. McGill), pp. 105-122. Oxford University Press, Oxford.
- McGill, B.J., Etienne, R.S., Gray, J.S., Alonso, D., Anderson, M.J., Benecha, H.K., ... White, E.P. (2007) Species abundance distributions: moving beyond single prediction theories to integration within an ecological framework. *Ecology Letters*, **10**, 995-1015.
- R Core Team (2017) R: A language and environment for statistical computing. R foundation for statistical computing, Vienna, Austria.
- Ugland, K.I., Lambshead, P.J.D., McGill, B., Gray, J.S., O'Dea, N., Ladle, R.J. & Whittaker, R.J. (2007) Modelling dimensionality in species abundance distributions: description and evaluation of the Gambin model. *Evolutionary Ecology Research*, **9**, 313-324.
- Vergnon, R., van Nes, E.H. & Scheffer, M. (2012) Emergent neutrality leads to multimodal species abundance distributions. *Nature Communications*, **3**, 663.

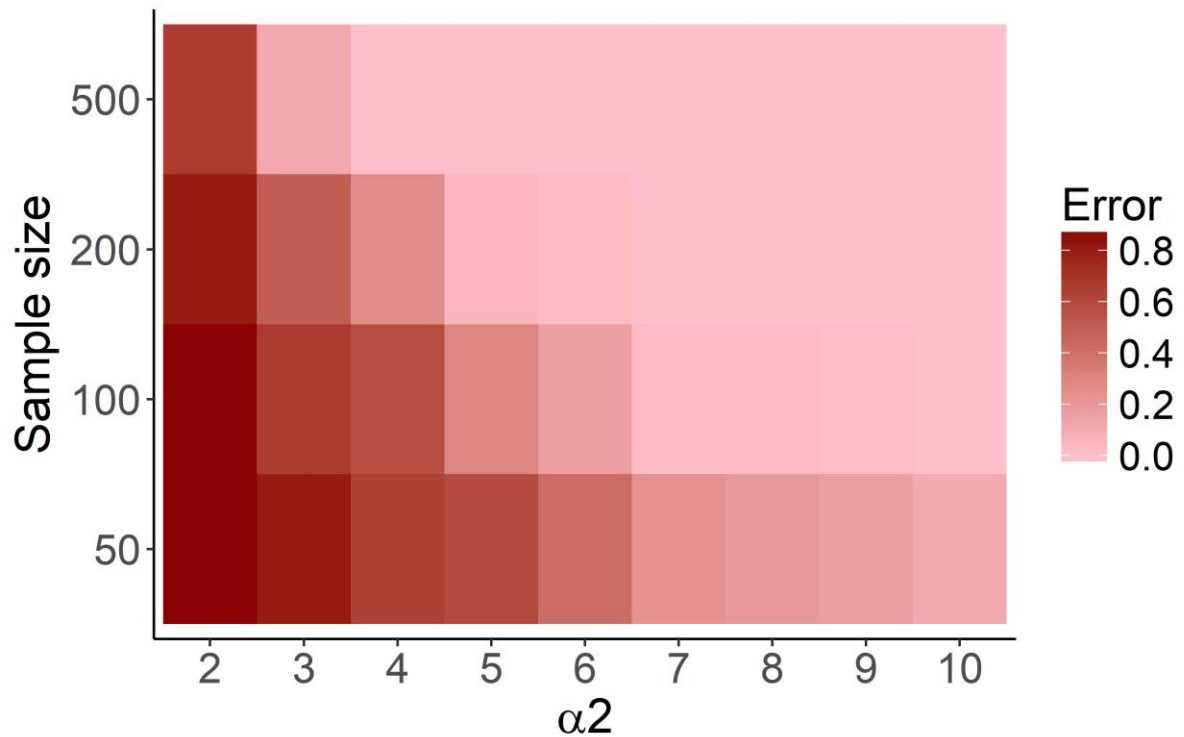
## SUPPORTING INFORMATION

Please see the online Supporting information tab for this article.

## FIGURES



**FIGURE 1** The fit of the unimodal (blue circles), bimodal (red triangles) and trimodal (black diamonds) gambin models to two horse fly species abundance distribution datasets (black bars) from Brazil. (a) horse fly data from 33 localities across Brazil (number of unique species = 164; total number of individuals = 78,755), and (b) data from one individual locality and one type of sampling (number of unique species = 58; total number of individuals = 1943; see Appendix S3). In (a) the bimodal model provides the best fit according to BIC, whilst the unimodal model provided the best to (b).



**FIGURE 2** The multimodal SAD error rate (expressed as a percentage) for an information theoretic model comparison test. For the test, a bimodal SAD was simulated, with one  $\alpha$  parameter fixed at 0.5 and the second ( $\alpha_2$ ) set to vary between 2 and 10 in units of 1. The number of species (sample size) was set to: 50, 100, 200, 500. The unimodal and bimodal gambin models were then fitted to this simulated SAD and the best model fit determined using BIC. The error rate percentage relates to the proportion of times the unimodal model provided a better fit than the bimodal model (i.e. a higher error rate percentage indicates that the unimodal model erroneously provided a better fit to the bimodal SAD).

432

## SUPPORTING INFORMATION

433 **Extension of the gambin model to multimodal species abundance distributions**

434 Thomas J. Matthews, Michael K. Borregaard, Colin S. Gillespie, Karl I. Ugland , François  
 435 Rigal, Rodrigo Ferreira Krüger, Roberta Marques, Jon P. Sadler, Paulo A.V. Borges,  
 436 Yasuhiro Kubota, Robert J. Whittaker

437

438 **Appendix S1: Derivation of the likelihood functions for multimodal gambin models**

439 As outlined in the main paper, the two-component gambin model is simply the mixture of  
 440 two gambin distributions. In order to allow for the subdivision of all of the observed objects  
 441 (*species* in the context of SADs) ( $y_{obs}$ ), a parameter ( $w_1$ ) is needed that describes the fraction  
 442 of objects belonging to the first distribution ( $w_i$  is analogous to the  $\rho$  parameter in the  
 443 multimodal PLN context). The fraction of objects belonging to the second component ( $w_2$ ) is  
 444  $1 - w_1$ . Thus, the expected number of observed objects is split into two components,  
 445 consisting of  $w_1 * y_{obs}$  and  $w_2 * y_{obs}$  objects, respectively. Thus,  $y_{obs} = (w_1 * y_{obs}) + (w_2 * y_{obs})$ .  
 446 With no extra information, we may therefore assume that the number of objects in the  $k$ -th  
 447 interval ( $k = 1, 2, \dots, i$ ) are  $w_1 * y_k$  and  $w_2 * y_k$ .

448 At present, the gambin distribution can only be fitted to data binned into octaves, and the  
 449 binning process used follows that used in previous gambin papers, whereby octave 0 contains  
 450 the number of species with 1 individual, octave 1 the number of species with 2 or 3  
 451 individuals, and so on.

452 The likelihood function for the two-component gambin distribution can then be derived as  
 453 follows. Following Ugland et al. (2007), for a standard one-component gambin model, we fix  
 454 the scale parameter of a gamma distribution to 1, and the 99% point for  $\alpha$  is defined as the  
 455 number  $c_1(\alpha)$  that covers an area of 0.99:

457

$$456 \int_0^{c_1(\alpha)} \frac{1}{\Gamma(\alpha)} x^{\alpha-1} e^{-x} dx = 0.99$$

458

459 The interval is subdivided between 0 and  $c_1(\alpha)$  into 100 equal sub-intervals:  $(k / 100) * c_1(\alpha)$   
 460 for  $k = 1, \dots, 100$ . The frequencies of the relative successes will be arranged according to:

$$461 P\left(\frac{k-1}{100} < \text{Rel. success} < \frac{k}{100}\right) = \left[G_{\alpha,1}\left(\frac{k}{100}c_1(\alpha)\right) - G_{\alpha,1}\left(\frac{k-1}{100}c_1(\alpha)\right)\right] = \delta_{k,\alpha} \quad (1)$$

462 where  $G$  is the gamma distribution function. The probability of observing a species in octave  
 463  $s$  is:

$$464 Pr(S = s | \alpha, n_{oct}) = \sum_{j=1}^{100} \delta_{j,\alpha} \binom{N}{s} p_j^s (1 - p_j)^{N-s} \quad (2)$$

465 where  $n_{\text{oct}}$  = the number of octaves,  $N = n_{\text{oct}} - 1$ , and  $p_j = j / 100$ .

466

467 Hence the two-component gambin model is given by:

$$\begin{aligned} 468 \quad Pr(S = s | \alpha_1, \alpha_2, n_{\text{oct},1}, n_{\text{oct},2}, w_1) &= \sum_{j=1}^{100} w_1 \delta_{j,\alpha_1} \binom{N_1}{s} p_j^s (1 - p_j)^{N_1-s} + \\ 469 \quad w_2 \delta_{j,\alpha_2} \binom{N_2}{s} p_j^s (1 - p_j)^{N_2-s} \quad (3) \end{aligned}$$

470 where  $w_i$  is the fraction of species in  $i$ -th component,  $\alpha_i$  is the  $\alpha$  parameter of the  $i$ -th  
471 component,  $n_{\text{oct},i}$  is the max octave of the  $i$ -th component,  $N_i = n_{\text{oct},i} - 1$ , and  $p_j = j / 100$ .

472 It is relatively straightforward to extend the above approach for fitting the two-component  
473 gambin model by maximum likelihood, to fitting gambin models with  $g$  components (where  
474 components correspond to the number of modes):

475

$$476 \quad Pr(S = s | \theta) = \sum_{j=1}^{100} \sum_{k=1}^K w_k \delta_{j,\alpha_k} \binom{N_k}{s} p_j^s (1 - p_j)^{N_k-s} \quad (4)$$

477 where  $w_K = 1 - \sum_{k=1}^{K-1} w_k$ ,  $N_k = n_{\text{oct},k} - 1$ ,  $p_j = \frac{j}{100}$ , and

478  $\theta = (\alpha_1, \dots, \alpha_K, n_{\text{oct},1}, \dots, n_{\text{oct},K}, w_1, \dots, w_{K-1})$ .

479

480 If we observe data  $s_1, \dots, s_n$  the associated likelihood function is:

$$481 \quad L(\theta; s_1, \dots, s_n) = \prod_{i=1}^n Pr(S = s_i | \theta) \quad (5)$$

482

483

484

485

486

487

488

489

490

491

492

## 493 **Appendix S2: Overview of the Brazilian Horsefly Dataset**

494 The empirical dataset used in the main paper comprises abundance records of horse flies  
495 (Diptera, Tabanidae) from a variety of locations in South America, Central America and  
496 Mexico, with a particular focus on Brazil. Data were sourced from the literature (e.g. Barbosa  
497 et al., 2005) and the subset of the dataset focused on Brazil comprised data from 33 localities  
498 across the country. The data come from a wide variety of sources incorporating different  
499 sampling intensities, sampling methods (e.g. malaise traps, canopy traps, netting, light traps,  
500 baited traps) and sampling extents. Where the same species was recorded in multiple  
501 localities, we summed the abundance values across these records for use in constructing the  
502 SAD (number of unique species = 164; total number of individuals = 78,755). We then took a  
503 subset of the dataset relating to one individual locality and one type of sampling (Centro de  
504 Instrução de Guerra na Selva, close to the city of Manaus, 02°45'33"S, 59°51'03"W; canopy  
505 sampling only; number of unique species = 58; total number of individuals = 1943; Ferreira-  
506 Kepler et al. 2010). Both sets of data are available in the gambin R package, using the  
507 command: data(fly).

508

509 Barbosa, M.G.V., Henriques, A.L., Rafael, J.A. & da Fonseca, C.R.V. (2005) Diversidade e  
510 similaridade entre habitats em relação às espécies de Tabanidae (Insecta: Diptera) de  
511 uma floresta tropical de terra firme (Reserva Adolpho Ducke) na Amazônia Central,  
512 Brasil. *Amazoniana*, **18**, 251-266.

513

514 Ferreira-Keppler, R.L., Rafael, J.A. & Guerrero, J.C.H. (2010) Sazonalidade e uso de  
515 ambientes por espécies de Tabanidae (Diptera) na Amazônia Central, Brasil.  
516 *Neotropical Entomology*, **9**, 645-654.

517

518

519

520

521

522

523

524

525

526

527

## Appendix S3: Additional empirical examples of multimodal gambin fits

### *Model results of gambin fits to 275 woody plant SAD datasets*

The results are available as a separate excel file ('Kubota\_BIC\_values').

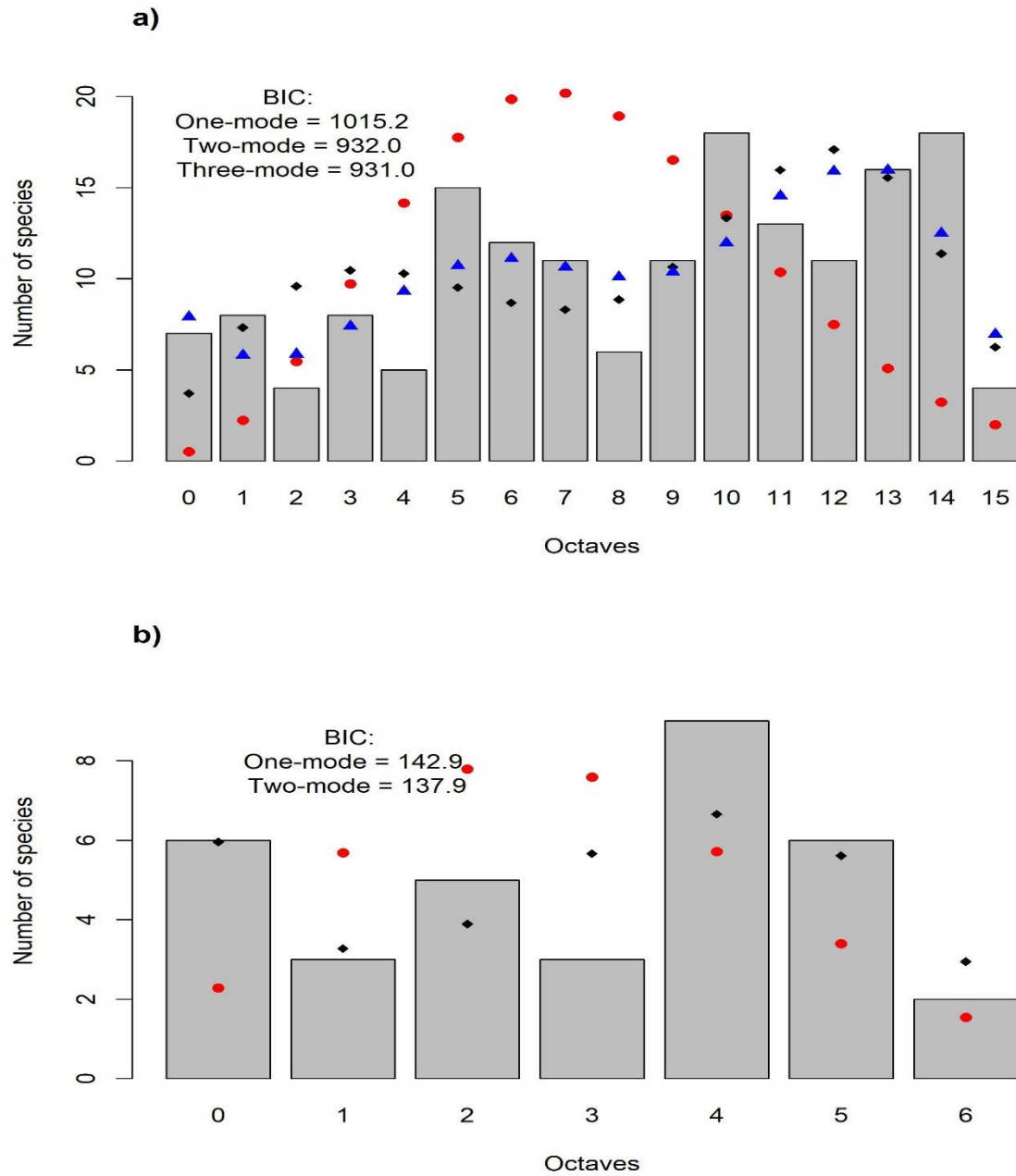
### *Other macroecological distributions*

To illustrate potential applications of multimodal gambin models in regards to summarising macroecological distributions other than species abundance distributions we fitted multimodal gambin models to a (a) species range-size distribution dataset, and (b) frequency distribution of species occupancy. For (a), we fitted one, two and three-component gambin models to the global range size distribution of 167 marine mammal species, including seals, whales, dolphins and dugongs. The data were downloaded from IUCN (2016). For (b), we fitted one and two-component gambin models to data of the occupancy of intestinal helminths in three species of grebe (total of 84 individuals) sampled in Alberta, Canada (data were pooled from Tables 2-4 in Stock & Holmes, 1988; see also Vandermeer & Goldberg, 2003, Chapter 5). In both cases, models were compared using BIC. The model fits and BIC values are presented in Figure S1 below. For the marine mammal range size distribution, the three-component model had the lowest BIC value, and for the intestinal helminths occupancy distribution the two-component model provided the best fit. The bimodal fit in (b) is in accordance with the core-satellite hypothesis of Hanski (1982; see also Vandermeer & Goldberg, 2003, Chapter 5).

### *References*

- Hanski, I. (1982) Dynamics of regional distribution: the core and satellite species hypothesis. *Oikos*, **38**, 210-221.
- IUCN (2016) The IUCN Red List of Threatened Species. Version 2016-1. <http://www.iucnredlist.org>. Downloaded on 21/05/2018.
- Stock, T.M. & Holmes, J.C. (1988) Functional relationships and microhabitat distributions of enteric helminths of grebes (Podicipedidae): the evidence for interactive communities. *The Journal of Parasitology*, **74**, 214-227.
- Vandermeer, J.H. & Goldberg, D.E. (2003) *Population ecology: first principles*. Princeton University Press, Princeton.





**FIGURE S1** The fits of different gambin models to (a) the global range size distribution of 167 marine mammal species, and (b) the occupancy distribution of intestinal helminths in three species of grebe (total of 84 individuals) sampled in Alberta, Canada. In (a) the one (red circles), two (black diamonds) and three-component (blue triangle) gambin models have been fitted, and in (b) the one (red circles) and two-component (black diamonds) gambin models have been fitted. The BIC values of the model fits are provided in the plots.

## Appendix S4: Simulation analyses – methods, results, discussion and supplementary figures.

To test the error rates of our new models, and any sensitivity to the number of species and individuals (Locey & White, 2013), we ran three separate simulations.

### *Sensitivity of model fits to the number of species*

To test how the performance of the bimodal gambin model varied with the number of species in the sample, we first sampled  $n$  species from a bimodal gambin distribution (that represented a true ecological community in this simulation) with set parameter values  $\alpha_1$  and  $\alpha_2$ , using the ‘*rgambin*’ function. The bimodal model was fitted to this  $i$ th sample and the  $i$ th sample parameter values ( $\alpha_{i1}$  and  $\alpha_{i2}$ ) stored. This procedure was then repeated with a different value of  $n$ , with  $n$  increasing from 20 to 1000 species in increments of 20 (i.e. up to  $i = 50$ ). As sampling random values from a gambin distribution with set parameter values is stochastic, this incremental sampling and model fitting process was then repeated 100 times with the same  $\alpha_1$  and  $\alpha_2$  values. We took the median sample parameter estimates of the 100 iterations (i.e. the median of the 100 parameter estimates for a given  $n$ ) and compared these to the  $\alpha_1$  and  $\alpha_2$  values of the underlying distribution. It was necessary to use the median rather than the mean as when  $n$  was very low the model fit would occasionally not converge and unreasonably high  $\alpha$  values would be generated. For practical applications, this issue of non-convergence may be avoided by simply fixing the lower and upper bounds of the  $\alpha$  parameters; for example, between 0.1 and 100. The  $\alpha_1$  and  $\alpha_2$  values were set to 1 and 5, the max octaves were set to 5 and 13, and the weights were set to 0.5. We experimented with different community  $\alpha_1$  and  $\alpha_2$  values, but the results were qualitatively similar and so we only present the results using  $\alpha_1$  and  $\alpha_2$  values of 1 and 5, below.

When the simulation was run using the bimodal gambin model and  $\alpha_1$  and  $\alpha_2$  values of 1 and 5, the estimation of the  $\alpha_1$  parameter was relatively insensitive to the number of species in the sample (Fig. S2). That is, the sampled data provided accurate estimations of  $\alpha_1$  regardless of the number of species. The estimation of  $\alpha_2$  was more sensitive when the number of species in the sample was very low (Fig. S2). In particular, when the number of species in the sample was below 50, the model tended to overestimate the value of  $\alpha_2$ . Using different community  $\alpha_1$  and  $\alpha_2$  values resulted in the same pattern emerging.

### *Sensitivity of model fits to the total number of individuals*

In order to test the sensitivity of the bimodal gambin model to the total number of individuals, we followed the approach of McGill (2011) and used two well specified SAD datasets from the literature: (1) the 2005 BCI tree dataset (number of species = 229; number of individuals = 20, 852), comprising all trees with DBH greater than 10 cm from a 50 ha forest plot in Panama (Hubbell *et al.*, 2005), and (2) a coral reef dataset (number of species = 154; number of individuals = 44, 255) from Australia (Dornelas & Connolly, 2008). First, for

each community SAD we fitted and examined the bimodal gambin model to determine whether it provided a good fit to both datasets. We then employed an iterative subsampling procedure whereby we first subsampled  $x_1$  individuals from each community SAD, fitted the bimodal model and recorded the two  $\alpha$  parameter ( $\alpha_1$  and  $\alpha_2$ ) estimates. We then increased  $x_1$  by an increment  $y$  and subsampled  $x_2$  individuals from each community SAD, and so on until  $x_i$  was equal to the number of individuals in the community SAD. We set  $x_1$  to 100 for both datasets, and  $y$  to 400 for the BCI dataset and 800 for the coral reef dataset, given their differences in terms of number of individuals. We then plotted the  $\alpha_1$  and  $\alpha_2$  values from these subsamples against  $x_i$  to determine the influence of the number of sampled individuals on the parameter estimates. As sampling from a distribution is a random process, we re-ran the subsampling 100 times for each value of  $x_i$  and took the median of the two  $\alpha$  parameter estimates (i.e. the median of 100  $\alpha_1$  values and 100  $\alpha_2$  values for each  $x_i$ ). For the purposes of this analysis, for the coral reef dataset we edited the maximum likelihood optimisation procedure so that both of the two compound distributions in the bimodal model had the same max octave (i.e. the max octave of the empirical distribution). This was because the maximum likelihood value without this change corresponded to a model in which one of the compound distributions only covered the first two octaves; and thus did not provide a useful example for our tests. Fixing the max octave for both compound distributions resulted in a model with a slightly lower likelihood, but provided a better test case for the purposes of this simulation.

The two  $\alpha$  parameter estimates from the bimodal model were 0.50 and 6.41, and 1.07 and 11.08, for the BCI and coral datasets, respectively. Plotting the fitted values indicated that the bimodal model provided relatively good fits to both datasets (see Fig. S3a, b). Results from the subsampling analysis were similar for both the BCI dataset and the coral dataset (Fig. S3c, d). In regards to the  $\alpha$  parameter of the first component in the bimodal model (i.e.  $\alpha_1$ ), the parameter estimate was relatively accurate regardless of the sample size; however, at very low sample sizes the model tended to overestimate the  $\alpha_1$  parameter when the BCI dataset was used (Fig. S3c). In regards to the  $\alpha_2$  parameter, in both cases the parameter estimate was strongly influenced by the number of individuals in the sample, such that at low sample sizes the model-estimated  $\alpha_2$  parameter value varied quite considerably from the true value (Fig. S2c, d). As sample size increased, the  $\alpha_2$  parameter estimate appeared to converge on the values for the full dataset; however, this convergence was not asymptotic.

### ***Determining the error rate***

When working with multimodal SAD models, and ecological models more generally, it is useful to know the error rate of the model fits, i.e. the proportion of times that a multimodal model provides a better fit than a unimodal model to a unimodal dataset (i.e. false positive) and the proportion of times a unimodal model provides a better fit than a multimodal model to a multimodal dataset (i.e. false negative) (see Antão *et al.*, 2017), and under what conditions the error rate varies. To this end, we simulated both 100 unimodal and 100

bimodal gambin distributions with given  $\alpha$  parameters and number of species (i.e. it is known *a priori* how many components each distribution has). We then fitted both the unimodal and bimodal gambin models to each distribution and calculated the BIC values (see Burnham & Anderson, 2002) of the model fits. It was then possible to determine the number of times a bimodal gambin model provided a better fit than a unimodal model, based on BIC, to a unimodal sample (false positive), and *vice versa*. The total error rate was expressed as a percentage.

For the unimodal simulation, the number of species, the  $\alpha$  value and the max octave of the simulated datasets were set to 100, 1 and 10, respectively. The weights for the simulated bimodal distribution were set to 0.3 and 0.7 (for the first and second compound distributions, respectively), and the max octaves were set to 7 and 10 (again for the first and second compound distributions, respectively). These values were chosen as experimental analysis revealed they produced realistic looking ecological SADs. We experimented with different starting  $\alpha$  values and number of species to determine whether/how the error rate changes when the modes in the bimodal distribution become closer together. To achieve this, in each simulation the  $\alpha_1$  value (i.e. the component distribution corresponding to the relatively rarer species) was set to 0.5, whilst the  $\alpha_2$  value was iteratively changed across simulations, from 2 to 10, in units of 1. The number of species were set to 50, 100, 200 and 500.

When a unimodal gambin distribution was simulated, the error rate (false positive) was 7.0%. When a bimodal gambin distribution was simulated (number of species varied between 50 and 500;  $\alpha_1$  kept constant at 0.5;  $\alpha_2$  varied from 2 to 10), the mean error rate (false negative) depended on the sample size and the difference between the  $\alpha_1$  and  $\alpha_2$  values (Fig. 2 in the main paper). When the difference between  $\alpha_1$  and  $\alpha_2$  was relatively large, the error rate (i.e. the proportion of times BIC selected the unimodal gambin model as the best model) was very low (e.g. 0%) regardless of the sample characteristics. In contrast, when the difference between the  $\alpha_1$  and  $\alpha_2$  values was very small, the error rate was high (e.g. 81%) regardless. In between these extremes, the error rate at a given level of difference between  $\alpha_1$  and  $\alpha_2$  depended to some extent on the number of species (Fig. 2 in the main paper).

### ***Discussion of the Simulation Analyses Results***

The results of our simulations involving varying the number of species in a sample indicated that the  $\alpha$  parameter estimates of the bimodal gambin model were relatively insensitive to the number of species in the sample. There was some bias (see also McGill, 2011 for a discussion of this issue in SAD analyses more generally), such that at low sample species richness the  $\alpha_2$  parameter tended to be overestimated. However, it was only when species richness was very low ( $< 30$ ) that the sample  $\alpha_2$  value diverged substantially from the  $\alpha_2$  value. The reason for this observation is likely due to the fact that reducing richness in the simulated samples results in a concomitant reduction in the number of individuals in the sample (discussed further below).

In contrast to the number of species in the sample, it is apparent that to get truly accurate parameter estimates (accurate in the sense that they are close to the parameter values of the population) for the bimodal model fits, the number of individuals in the sample needs to be quite large (see Fig. S3 and Fig. S4 below). This has been reported previously for other SAD models. For example, McGill (2011, p. 114) found, using a similar subsampling method, that “most metrics only began to come close to their true values (even to within  $\pm 50\%$ ) with at least 1000 individuals sampled and in many cases only with 10,000 individuals sampled.” For example, accurate determination of the  $\alpha$  parameter from the unimodal gambin model was determined to require sample sizes of 1000 individuals (McGill, 2011). Furthermore, McGill’s analysis was focused primarily on unimodal distributions, and arguably the problem can be seen to be even more acute for compound distributions, due to the generally larger number of parameters and more complex optimisation algorithms needed to find the maximum likelihood estimates of the parameters.

In regards to our simulations using the bimodal gambin model, it is the  $\alpha_2$  parameter that is most sensitive to sample size, and although  $\alpha_2$  increases with increasing sample size, albeit with scatter at very small sample sizes, it does not perfectly asymptotically converge on the true value. This issue appears to be largely caused by sampling problems. Primarily, when smaller and smaller samples are taken from a community SAD that is distinctly multimodal, the multiple modes become more difficult to detect. Thus, this is a problem for all multimodal SAD models, although it is potentially more of an issue with gambin due to the prior binning of the data. Figure S4 illustrates this sampling effect. In Fig.S4d the fit of the bimodal gambin model to the BCI dataset is shown. The fit of the bimodal model to samples of 100, 1000 and 10,000 individuals from the BCI dataset (Fig.S4a, b, c) are also shown. It can be seen that, as sample size decreases, the second mode (i.e. the mode representing the relatively more common species) shifts to the left, whilst the first mode (i.e. the mode representing the singleton class) remains static (Fig. S4). This fits with Preston’s (1948) concept of the veil line; it is only by increasing the size of the sample that the full empirical distribution is revealed. This explains why the  $\alpha_1$  parameter is relatively insensitive to variations in sample size, whilst the shifting of the second mode to the left of the distribution explains why  $\alpha_2$  tends to decrease with sample size. The shifting of the second mode with variation in sample size is not a ‘smooth’ continuous process, which is why  $\alpha_2$  does not converge towards the true value in a smooth, asymptotic fashion. The reason for this likely relates to the changing empirical max octave value of the sample. Starting with a small sample size (i.e. a low proportion of the number of individuals in the community), as the number of sampled individuals increases at some point a new max octave is added (i.e. there are species now in the sample with higher abundance than in samples of smaller sizes; e.g. compare the different empirical max octaves in Fig. S4a-d). The addition of a new max octave acts to stretch out the gambin distribution, which in turn influences the  $\alpha$  values.

The sensitivity of the  $\alpha$  parameter to sample size was found in previous work focused on the unimodal gambin model (Matthews *et al.*, 2014). However, in contrast to the unimodal model, in regards to the bimodal model we do not think that this sensitivity is necessarily a substantive issue, depending on the aim of the study. With the unimodal gambin model, the  $\alpha$

value can be used as a type of diversity metric to compare SAD shape across communities (e.g. Arellano *et al.*, 2017). Thus, we have previously advised that subsampling should be used to ensure constant sample size in comparative analyses using the unimodal model (Matthews *et al.*, 2014). However, for multimodal gambin models the meaning of the  $\alpha$  values is not as clear: partly due to the fact that the max octave of the first distribution (in the context of the bimodal model) is allowed to vary and thus will vary between different samples. Rather, the benefit of multimodal gambin models is to provide a simple, quick and easy to use test for determining whether empirical SADs are multimodal; the  $\alpha$  parameter values can then be used to simply provide a rough idea of the shape of the component distributions in certain cases.

Furthermore, according to our simulations, unless sample size is very low,  $\alpha_2$  is still relatively large compared to  $\alpha_1$  and thus indicates to the user the presence of two components in the overall distribution. It is also worth noting that, as gambin is a statistical distribution rather than an ecological model, obtaining exact parameter estimates, particularly for multimodal models, is arguably less important than for other, ecological models (e.g. the migration parameter,  $m$ , in various neutral models; e.g. Hubbell, 2001; Chust *et al.*, 2013).

#### *Acknowledgments and references for Appendix S4*

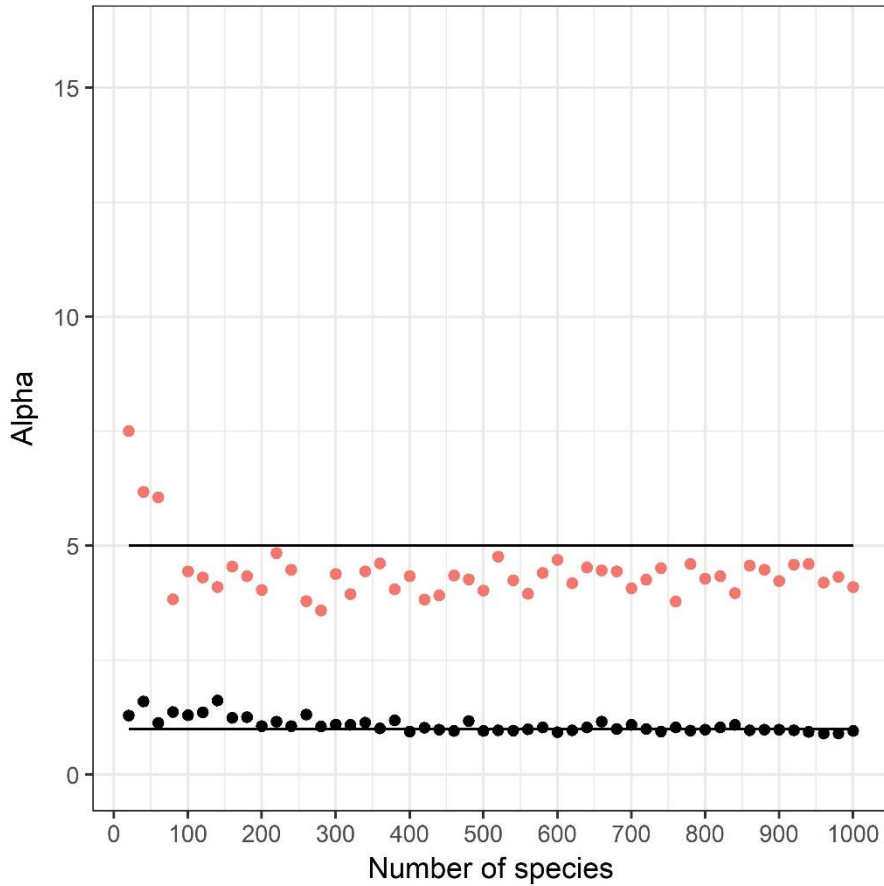
The BCI forest dynamics research project was made possible by National Science Foundation grants to Stephen P. Hubbell: DEB-0640386, DEB-0425651, DEB-0346488, DEB-0129874, DEB-00753102, DEB-9909347, DEB-9615226, DEB-9615226, DEB-9405933, DEB-9221033, DEB-9100058, DEB-8906869, DEB-8605042, DEB-8206992, DEB-7922197, support from the Center for Tropical Forest Science, the Smithsonian Tropical Research Institute, the John D. and Catherine T. MacArthur Foundation, the Mellon Foundation, the Small World Institute Fund, and numerous private individuals, and through the hard work of over 100 people from 10 countries over the past two decades. The plot project is part of the Center for Tropical Forest Science, a global network of large-scale demographic tree plots.

Antão, L.H., Connolly, S.R., Magurran, A.E., Soares, A. & Dornelas, M. (2017) Prevalence of multimodal species abundance distributions is linked to spatial and taxonomic breadth. *Global Ecology and Biogeography*, **26**, 203-215.

Arellano, G., Umaña, M.N., Macía, M.J., Loza, M.I., Fuentes, A., Cala, V. & Jørgensen, P.M. (2017) The role of niche overlap, environmental heterogeneity, landscape roughness and productivity in shaping species abundance distributions along the Amazon–Andes gradient. *Global Ecology and Biogeography*, **26**, 191-202.

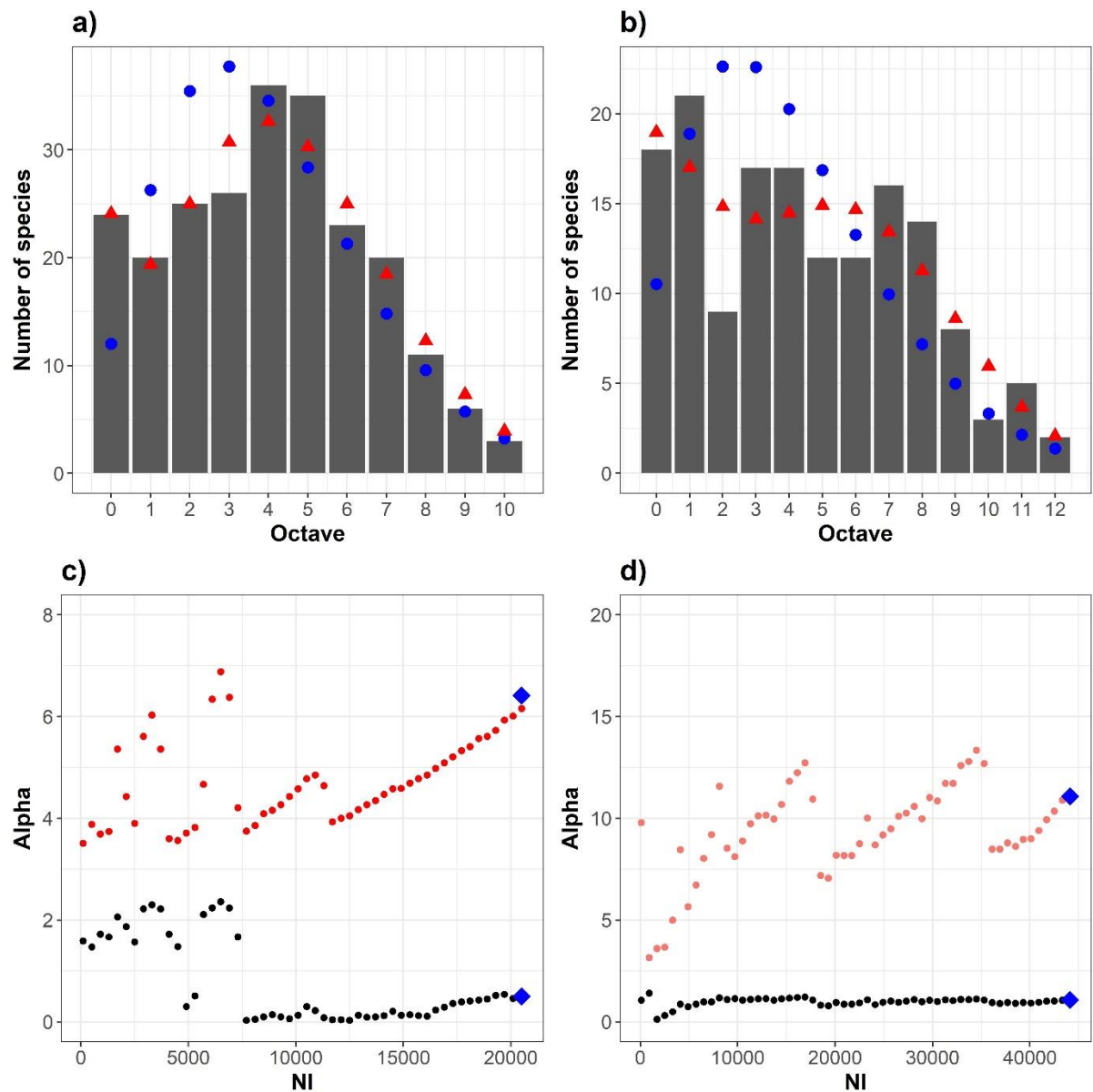
Burnham, K.P. & Anderson, D.R. (2002) *Model selection and multi-model inference: a practical information-theoretic approach*, 2nd edn. Springer, New-York.

- Chust, G., Irigoien, X., Chave, J. & Harris, R.P. (2013) Latitudinal phytoplankton distribution and the neutral theory of biodiversity. *Global Ecology and Biogeography*, **22**, 531-543.
- Dornelas, M. & Connolly, S.R. (2008) Multiple modes in a coral species abundance distribution. *Ecology Letters*, **11**, 1008-1016.
- Hubbell, S.P. (2001) *The unified neutral theory of biodiversity and biogeography*. Princeton University Press, Princeton.
- Hubbell, S.P., Condit, R. & Foster, R.B. (2005) *Barro Colorado Forest census plot data*. Available at: <https://ctfs.arnarb.harvard.edu/webatlas/datasets/bci>. (accessed April 3rd 2013).
- Locey, K.J. & White, E.P. (2013) How species richness and total abundance constrain the distribution of abundance. *Ecology Letters*, **16**, 1177-1185.
- Matthews, T.J., Borregaard, M.K., Ugland, K.I., Borges, P.A.V., Rigal, F., Cardoso, P. & Whittaker, R.J. (2014) The gambin model provides a superior fit to species abundance distributions with a single free parameter: evidence, implementation and interpretation. *Ecography*, **37**, 1002-1011.
- McGill, B.J. (2011) Species abundance distributions. *Biological diversity: frontiers in measurement and assessment* (ed. by A.E. Magurran and B.J. McGill), pp. 105-122. Oxford University Press, Oxford.
- Preston, F.W. (1948) The commonness, and rarity, of species. *Ecology*, **29**, 254-283.
- R Core Team (2017) *R: A language and environment for statistical computing*. Version 3.4.0. R foundation for statistical computing, Vienna.

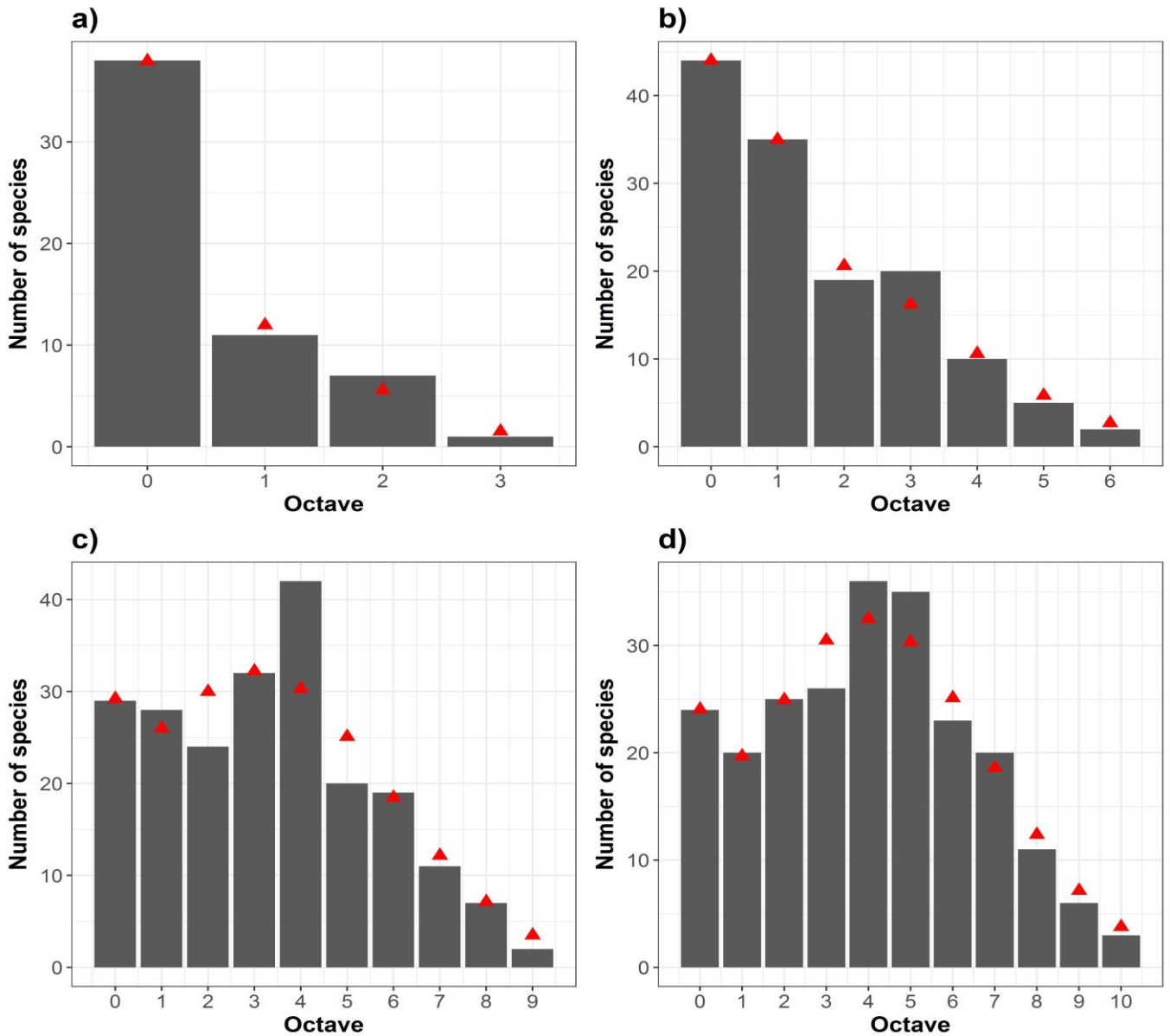


**FIGURE S2** Variation in the two  $\alpha$  parameter estimates ( $\alpha_1$  = black circles;  $\alpha_2$  = red circles) of a bimodal gambin model with the number of species in the sample. Samples were taken from a bimodal distribution with  $\alpha_1$  and  $\alpha_2$  parameters set to 1 and 5 (solid black lines), respectively. For each number of species value, the subsampling was repeated 100 times and the median  $\alpha_1$  and  $\alpha_2$  values taken.





**FIGURE S3** The fits of the unimodal (blue circles) and bimodal (red triangles) gambin models to two empirical SADs: (a) the 2005 BCI tree dataset (number of species = 229; number of individuals = 20,852) from Panama, and (b) a coral reef dataset (number of species = 154; number of individuals = 44,255) from Australia. Subsamples of these datasets were created by subsampling a varying number of individuals (NI) in each case. The bimodal gambin model was fitted to each subsample and the average  $\alpha_1$  (black circles) and  $\alpha_2$  (red circles) parameters (average of 100 iterations for each NI value) were stored. These values are plotted against NI for the BCI data (c) and the coral data (d). The blue diamonds in (c) and (d) are the true parameter values, i.e. the parameters of the model fits to (a) and (b). The bimodal model in (b) was fitted using the standard function in the gambin R package that allows the max octave of the first component distribution to vary (i.e. it does not have to equal to max octave of the empirical distribution). However, to enable a better test of the effect of the NI on the  $\alpha$  estimates (d), it was necessary to use a different function that fixed the max octave of both component distributions.



**FIGURE S4** The fit of the bimodal (red triangles) gambin model to the SAD (d) of the 2005 BCI tree dataset (number of individuals = 20,852), and to three samples from this dataset of varying size: (a) 100 individuals, (b) 1000 individuals, and (c) 10,000 individuals. The two  $\alpha$  values for each model fit are: (a) 0.46 and 5.90, (b) 2.31 and 3.43, (c) 0.57 and 5.42, and (d) 0.50 and 6.41. The  $P$ -value of the  $X^2$  goodness of fit test was non-significant ( $> 0.55$  in each case) for all model fits.