Received Date : 24-Oct-2013

Accepted Date : 06-Mar-2014

Article type : Research Article

Editor : Jana McPherson

Title: Understanding co-occurrence by modelling species simultaneously with a Joint Species Distribution Model (JSDM)

Running Title: Joint species distribution models

Authors: Laura J. Pollock[1,3*], Reid Tingley[2,3*], William K. Morris[1,2,3], Nick Golding[4], Robert B. O'Hara[5], Kirsten M. Parris[1,2,3], Peter A. Vesk[1,2,3] and Michael A. McCarthy[1,2,3]

[1]National Environmental Research Program (NERP) Environmental Decisions Hub

[2]ARC Centre of Excellence for Environmental Decisions (CEED)

[3]School of Botany, University of Melbourne, Victoria, Australia

[4]Spatial Ecology and Epidemiology Group, Department of Zoology, University of Oxford, Oxford, UK

[5]Biodiversity and Climate Research Centre (BiK-F), Senckenberganlage 25, 60325 Frankfurt am Main, Germany

*Both authors contributed equally to this work

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1111/2041-210X.12180

Corresponding Author: M. McCarthy, School of Botany, University of Melbourne, Victoria 3010, Australia, mamcca@unimelb.edu.au, +61 3 8344 6856

Abstract

1. A primary goal of ecology is to understand the fundamental processes underlying the geographic distributions of species. Two major strands of ecology—habitat modelling and community ecology—approach this problem differently. Habitat modellers often use species distribution models (SDMs) to quantify the relationship between species' and their environments without considering potential biotic interactions. Community ecologists, on the other hand, tend to focus on biotic interactions and, in observational studies, use co-occurrence patterns to identify ecological processes. Here, we describe a joint species distribution model (JSDM) that integrates these distinct observational approaches by incorporating species co-occurrence data into a SDM.
2. JSDMs estimate distributions of multiple species simultaneously, and allow decomposition of species co-occurrence patterns into components describing shared environmental responses and residual patterns of co-occurrence. We provide a general description of the model, a tutorial, and code for fitting the model in R. We demonstrate this modelling approach using two case studies: frogs and eucalypt trees in Victoria, Australia.
3. Overall, shared environmental correlations were stronger than residual correlations for both frogs and eucalypts, but there were cases of strong residual correlation. Frog species generally had positive residual correlations, possibly due to the fact these species occurred in similar habitats that were not fully described by the environmental variables included in the JSDM. Eucalypt species that interbreed had similar environmental responses, but had negative residual co-occurrence. One explanation is that interbreeding species may not form stable assemblages despite having similar environmental affinities.

4. Environmental and residual correlations estimated from JSDMs can help indicate whether co-occurrence is driven by shared environmental responses or other ecological or evolutionary process (e.g. biotic interactions), or if important predictor variables are missing. JSDMs take into account the fact that distributions of species might be related to each other, and thus, overcome a major limitation of modelling species distributions independently.

Key-words: amphibians, biotic interactions, community assembly, correlated residuals, *Eucalyptus*, frogs, species covariance

Introduction

The geographic distribution of a species is influenced by its environmental tolerances, as well as by interactions with other species (Hutchinson 1957), but decomposing the roles of abiotic and biotic factors on species' distributions is far from routine. Species distribution models (SDMs) that correlate the occurrence or abundance of a species with abiotic variables (e.g. climate, topography) are typically used to investigate species-environment relationships (Austin 2002). However, most SDMs only implicitly consider interactions between species (Dormann *et al.* 2012), despite the potentially important influence of biotic interactions on species' ranges (Davis *et al.* 1998; Wisz *et al.* 2013).

On the other hand, community ecology studies that tackle questions of co-occurrence tend to focus on interactions between species (e.g. trophic dynamics, facilitation, or competition). The environment is often inferred to be important if species within local communities are functionally similar (relative to null or randomized communities) (e.g. Webb *et al.* 2002). In these randomizations, co-occurrence is often represented by an index (Hardy 2008) that does not account for the amount of co-occurrence that can be attributed to shared environmental responses among species. However, studies are beginning to link the fields of community assembly and species distribution modelling (for a review see Kissling *et al.* 2012). For

example, Helmus et al. (2007) used the residuals from an SDM to calculate a co-occurrence index, thereby considering the effect of environmental variables on co-occurrence estimates.

Likewise, studies that use SDMs are beginning to consider species interactions by restricting the predicted distribution of one species to that of another (Schweiger *et al.* 2012) or by adding the occurrence or abundance of other species as predictors alongside abiotic variables (e.g. Leathwick & Austin 2001; Leathwick 2002; Meier *et al.* 2010; Pellissier *et al.* 2010). The addition of biotic interaction terms has generally improved the predictive performance of SDMs (Araujo & Luoto 2007; Heikkinen *et al.* 2007), and in some cases, biotic predictors have outperformed abiotic variables (Meier et al. 2010). However, this approach only models unidirectional interactions between species, and confounds the influence of species interactions and environmental covariates (Kissling *et al.* 2012).

Similarities in environmental responses of species can be accommodated in multispecies SDMs (Ovaskainen & Soininen 2011; Pollock, Morris & Vesk 2012), and such responses to environmental gradients can be modelled as a function of species traits (Pollock, Morris & Vesk 2012). However, not all features that influence co-occurrence, particularly biotic interactions, will be captured by environmental variables. In this case, residual patterns of co-occurrence will exist. For example, two species might have a 0.5 probability of occurrence at a site, in which case each of the four combinations of co-occurrence (both species present, both absent, and one or the other present) would be equally likely if the species occurred independently. However, if the species were perfectly positively associated (taken as one extreme for illustrative purposes), then they would occur together at 50% of sites, and both would be absent from 50% of sites. Alternatively, with perfect negative co-occurrence, one species would be present at 50% of sites while the other species would only be present at the other 50% of sites, and the species would never occur together. Such residual patterns of co-occurrence can be thought of as correlations in the random (Bernoulli-distributed) occurrence of species.

Hierarchical generalized linear models provide a flexible way to include multiple species in a single SDM and incorporate uncertainties that are common in species distribution data (Gelfand *et al.* 2003). Multispecies models result in more precise estimates of model parameters for rare species because parameters can "borrow strength" from those of common species (Ovaskainen & Soininen 2011; Pollock, Morris & Vesk 2012). Despite these potential benefits, hierarchical multispecies GLMs also usually ignore interactions between species, as they assume that each species' response to the environment represents an independent draw from a common distribution of possible responses. In practice, however, interactions between species will induce unmodeled dependence in the residuals of such a model. These residual correlations violate a primary model assumption if not accounted for, but more importantly, can be used to gain insights into the relative roles of biotic and abiotic constraints on species co-occurrence patterns.

Here we describe a joint species distribution model (JSDM) that introduces correlated occurrence into a hierarchical multivariate probit regression model. The statistical foundation of this general method was introduced over 15 years ago (Chib & Greenberg 1998), but has rarely been applied in the ecological literature. In fact, only 4 of 458 papers that have cited Chib & Greenberg's (1998) seminal work have dealt with ecological problems, and to the best of our knowledge, only two studies (Latimer *et al.* 2009; Clark *et al. In press*) have used a multivariate probit model to fit SDMs (but see Ovaskainen et al. (2010) and Sebastián-González et al. (2010) for a similar approach using multivariate logistic models).

In contrast to these earlier applications, we provide a general introduction to the use and interpretation of these models in ecology. We include a step-by-step tutorial on how to fit and assess multivariate probit models in a Bayesian framework, and include code for running these types of models in R (R Core Team 2013) (See Appendix S1). To illustrate our approach, we examine co-occurrence patterns in natural communities using case studies on

frogs and trees in Victoria, Australia. We demonstrate how these models can provide insights into the underlying causes of similarities and dissimilarities in distributions among species.

Materials and methods

MODEL DESCRIPTION

We model species co-occurrence using a multivariate probit regression model (Chib & Greenberg 1998). Probit regression is a generalized linear model similar to logistic regression (McCullagh & Nelder 1989). Probit regression relates a linear predictor, the standard regression equation used in generalized linear models, to probabilities with a standard normal cumulative distribution function or probit link. In contrast, a logistic regression uses a logit link function.

An alternative way of parameterising a probit model is indirectly with a latent variable formulation, rather than using a probit link directly. Latent (or unobserved) variables are superficially similar to link functions as both are used to relate a continuous linear predictor to discrete binary response data. If we consider a site by species dataset, Y_{ij}, species j is present at site i when a latent variable, Z_{ij}, is greater than zero (and absent if less). Here Z_{ij} is a normal random variate with mean L_{ij} and a standard deviation of 1.

We represent this graphically for two hypothetical species (Fig. 1), with the probability of presence being the shaded area under the density function for values of $Z_{ij} > 0$. The mean of the normal distribution, L_{ij}, is the analogue of the linear predictor in a standard probit regression. A large positive value of L_{ij} implies a high probability of presence, while a large negative value implies a low probability of presence. For example, the probability of presence is 0.69 if $L_{ij} = 0.5$, and is 0.16 if $L_{ij} = -1$ (Fig. 1).

If the latent variable Z_{ij} is independent of the other latent variables in the model (i.e., there is independence among sites and species), then it is a standard probit regression. However, if the

latent variables are correlated, indicating that species presences and absences are not independent, then a multivariate normal distribution must be used to model the values of Z_{ij}. The number of dimensions of the multivariate normal distribution is the number of species being modelled. For example, correlation in a latent bivariate normal distribution influences the joint probabilities of presence and absence of two species, with the probability of joint presence or absence increasing with the correlation coefficient (Fig. 2). However, the probability of presence of each species, unconditional of the presence of the other, is unaffected by the correlation. In our hypothetical example, the probability of presence of species 1 is 0.69 (Fig. 1) regardless of the correlation (summing the probabilities in the two right-hand quadrants in each panel of Fig. 2). Similarly, the probability of presence of species 2 remains 0.16 as the correlation changes.

In the Chib and Greenberg (1998) model, the probability of presence changes when the location of the bivariate normal changes, while the correlations defining the multivariate normal can stay the same. For example, if the mean of the bivariate normal changes from (0.5, –1, as in Fig. 2) to (0.5, –0.5, as in Fig. 3), the probability of presence of species 2 increases to 0.31 (L_{i2} changes from –1 to –0.5), but the probability of presence of species 1 remains 0.69 (L_{i1} remains 0.5). Thus, associations among species are modelled by changing the correlations of the latent multivariate normal distribution, while the (joint) probabilities of presence are modelled by changing the locations of the distribution.

While illustrated schematically here using two species and a bivariate normal distribution, the approach to modelling correlated occurrences extends to any number of J species by using a J-dimensional multivariate normal distribution (Chib and Greenberg 1998). The relationship between correlated normal distributions and correlated Bernoulli events has been used previously to *simulate* correlated fire events (McCarthy & Lindenmayer 1998; McCarthy & Lindenmayer 2000). Here we use it as a basis to *estimate* correlations in the occurrence of species.

MODEL DETAILS

We fit a multivariate model where the probability of occurrence is the probability density of a latent variable exceeding a threshold (eqn. 1). The response is species occurrence, represented by the matrix Y with dimensions n sites by J species with elements Y_{ij}. If the j^{th} species is found at the i^{th} site, then Y_{ij} is one (or zero if absent). The response is predicted by a data matrix (X) that has dimensions n sites by K predictors. All elements of the first column vector of X are ones, which accounts for the model intercept terms, and the remaining column vectors are $K-1$ environmental variables centred on zero and scaled by their standard deviations.

$$\begin{aligned} \Pr(Y_{ij} = 1) &= \Pr(Z_{ij} > 0), \text{ for } i = 1, \ldots, n; j = 1, \ldots, J \\ Z_i &\sim \mathcal{N}_j(X_i B^*, \Sigma) \\ B^*_{jk} &\sim \mathcal{N}(\mu_k, \sigma_k), \text{ for } k = 1, \ldots, K, \end{aligned} \quad (1)$$

The probability that the j^{th} species is present at the i^{th} site equals the probability that the equivalent element of a latent variable matrix, Z_{ij}, is greater than zero, i.e. $Z_{ij} > 0$. The row vectors of the latent variable matrix, Z_i, follow J-dimensional multivariate normal distributions. Each multivariate normal distribution has the same variance-covariance matrix, $\sum$. The mean vector of each multivariate normal distribution is the inner product of the corresponding row vector of the predictor data matrix, X_i, and an unscaled J by K coefficient matrix B^* (equivalent to L_{ij} above). The first column vector of B^* is the unscaled species intercept terms and the remaining K-1 columns are unscaled regression coefficient vectors for the k^{th} environmental variable. The elements of the coefficient matrix, B_{jk}^* are modelled hierarchically by drawing them from normal distributions common to the k^{th} column, with mean μ_k, and standard deviation σ_k.

Our motivation for using a hierarchical approach to estimate the regression coefficients is both ecologically and computationally driven. Having a hierarchical structure to estimate environmental responses of individual species has previously been demonstrated to have

desirable properties for fitting multispecies distribution models (Latimer et al. 2009; Pollock, Morris & Vesk 2012). But in this case there are also advantages of the hierarchical estimation technique that flow on to the correlated occurrence component of the JSDM (see section MODEL FITTING).

A multivariate normal distribution is defined by a variance-covariance matrix, $\sum$, which governs the correlations among variates. Because this approach is based on probit regression, all standard deviations are equal to 1, by definition. In this case, the variance-covariance matrix is a correlation matrix. Specifying a prior for the correlation matrix is not straightforward because elements of correlation matrices are related to each other. The inverse Wishart distribution has the necessary constraints for a variance-covariance matrix (it is positive definite), but this does not constrain the standard deviations to be one. To ensure that the variance-covariance matrix $\sum$ conforms to a correlation matrix, the covariance terms must be divided by the corresponding standard deviations (this is the definition of a correlation coefficient).

As Chib and Greenberg (1998) show, this re-scaling of the variance-covariance matrix so that it becomes a correlation matrix also requires a re-scaling of the coefficients B^* so that they can be interpreted as regular probit regression coefficients. Thus, the scaled probit regression coefficients, B, are calculated by dividing B^* by the standard deviations of the variance-covariance matrix, which are the square root of the diagonal elements (Σ_{jj}). These scaled regression coefficients B_{jk} correspond to the regression coefficients of probit regression for the response of species j to environmental variable k. Thus, the probit of the probability of occupancy of species j at site i

is: $$\text{probit}(\Pr(Y_{ij} = 1)) = B_{j1} + B_{j2}X_{i2} + B_{j3}X_{i3} + \ldots + B_{jK}X_{iK} \quad (2)$$

We can use the output of the model to decompose species correlations into: (a) residual correlation and (b) correlation due to similar environmental responses, which may be used to generate hypotheses about mechanisms that explain why species occur together (or not). For example, strong correlations due to the environment may suggest habitat filtering. Strong residual correlations may hint at a biological interaction between species (e.g. facilitation or competition). Residual correlation may also indicate the need for additional explanatory variables.

1. Correlation parameters

$$P_{jj'} = \frac{\Sigma_{jj'}}{\sqrt{\Sigma_{jj}}\sqrt{\Sigma_{j'j'}}}, \quad \text{for } j = 1, \ldots, J;\ j' = 1, \ldots, J, \tag{3}$$

A correlation matrix *P* can be calculated by rescaling the variance-covariance matrix. To calculate the correlation in the latent distribution between species *j* and species *j*', we divide their covariance by the product of their standard deviations (eqn 3).

2. Correlation due to environment

$$\mathbb{P}_{jj'} = \frac{\sum_{k=1}^{K} B_{jk}B_{j'k} \sum_{k=1}^{K}\sum_{k'=1,k'\neq k}^{K} B_{jk}B_{j'k'}Cov(X_k, X_{k'})}{\sqrt{\left(\sum_{k=1}^{K} B_{jk}^2 \sum_{k=1}^{K}\sum_{k'=1,k'\neq k}^{K} B_{jk}B_{jk'}Cov(X_k, X_{k'})\right)\left(\sum_{k=1}^{K} B_{j'k}^2 \sum_{k=1}^{K}\sum_{k'=1,k'\neq k}^{K} B_{j'k}B_{j'k'}Cov(X_k, X_{k'})\right)}},$$
$$\text{for } j = 1, \ldots, J;\ j' = 1, \ldots, J, \tag{4}$$

We can also calculate a second correlation matrix, $\mathbb{P}_{jj'}$, that accounts for the component of between species correlation due to their shared environmental responses. Equation 4 shows that the environmental correlation between species *j* and *j'* is a function of those species' scaled regression coefficient vectors B_{jk} and $B_{j'k}$ and the covariances of the *k* environmental variables, assuming the environmental data has been centred and scaled appropriately as above.

We include a tutorial and R code for the models as described above in Appendix S1. The R package 'BayesComm' is also available to run a non-hierarchical version of the model described above. This package returns residual correlations between species ($P_{jj'}$), but the current version does not calculate correlations due to shared environmental responses ($\mathbb{P}_{jj'}$). 'BayesComm' is available at http://cran.r-project.org/web/packages/BayesComm/index.html (Golding 2013a; Golding 2013b).

MODEL FITTING

All models were fit with the Markov Chain Monte Carlo Bayesian modelling software JAGS v3.4.0 run through R v3.0.2 via the package R2jags v0.03-11 (R Core Team 2013; Plummer 2014). For both case studies, we ran three chains for 1,000,000 iterations, with the first 15,000 discarded as burn-in. The remaining samples were thinned by a factor of 1000 meaning we retained 985 samples per-chain for post-processing.

We used vague priors for all model parameters in both case studies. We used vague normal priors (mean=0, sd=100) for the elements of μ_k and uniform priors in the interval 0 to 100 for the standard deviations, σ_k. For the variance-covariance matrix we used an inverse-Wishart prior with J+1 degrees of freedom and a J by J identity matrix as the scale matrix. Using these parameters for the inverse-Wishart distribution implies a uniform prior on the off-diagonal elements of P, the correlation coefficients (Gelman & Hill 2007). We found that without regularising the unscaled matrix B^* by applying the hyperprior to the column vectors, the model would not converge without a more informative prior on $\sum$. However, using the hyperparameters μ_k and σ_k allows minimal prior information to be applied to correlation coefficients by setting the degrees of freedom parameter at J+1. We considered model runs converged where after the burn-in, all elements of the parameter matrix B and the off diagonal elements of P had potential scale reduction factor values of less than 1.1.

COMPARISON TO CO-OCCURRENCE INDICES

We plotted environmental and residual correlations from our model against values calculated from two co-occurrence indices commonly used in community ecology: Schoener's Index and a modified version of Dice's Index (Hardy 2008) using 'species.dist' in the picante package (v. 1.6-1) in R (Kembel *et al.* 2010). Co-occurrence indices are often used to infer ecological processes such as potential species interactions, but unlike JSDMs, these indices are not capable of disentangling the influences of shared environmental responses and residual correlations on co-occurrence. Comparing the output of JSDMs to typical co-occurrence indices therefore provides an assessment of how well co-occurrence indices capture these two processes.

CASE STUDY 1: FROG COMMUNITIES IN GREATER MELBOURNE

Our first case study uses data on the occurrence of seven frog species (see Table S1 for a species list) at 104 lentic ponds in parks and gardens around Greater Melbourne, Victoria, Australia. At each site, nocturnal visual searches and acoustic monitoring were conducted three times over two breeding seasons. Anuran assemblages in the study area are strongly influenced by pond size, road cover, and the presence of vertical walls surrounding ponds (Parris 2006). We, therefore, used these three environmental variables in our analyses. Pond surface areas were measured in the field or from aerial photographs. Road cover was quantified by calculating the proportion of a 500 m radius surrounding each pond that was covered by sealed roads. The presence or absence of a vertical wall at each pond was determined during field surveys. For further details see Parris (2006).

CASE STUDY 2: EUCALYPT COMMUNITIES IN THE GRAMPIANS NATIONAL PARK

The *Eucalyptus* dataset includes 12 taxa (see Table S2 for a taxon list) recorded in 458 plots spanning elevation gradients in the Grampians National Park, Victoria, which is known for high species diversity and endemism. The Park has three mountain ranges interspersed with

alluvial valleys and sand sheet, and has a semi-Mediterranean climate with warm, dry summers and cool, wet winters. Plots were based on a nearest-neighbor sampling approach intended to be at a spatial scale in which species interact. Species and ecological traits are tied to environmental gradients, especially soil type and geology (Enright, Miller & Crawford 1994; Pollock, Morris & Vesk 2012). Here, we use six environmental variables previously found to be important to the focal species. Rock cover, soil sand and loam content were quantified in field plots. Valley bottom flatness identifies areas with poor water drainage that accumulate sediment, and was derived from a digital elevation model (Gallant & Dowling 2003). Annual precipitation and temperature variability were estimated using BIOCLIM (Houlder *et al.* 2000). For a further description of the site and environmental variables see Pollock, Morris & Vesk (2012).

Results

Our analyses demonstrate the value of partitioning the effects of the environment ($\mathbb{P}_{jj'}$) from residual interactions between species ($P_{jj'}$). This partitioning revealed contrasting patterns of co-occurrence in the frog and eucalypt case studies (Fig. 4). Specifically, frog species tended to respond similarly to environmental conditions, and have positive residual correlations in co-occurrence (Fig. 5), whereas eucalypt species had much more variable covariance patterns, with numerous cases of both negative and positive correlations in environmental and residual co-occurrence (Figs. 5). In both case studies, environmental correlations tended to be stronger than residual correlations (Figs. 4 and 5).

Many eucalypt species rarely or never co-occur simply because they occupy distinct habitats (negatively correlated estimates of $\mathbb{P}_{jj'}$ in Fig. 5). The more interesting cases are those species that occupy similar environments, yet co-occur more or less than expected. For example, the two species with a particularly high positive residual co-occurrence are from different subgenera, whereas the species with similar environmental responses but negative residual co-occurrence are closely related and able to interbreed (Fig. 5).

COMPARISON OF OUR RESULTS TO TYPICAL CO-OCCURRENCE INDICES

Co-occurrence tends to be positively correlated with environmental correlation in the case of the Schoener Index (Fig. 6) and a modified Dice Index (Fig. S1), although the relationships are generally weak and non-linear. Co-occurrence has no clear relationship with residual correlation (Fig. 6, Fig. S1), indicating the JSDM captures complex interactions that the simple co-occurrence metrics do not.

Discussion

THE IMPORTANCE OF INCLUDING RESIDUAL CORRELATIONS IN SDMs

Species distribution models (SDMs) are widely used to address issues in ecology, evolution and conservation, but current approaches to fitting SDMs make a range of limiting assumptions (Davis *et al.* 1998; Guisan & Thuiller 2005). Here we describe an approach that can help overcome two of the most important of these assumptions – that all relevant environmental covariates are included, and that species distributions are independent of interactions with other species. However, like any correlative method that attempts to partition environmental and residual effects, our approach cannot fully disentangle which of these assumptions is being violated. Residual correlations may be due to missing environmental covariates, or ecological (e.g. facilitation) or evolutionary mechanisms (e.g. allopatric speciation). Nevertheless, examining residual co-occurrence patterns in light of the natural history of the species involved may highlight important environmental variables that are missing from a model, or may point to cases where further research would provide additional insights into biotic interactions.

Clark et al. (*In press*) recently demonstrated that JSDMs reduced inflated estimates of community abundance obtained from aggregating independent SDMs, resulting in more realistic predictions of forest response to climate change. In addition to potentially improving estimates of the responses of species to climate change, this modelling approach may help

determine whether SDMs should be used to interpolate or extrapolate species' distributions across geographic or environmental space more generally. Most SDMs assume that species distributions are in equilibrium with current environmental conditions. However, when there are strong residual correlations between species, projections of species' distributions necessarily assume that interactions will remain constant, which is a questionable assumption given that species will encounter novel biotic communities in different environments. Similarly, if residual correlations are due to missing environmental covariates, then projections of species' distributions might also be suspect.

Most current approaches to incorporating biotic interactions in SDMs involve adding the occurrence or abundance of other species as predictors (Araujo & Luoto 2007; Heikkinen *et al.* 2007; Meier *et al.* 2010). However, adding species as predictors assumes unidirectional interactions and induces multicollinearity within a model when the distribution of a predictor species is governed by similar abiotic variables (Kissling *et al.* 2012). In contrast, the approach presented here directly estimates reciprocal interactions. Incorporating interactions in the residuals, rather than in the mean response, avoids issues of multicollinearity. However, JSDMs do not explicitly model species interactions. A more direct approach to understanding strong unidirectional interactions (e.g. commensalism) would be to relate one species' population dynamics or performance directly to the occurrence of another species via the model mean.

Our analyses demonstrate limitations of using a co-occurrence index to identify potential ecological processes in community ecology studies. There are positive relationships (though subject to uncertainty) between co-occurrence and environmental correlations for frogs and eucalypts (Fig. 6, Fig. S1), which is expected given co-occurring species share environmental responses. However, there are no strong relationships between the co-occurrence indices and residual correlations estimated from JSDMs (Fig. 6, Fig. S1). JSDMs go a step beyond co-occurrence indices because they are a model-based approach that decomposes co-occurrence

into environmental and residual components. Residual correlation does not necessarily indicate a species interaction, but strong residual correlation between species presents a case for further investigation.

IDENTIFYING POTENTIAL BIOTIC INTERACTIONS FROM MODEL OUTPUTS

An important component of the JSDM presented here is that it can partition out the contribution of environmental variables on co-occurrence. The environmental effect itself is important because it highlights the potential role of habitat filtering in community assembly. Conversely, the residual correlation beyond the environmental effect may indicate that other ecological or evolutionary processes are important, though correlated responses to unmeasured covariates cannot be excluded. Previous studies have also examined residual correlations between species after accounting for the effects of environmental covariates using similar models (Latimer *et al.* 2009; Ovaskainen, Hottola & Siitonen 2010; Sebastián-González *et al.* 2010; Clark *et al. In press*), but to the best of our knowledge, no previous studies have explicitly quantified the contribution of shared environmental responses and residual co-occurrence. Our model identified sets of frog and eucalypt species that occurred together more and less than expected given shared responses to the environmental variables we considered. Below, we discuss several potential reasons for this pattern with respect to previous studies and the ecologies of these communities.

Our analysis of frog co-occurrence patterns revealed that all species generally responded similarly to pond area, road density, and the presence of a pond barrier, and the magnitude and direction of these effects are consistent with earlier studies of species richness in the study area (Parris 2006), as well as findings from a wide range of studies on the occurrence of pond-breeding amphibians (Chardon 1998; Popescu & Gibbs 2009; Heard *et al.* 2013). However, many of the residual correlations between species were positive, suggesting that species co-occurred more than expected given their shared responses to environmental variables. Facilitative interactions between frogs seem unlikely, but positive residual

correlations between frog species could have been due to a shared response to an abiotic variable that was not considered in our model, such as the presence of fish (Hamer and Parris (2013).

In contrast to the frog case study, analyses of eucalypt communities revealed considerable variability in shared environmental responses and residual correlations among species. Two eucalypt species that co-occur more than expected in neighbourhood plots given their responses to environmental variables are *Eucalyptus arenacea* and *E. goniocalyx*, which are from different subgenera (Fig. 5). Co-dominance of eucalypts from different subgenera is a common pattern known as Pryor's rule (Pryor 1953). A possible explanation for this pattern is that species from different subgenera are able to differentiate resource use, thereby alleviating competition (Austin, Cunningham & Wood 1983). Another potential explanation is that species from different subgenera are not able to interbreed (Ellis, Sedgley & Gardner 1991). If species are able to interbreed and one species has a reproductive advantage (e.g. if selection favouring one species leads to more pollen output from that species), then, with continued back-crossing, all individuals begin to resemble the favoured species (Levin 2006). Simulations suggest interbreeding species usually form unstable assemblages because one species tends to gain a reproductive advantage over the other (Currat *et al.* 2008). Interbreeding (i.e. hybridization) has influenced the evolution of eucalypts, and may be a mechanism for dispersal (Potts & Reid 1988). In our study, the species pairs that are reproductively compatible (black dots in Fig. 5) occupy similar environments, but have negatively correlated residuals (bottom right quadrant of Fig. 5). In other words, these species co-occur less frequently together than we would expect given their similar habitat preferences.

FURTHER REFINEMENTS AND APPLICATIONS OF THE MODEL

One advantage of the hierarchical modelling framework used here is that it can be easily modified to account for additional complexities and uncertainties in the data. Additional

correlations between species (e.g. functional similarity or phylogenetic relatedness) could be incorporated into the model. For instance, similarity in specific leaf area (SLA) between eucalypt species increases with shared environmental space, yet slightly decreases with increasing residual correlation (Fig. 7). Specific leaf area is functionally related to species occurrences along the environmental gradients studied here, explaining the positive correlation between SLA similarity and shared environmental space (Pollock, Morris & Vesk 2012). Other species traits may be related to residual co-occurrence. For example, flowering time might be related to residual correlation if temporal niche partitioning is important. In these cases, incorporating functional traits as predictors of environmental responses in the model (e.g. Pollock, Morris & Vesk 2012) may reveal additional ecological insights.

Our model could also be extended to incorporate imperfect detection probabilities (MacKenzie *et al.* 2002) or spatial random effects (Parris 2006). In cases where data are available for more than one time period, our approach could also be used to analyse correlates of co-occurrence in community time-series, where the effects of biotic interactions may be more readily identified (Mutshinda, O'Hara & Woiwod 2011; Kissling *et al.* 2012). For example, Sebastián-González *et al*. (2010) used a similar approach to analyse the effects of heterospecific attraction on the temporal dynamics of seven waterbird species. Our model thus offers a flexible approach for examining a wide range of questions in theoretical and applied ecology, and could be adapted to suit a variety of applications.

Acknowledgements

This paper was much improved by discussions with Dave Harris and by suggestions from Otso Ovaskainen and two anonymous reviewers.

Data Accessibility

R scripts: uploaded as online supporting information

Supporting Information

Additional Supporting Information may be found in the online version of this article.

Appendix S1. Tutorial for fitting a Joint Species Distribution Model (JSDM) in R.

Appendix S2. Descriptions of Schoener's and Dice's co-occurrence indices.

Table S1. List of frog species included in case study 1.

Table S2. List of eucalypt species included in case study 2.

Figure S1. Median (± 95% credible intervals) residual correlations ($P_{jj'}$) and environmental correlations ($\mathbb{P}_{jj'}$) compared to a modified Dice's co-occurrence index for frogs (left panels) and eucalypts (right panels).

References

Araujo, M.B. & Luoto, M. (2007) The importance of biotic interactions for modelling species distributions under climate change. *Global Ecology and Biogeography,* **16,** 743-753.

Austin, M.P. (2002) Spatial prediction of species distribution: an interface between ecological theory and statistical modelling. *Ecological Modelling,* **157,** 101-118.

Austin, M.P., Cunningham, R.B. & Wood, J.T. (1983) The subgeneric composition of Eucalypt forest stands in a region of south-eastern Australia. *Australian Journal of Botany,* **31,** 63-71.

Chardon, V. (1998) Effects of habitat fragmentation and road density on the distribution pattern of the moor frog *Rana arvalis*. *Journal of Applied Ecology,* **35,** 44-45.

Chib, S. & Greenberg, E. (1998) Analysis of multivariate probit models. *Biometrika,* **85,** 347-361.

Clark, J.S., Gelfand, A.E., Woodall, C.W. & Zhu, K. (*In press*) More than the sum of the parts: Forest climate response from Joint Species Distribution Models. *Ecological Applications,* http://dx.doi.org/10.1890/13-1015.1.

Currat, M., Ruedi, M., Petit, R.J. & Excoffier, L. (2008) The hidden side of invasions: massive introgression by local genes. *Evolution,* **62,** 1908-1920.

Davis, A.J., Jenkinson, L.S., Lawton, J.H., Shorrocks, B. & Wood, S. (1998) Making mistakes when predicting shifts in species range in response to global warming. *Nature***,** 783-786.

Dormann, C.F., Schymanski, S.J., Cabral, J., Chuine, I., Graham, C.H., Hartig, F., Kearney, M., Morin, X., Romermann, C., Schroder, B. & Singer, A. (2012) Correlation and process in species distribution models: bridging a dichotomy. *Journal of Biogeography,* **39,** 2119-2131.

Ellis, M.F., Sedgley, M. & Gardner, J.A. (1991) Interspecific pollen-pistil interaction in *Eucalyptus* L'Hér. (Myrtaceae): the effect of taxonomic distance. *Annals of Botany,* **68,** 185-194.

Enright, N.J., Miller, B.P. & Crawford, A. (1994) Environmental correlates of vegetation patterns and species richness in the northern Grampians, Victoria. *Australian Journal of Ecology,* **19,** 159-168.

Gallant, J.C. & Dowling, T.I. (2003) A mutliresolution index of valley bottom flatness for mapping depositional areas. *Water Resources Research,* **39,** 1347.

Gelfand, A.E., Silander Jr., J.A., Wu, S., Latimer, A., Lewis, P.O., Rebelo, A.G. & Holder, M.T. (2003) Explaining species distribution patterns through hierarchical modeling. *Bayesian Analysis,* **1,** 1-35.

Gelman, A. & Hill, J. (2007) *Data Analysis Using Regression and Multilevel/Hierarchical models*. Cambridge University Press, New York.

Golding, N. (2013a) BayesComm: Bayesian community ecology analysis. R package version 0.1-0. http://CRAN.R-project.org/package=BayesComm.

Golding, N. (2013b) Mapping and understanding the distributions of potential vector mosquitoes in the UK: New methods and applications. Doctor of Philosophy, University of Oxford.

Guisan, A. & Thuiller, W. (2005) Predicting species distribution: offering more than simple habitat models. *Ecology Letters,* **8,** 993-1009.

Hamer, A.J. & Parris, K.M. (2013) Predation modifies larval amphibian communities in urban wetlands. *Wetlands,* **33,** 641-652.

Hardy, O.J. (2008) Testing the spatial phylogenetic structure of local communities: statistical performances of different null models and test statistics on a locally neutral community. *Journal of Ecology,* **96,** 914-926.

Heard, G.W., McCarthy, M.A., Scroggie, M.P., Baumgartner, J.B. & Parris, K. (2013) A Bayesian model of metapopulation viability, with application to an endangered amphibian. *Diversity and Distributions,* **19,** 555-566.

Heikkinen, R.K., Luoto, M., Virkkala, R., Pearson, R. & Korber, J.-H. (2007) Biotic interactions improve prediction of boreal bird distributions at macro-scales. *Global Ecology and Biogeography,* **16,** 754-763.

Helmus, M.R., Savage, K., Diebel, M.W., Maxted, J.T. & Ives, A.R. (2007) Separating the determinants of phylogenetic community structure. *Ecology Letters,* **10,** 917-925.

Houlder, D.J., Hutchison, M.F., Nix, H.A. & McMahon, J.P. (2000) ANUCLIM User Guide, Version 5.1. Centre for Resource and Environmental Studies, Australian National University, Canberra.

Hutchinson, G.E. (1957) Concluding remarks. *Cold Spring Harbor Symposia on Quantitative Biology,* **22,** 415-427.

Kembel, S.W., Cowan, P.D., Helmus, M.R., Cornwell, W.K., Morlon, H., Ackerly, D.D., Blomberg, S.P. & Webb, C.O. (2010) Picante: R tools for integrating phylogenies and ecology. *Bioinformatics,* **26,** 1463-1464.

Kissling, W.D., Dormann, C.F., Groeneveld, J., Hickler, T., Ingolf, K., McInerny, G.J., Montoya, J.M., Römermann, C., Schiffers, K., Schurr, F.M., Singer, A., Svenning, J.-C., Zimmermann, N.E. & O'Hara, R.B. (2012) Towards novel approaches to modelling biotic interactions in multispecies assemblages at large spatial extents. *Journal of Biogeography,* **39,** 2163-2178.

Latimer, A.M., Banerjee, S., Sang, H., Mosher, E.S. & Silander Jr, J.A. (2009) Hierarchical models facilitate spatial analysis of large data sets: a case study on invasive plant species in northeastern United States. *Ecology Letters,* **12,** 144-154.

Leathwick, J.R. (2002) Intra-generic competition among Nothofagus in New Zealand's primary indigenous forests. *Biodiversity and Conservation,* **11,** 2177-2187.

Leathwick, J.R. & Austin, M.P. (2001) Competitive interactions between tree species in New Zealand's old-growth indigenous forests. *Ecology,* **82,** 2560-2573.

Levin, D.A. (2006) The spatial sorting of ecological species: Ghost of competition or of hybridization past? *Systematic Botany,* **31,** 8-12.

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

McCarthy, M.A. & Lindenmayer, D.B. (1998) Multi-aged mountain ash forest, wildlife conservation and timber harvesting. *Forest Ecology and Management,* **104,** 43-56.

McCarthy, M.A. & Lindenmayer, D.B. (2000) Spatially-correlated extinction in a metapopulation of Leadbeater's possum. *Biodiversity and Conservation,* **9**.

McCullagh, P. & Nelder, J.A. (1989) *Generalized linear models*. Chapman and Hall, London.

Meier, E.S., Kienast, F., Pearman, P.B., Svenning, J.C., Thuiller, W., Araújo, M.B., Guisan, A. & Zimmermann, N.E. (2010) Biotic and abiotic variables show little redundancy in explaining tree species distributions. *Ecography,* **33,** 1038-1048.

Mutshinda, C.M., O'Hara, R.B. & Woiwod, I.P. (2011) A multispecies perspective on ecological impacts of climatic forcing. *Journal of Animal Ecology,* **80,** 101-107.

Ovaskainen, O., Hottola, J. & Siitonen, J. (2010) Modeling species co-occurrence by multivariate logistic regression generates new hypotheses on fungal interactions. *Ecology,* **91,** 2514-2521.

Ovaskainen, O. & Soininen, J. (2011) Making more out of sparse data: hierarchical modeling of species communities. *Ecology,* **92,** 289-295.

Parris, K. (2006) Urban amphibian assemblages as metacommunities. *Journal of Animal Ecology,* **75,** 757-764.

Pellissier, L., Anne Bråthen, K., Pottier, J., Randin, C.F., Vittoz, P., Dubuis, A., Yoccoz, N.G., Alm, T., Zimmermann, N.E. & Guisan, A. (2010) Species distribution models reveal apparent competitive and facilitative effects of a dominant species on the distribution of tundra plants. *Ecography,* **33,** 1004-1014.

Plummer, M. (2014) rjags: Bayesian graphical models using MCMC. *R package version 3-12*.

Pollock, L.J., Morris, W.K. & Vesk, P.A. (2012) The role of functional traits in species distributions revealed through a hierarchical model. *Ecography*.

Popescu, V.D. & Gibbs, J.P. (2009) Interactions between climate, beaver activity, and pond occupancy by the cold-adapted mink frog in New York State, USA. *Biological Conservation,* **142,** 2059-2068.

Potts, B.M. & Reid, J.B. (1988) Hybridization as a dispersal mechanism. *Evolution,* **42,** 1245-1255.

Pryor, L.D. (1953) Genetic control in *Eucalyptus* distributions. *Proceedings of the Linnean Society of New South Wales,* **78,** 8-18.

R Core Team (2013) R: A language and environment for statistical computing. *R Foundation for Statistical Computing*. Vienna, Austria.

Schweiger, O., Heikkinen, R.K., Harpke, A., Hickler, T., Klotz, S., Kudrna, O., Kuhn, I., Poyry, J. & Settele, J. (2012) Increasing range mismatching of interacting species under global change is related to their ecological characteristics. *Global Ecology and Biogeography,* **21,** 88-99.

Sebastián-González, E., Sánchez-Zapata, J.A., Botella, F. & Ovaskainen, O. (2010) Testing the heterospecific attraction hypothesis with time-series data on species co-occurrence. *Proceedings of the Royal Society B: Biological Sciences,* **277,** 2983-2990.

Webb, C.O., Ackerly, D.D., McPeek, M.A. & Donoghue, M.J. (2002) Phylogenies and community ecology. *Annual Reviews of Ecology and Systematics,* **33,** 475-505.

Wisz, M.S., Pottier, J., Kissling, W.D., Pellissier, L., Lenoir, J., Damgaard, C.F., Dormann, C.F., Forchhammer, M.C., Grytnes, J.-A., Guisan, A., Heikkinen, R.K., Hoye, T.T., Kuhn, I., Luoto, M., Maiorano, L., Nilsson, M.-C., Normand, S., Ockinger, E., Schmidt, N.M., Termansen, M., Timmermann, A., Wardle, D.A., Aastrup, P. & Svenning, J.-C. (2013) The role of biotic interactions in shaping distributions and realised assemblages of species: implications for species distribution modelling. *Biological Reviews,* **88,** 15-30.

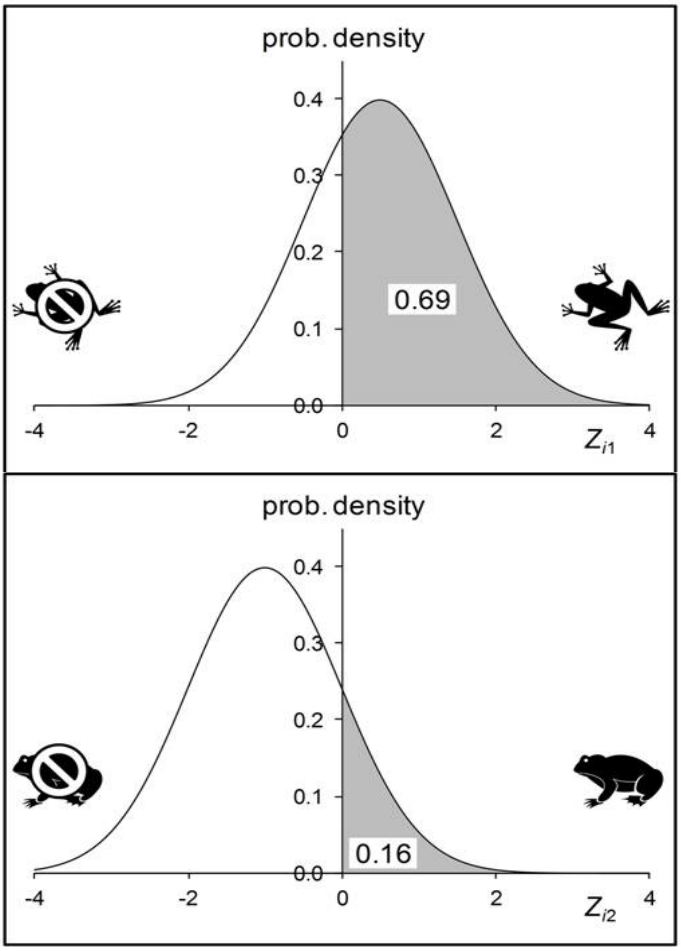

Figure 1. Probit regression for occurrence of two hypothetical species (j = 1, the tree frog, or j = 2, the toad) at a particular site i depicted using probability density functions of the latent normal variate Z_{ij}. The species would occur at the site when the latent random variable, which has a standard deviation of 1, is greater than 0. Thus, the mean of the latent variable (L_{i1} = 0.5, L_{i2} = -1.0) determines the probability of occurrence. The probability of occurrence equals the shaded area under the density function greater than zero (0.69 and 0.16). These representations of individual species ignore patterns of co-occurrence.

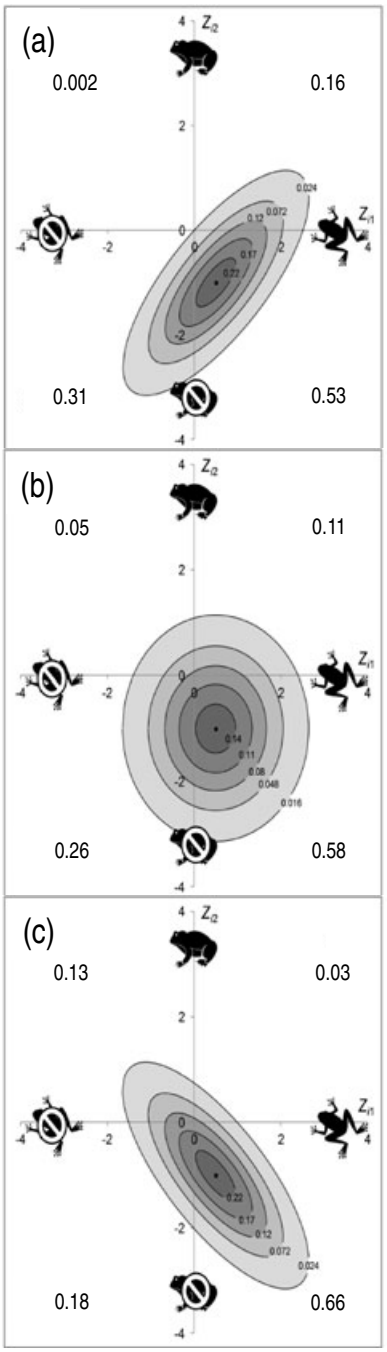

Figure 2. Co-occurrence patterns of the two species from Fig. 1, modelled using a bivariate normal distribution represented as contour plots of probability density, with correlation 0.75 (a), 0.0 (b), and –0.75 (c). The numbers on the contours (the concentric ellipses) are the probability densities that encompass 0.1, 0.3, 0.5, 0.7 and 0.9 of the volume under the bivariate normal distribution. Each species occurs at the

site when the corresponding random variate is greater than 0. Thus, species 1 (the tree frog) occurs when Z_{i1} is greater than zero (the right-hand quadrants), and species 2 (the toad) occurs when Z_{i2} is greater than zero (the upper quadrants). The joint probabilities of occurrence are indicated by the values in the corners. In all cases shown, the probability of occurrence of species 1 is 0.69 (the sum of the probabilities in the right-hand quadrants) because the mean of Z_{i1} (L_{i1}) remains 0.5, as in Fig. 1. Similarly, the probability of occurrence of species 2 remains 0.16 because the mean of Z_{i2} (L_{i2}) remains –1. The correlation changes the probabilities of co-occurrence, but not the unconditional probabilities of occurrence for each species.

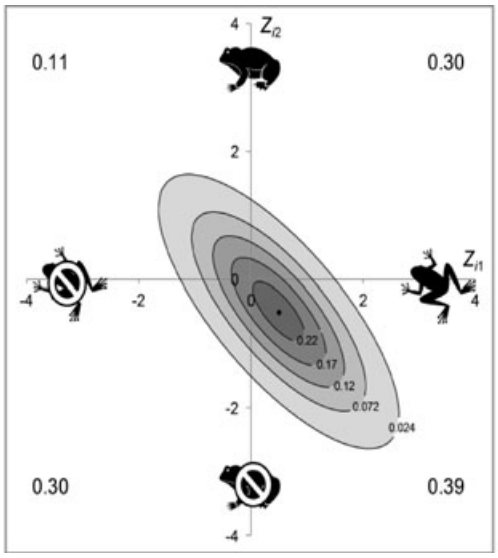

Figure 3. An equivalent representation of co-occurrence patterns of the two species in Fig. 2c, but with a higher probability of occurrence of species 2 (the toad at 0.41) because the mean of Z_{i2} (L_{i2}) has increased from –1 (in Fig. 2) to –0.5. The probabilities of co-occurrence of the species have also changed as a result, while the probability of occurrence of species 1 is the same (0.69) as the mean of Z_{i1} (L_{i1}) is unchanged.

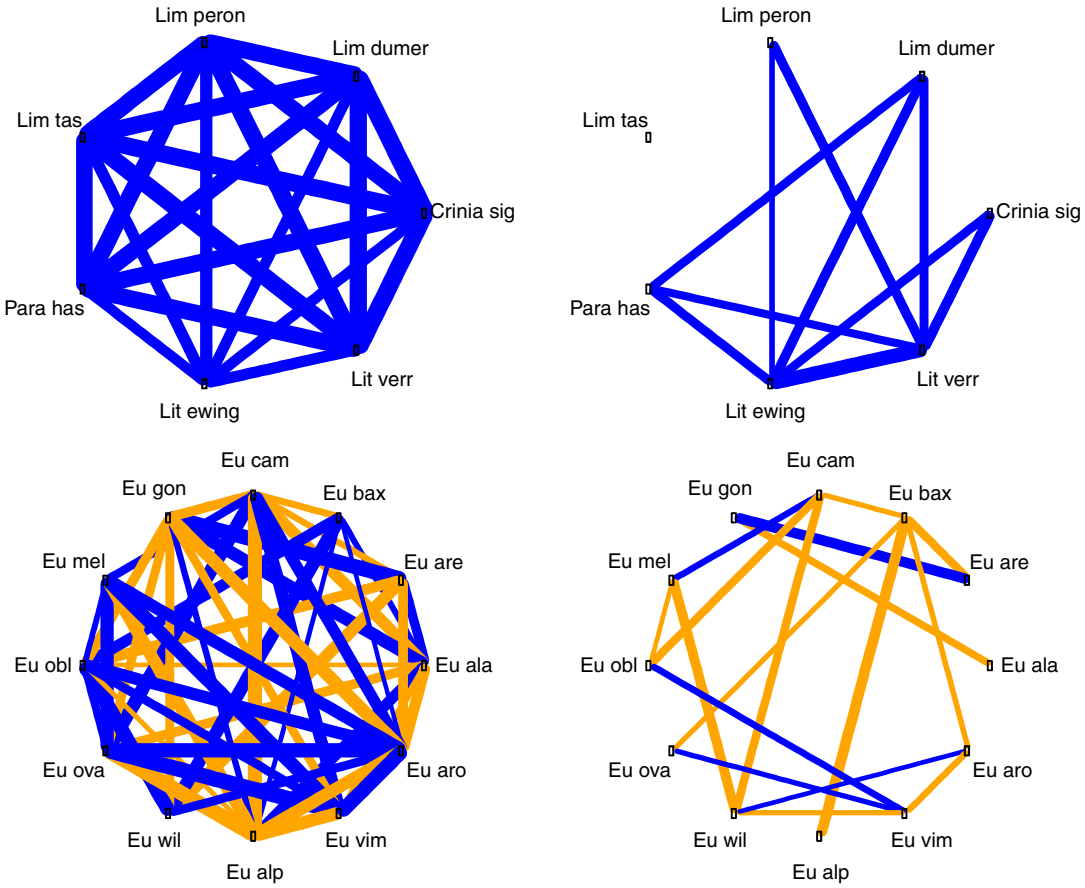

Figure 4. Network diagrams representing modelled environmental correlation ($\mathbb{P}_{jj'}$) (left panels) and residual correlation ($P_{jj'}$) (right panels) between species of frogs (top) and eucalypts (bottom). Blue (dark) lines are positive correlations between species and orange (pale) lines are negative correlations. Line thickness represents correlation strength. Only correlations in which the credible intervals do not cross zero are shown. See Tables S1 and S2 for full species names.

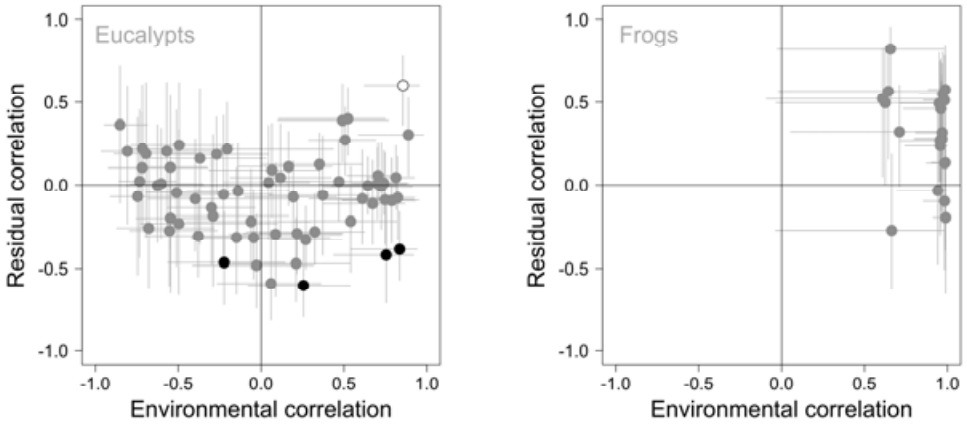

Figure 5. Modelled residual correlation ($P_{jj'}$) and environmental correlation ($\mathbb{P}_{jj'}$) between species pairs for eucalypts and frogs. Error bars represent 95% credible intervals. Black circles are eucalypt species pairs that interbreed. The open circle represents the pair *E. baxteri*-*E. goniocalyx*, species from different subgenera.

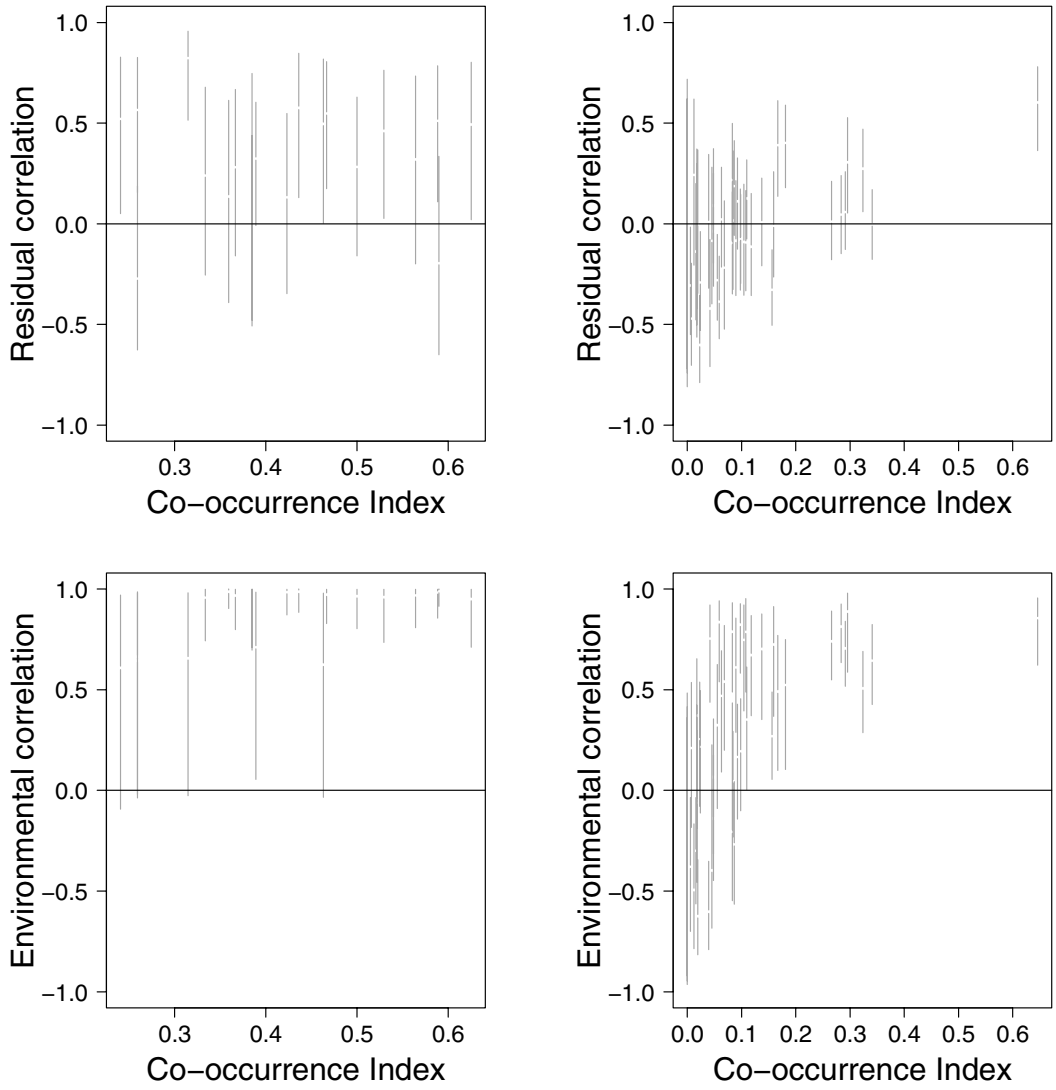

Figure 6. Median (± 95% credible intervals) residual correlations ($P_{jj'}$) and environmental correlations ($\mathbb{P}_{jj'}$) compared to a co-occurrence index (Schoener Index; see methods) for frogs (left panels) and eucalypts (right panels).

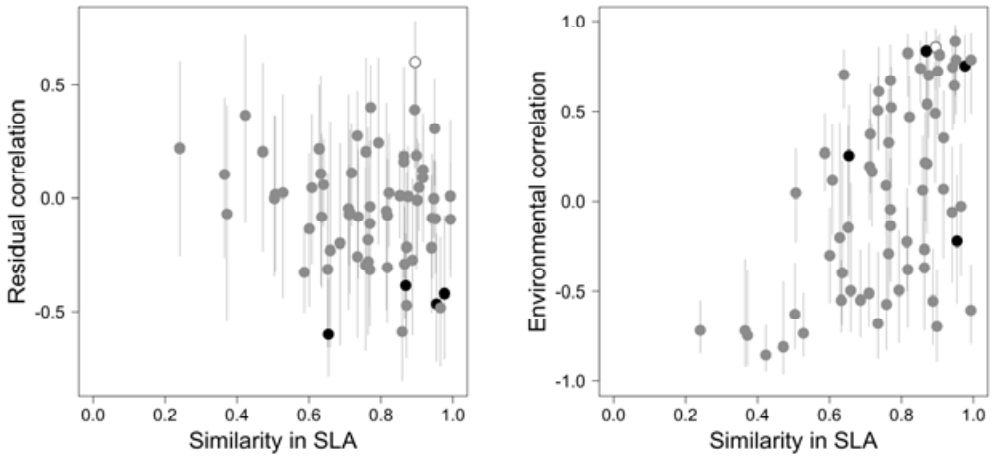

Figure 7. Relationship between residual correlation ($P_{jj'}$) and environmental correlation ($\mathbb{P}_{jj'}$) and similarity in ln-transformed mean specific leaf area (SLA) between pairs of eucalypt species. Black circles are eucalypt species pairs that interbreed. The open circle represents the pair *E. baxteri-E. goniocalyx*, species from different subgenera.