Author/s:
Tobler, MW;Kéry, M;Hui, FKC;Guillera-Arroita, G;Knaus, P;Sattler, T

Title:
Joint species distribution models with species correlations and imperfect detection

Date:
2019-08-01

Citation:
Tobler, M. W., Kéry, M., Hui, F. K. C., Guillera-Arroita, G., Knaus, P. & Sattler, T. (2019). Joint species distribution models with species correlations and imperfect detection. Ecology, 100 (8), https://doi.org/10.1002/ecy.2754.

Persistent Link:
https://hdl.handle.net/11343/254332

DR. MATHIAS W. TOBLER (Orcid ID : 0000-0002-8587-0560)

Article type : Articles

**Running header: JSDMs with imperfect detection**

# Joint species distribution models with species correlations and imperfect detection

Mathias W. Tobler[1*], Marc Kéry[2], Francis K. C. Hui[3], Gurutzeta Guillera-Arroita[4], Peter Knaus[2] & Thomas Sattler[2]

[1] San Diego Zoo Global, Institute for Conservation Research, 15600 San Pasqual Valley Rd. Escondido, CA, 92027, USA

[2] Swiss Ornithological Institute, Seerose 1, 6204 Sempach, Switzerland

[3] Research School of Finance, Actuarial Studies & Statistics, Australian National University, Acton, ACT 0200, Australia

[4] School of BioSciences, University of Melbourne, Parkville, VIC 3010, Australia

[*] corresponding author e-mail: mtobler@sandiegozoo.org

**Abstract**

Spatiotemporal patterns in biological communities are typically driven by environmental factors and species interactions. Spatial data from communities are naturally described by stacking models for all species in the community. Two important considerations in such multi-species or joint species distribution models (JSDMs) are measurement errors and correlations between species. Up to now, virtually all JSDMs have included either one or the other, but not both features simultaneously, even though both measurement errors and species correlations may be essential for achieving unbiased inferences about the distribution of communities and species co-occurrence patterns. We developed two presence-absence JSDMs for modeling pairwise species correlations while accommodating imperfect detection; one using a latent variable and the other using a multivariate probit approach. We conducted three simulation studies to assess the performance of our new models and to compare them to earlier latent variable JSDMs that did not consider imperfect detection. We illustrate our models with a large Atlas data set of 62 passerine bird species in Switzerland. Under a wide range of conditions, our new latent variable JSDM with imperfect detection and species correlations yielded estimates with little or no bias for occupancy, occupancy regression coefficients and the species correlation matrix. In contrast, with the multivariate probit model we saw convergence issues with large datasets (many species and sites) resulting in very long runtimes and larger errors. A latent variable model that ignores imperfect detection produced correlation estimates that were consistently negatively biased i.e., underestimated. We found that the number of latent variables required to adequately represent the species correlation matrix may be much greater than previously suggested, namely around $n/2$, where $n$ is community size. The analysis of the Swiss passerine dataset exemplifies how not accounting for imperfect detection will lead to negative bias in occupancy estimates and to attenuation in the estimated covariate coefficients in a JSDM. Furthermore, spatial heterogeneity in detection may cause spurious patterns in the estimated species correlation matrix if not accounted for. Our new JSDMs represent an important

extension of current approaches to community modeling to the common case where species presence-absence cannot be detected with certainty.

**Keywords:** BUGS; community modelling; detection probability; interaction; JSDM; latent variable; multivariate probit; occupancy model; passerine bird

## Introduction

The distribution and composition of species communities is shaped both by abiotic conditions and biotic interactions (Morin 2009). Species distribution models (SDMs, Elith and Leathwick 2009) have been widely used to study the environmental factors that influence the occurrence of species and to predict or forecast their distributions at larger spatial and/or temporal scales. While initially formulated for single species, SDMs have been recently extended to describe data recorded for multiple species by stacking single-species models, usually linked together via species-specific random effects, resulting in a type of hierarchical community model. Such models have often been referred to as joint species distribution models (JSDMs), because they jointly model multiple species. This stacking principle for community models has been invented and re-invented multiple times, coming from different perspectives.

In a first line of research, Dorazio and Royle (2005; see also Gelfand et al. 2005 and Dorazio et al., 2006) formulated a JSDM as a multi-species variant of an occupancy-detection model (MacKenzie et al. 2002), i.e., a hierarchical model containing two regressions, one to describe the true presence-absence of each species and the other to describe the observed detection/non-detection data, conditional on the latent presence-absence states of each species. This model accommodates imperfect detection of each species and allows covariates that influence the occurrence and/or the detection of a species to be introduced (Kéry and Royle 2016, chapter 11). It has since been extended to describe community dynamics (Dorazio et al.

2010) and to treat abundance as the response rather than presence-absence (Yamaura et al. 2011, Yamaura et al. 2012, Sollmann et al. 2015).

The original Dorazio-Royle community models do not contain parameters to capture residual correlations in occupancy probability that may arise as a consequence of biotic interactions among species or the effects of unmeasured covariates. However, species interactions often have an important impact on the distribution of species and the composition of communities through competition, facilitation, or predation (Cody and Diamond 1975, Begon et al. 2006, Morin 2009), and hence, it might seem desirable to include this feature of a community in these models.

A second line of research also formulated the modeling of a community as a stack of single-species models but focused on non-independent occurrence by explicitly addressing pairwise correlations between species (Latimer et al. 2009, Ovaskainen et al. 2010, Pollock et al. 2014, Hui et al. 2015, Warton et al. 2015). These models estimate the strength of positive or negative residual correlations in the apparent occupancy probability, i.e., the product of occupancy and detection probability (Kéry 2011) and they differ mostly in the precise manner in which the correlation is specified. Some authors have used multivariate logit or probit models that include an unstructured matrix of pairwise correlations for all species and therefore require a large number of parameters as species numbers increase (Latimer et al. 2009, Ovaskainen et al. 2010, Pollock et al. 2014). Others have proposed latent variable models as a computationally more efficient approximation to the models with a fully unstructured correlation matrix (Hui et al. 2015, Warton et al. 2015). Latent-variable models have the added advantage that they form the basis for model-based ordination (Hui et al. 2015, Warton et al. 2015). Regardless of the structure used for capturing correlations, a common feature of these recent developments is that they have failed to account for imperfect species detection, which has the potential to bias the estimation of virtually every descriptor of species distributions and of communities (MacKenzie 2005, Kéry 2011, Ruiz-Gutiérrez and Zipkin 2011, Guillera-Arroita et al. 2014,

Beissinger et al. 2016, Kéry and Royle 2016, chapter 11). Hence, it has been argued repeatedly that it would be desirable to incorporate this important feature of measurement error in real ecological data into such JSDMs as well (Beissinger et al. 2016, Warton et al. 2016).

Only a small number of papers have confronted the challenge of simultaneously modeling species correlations and imperfect detection, but usually their models were restricted to two or just a handful of species (MacKenzie et al. 2004, Richmond et al. 2010, Waddle et al. 2010, Sollmann et al. 2012, Dorazio et al. 2015, Rota et al. 2016b; but see Rota et al. 2016a). In this paper, we unify the two lines of research above by developing two JSDMs that account for *both* imperfect species detection and residual correlations in occurrence, allowing application to a much larger number of species. We describe a latent variable and a multivariate probit variant of a multi-species occupancy model with residual correlation, and thus in a straightforward fashion extend the work of Hui et al. (2015) and of Pollock et al. (2014), to accommodate a hallmark of all ecological data: imperfect detection (Iknayan et al. 2014, Beissinger et al. 2016, Kéry and Royle 2016). We use simulations to evaluate and compare the performance of our models under different sample sizes and illustrate their application with a large real-world dataset of 62 passerine bird species in Switzerland. We implement all our models in the BUGS language, thus making them accessible and, especially, easily generalizable to practitioners.

## Methods

### *Data requirements*

Our JSDMs require measurements of species presence-absence at the sampling sites ( $y_{ij}$, where $i = 1 \dots n$ refers to species, and $j = 1 \dots J$ refers to sites) (Kéry and Royle 2016, chapter 11). By writing 'measurements', we emphasize that these records are *not* necessarily the same as *true* presence and absence, because in practice, measurements are usually contaminated by two sorts of errors: false-positives, e.g., when one species is misidentified for another, and more

commonly false-negatives, when one species is overlooked at a site where it occurs (Kéry and Royle 2016, chapter 1). Here, we assume that false positives do not occur. Accounting for false negatives in the modelling of species occurrence (MacKenzie et al. 2002, Guillera-Arroita 2017, MacKenzie et al. 2018) typically requires repeated presence-absence measurements (also known as detection/non-detection data), such that we have $y_{ijk}$, where the additional index $k$ denotes the repeated measurement, for $k = 1...K$. Repeats need to take place over a relatively short time interval, such that the closure assumption is satisfied: that is, the true presence or absence $z_{ij}$ of species $i$ at site $j$ must not change over the duration of the $K$ measurements (if change is random estimation is still possible, only that the state variable should be interpreted as usage, rather than continuous presence, Mackenzie and Royle 2005) Not all sites need the same degree of replication or indeed any replication at all, i.e., we may have a site-specific *K*: $K_j$. In contrast, models that do not account for imperfect detection make implicit assumptions that either detection is perfect or that detection does not change across sampling sites. The inferences of these simpler models are then restricted to what has been called *apparent* rather than true occupancy probability (Kéry 2011, Lahoz-Monfort et al. 2014).

*Model description*

We extend two existing JSDMs to include a sub-model for imperfect detection: the latent-variable model (Hui et al. 2015, Warton et al. 2015) and the multivariate probit model (Pollock et al. 2014). Equivalently we could say that we extend existing multi-species occupancy models (Dorazio and Royle 2005, Gelfand et al. 2005, Dorazio et al. 2006) to include residual correlation in species occupancy probabilities. Next, we briefly describe this latter model and then show how we extend it to include species residual correlation either with a latent-variable construction or with a multivariate probit model.

*The Dorazio-Royle multi-species occupancy model* – Let the discrete latent variable $z_{ij}$ indicate the true presence state of species *i* at site *j*. For computational reasons (related to the modelling of the correlations), here we formulate the occupancy component of the model using a probit instead of a logit link, which is more customary for binomial responses in ecology. To implement the probit regression for each species, we can express $z_{ij}$ via a continuous normally-distributed latent variable $u_{ij}$ such that $z_{ij} = I(u_{ij} > 0)$, where I(.) is the indicator function which takes value 1 if the condition in brackets holds and zero otherwise (i.e. here $z_{ij}=1$ if $u_{ij} > 0$ and $z_{ij} = 0$ if $u_{ij} \leq 0$). The variance of $u_{ij}$ is constrained to be one for parameter identifiability reasons, and covariate effects can be incorporated into its mean as is analogous to standard linear regression. The occupancy component of the model can then be described as follows:

$$z_{ij} = I(u_{ij} > 0) ,$$

$$u_{ij} = Xocc_j \beta occ_i + \varepsilon_{ij},$$

$$\varepsilon_{ij} \sim \text{Normal}(0,1) ,$$

where $Xocc_j$ is a vector of environmental covariates for site *j* with the first element set to 1 for the intercept, and $\beta occ_i$ is the corresponding vector of species-specific regression coefficients for species *i*.

The detection part of the model describes the detection frequencies following a binomial distribution governed by the probability of detection $p_{ij}$, which can be expressed as a function of covariates e.g. using a logistic regression model as follows:

$$y_{ij} \sim \text{Binomial}(K_j, z_{ij} * p_{ij}) ,$$

$$\text{logit}(p_{ij}) = Xobs_j \beta obs_i,$$

where the response $y_{ij}$ is the number of sampling occasions out of $K_j$ when species *i* was detected at site *j*, $Xobs_j$ is a vector of detection covariates with the first element set to 1 for the intercept, and $\beta obs_i$ is a vector of species-specific regression coefficients related to the detection sub-model. This part of the model would be replaced by a set of independent Bernoulli trials if the

probability of detection is survey-specific (i.e., if binary, detection/non-detection data are modeled, as in our case study below). Typically, all regression coefficients are modelled hierarchically among species to allow improved estimates for rare species (Kéry and Royle 2008, Zipkin et al. 2009, Ovaskainen and Soininen 2011) and enhance rates of convergence in an MCMC-based analysis (see below). This means that species-level parameters are treated as random effects, e.g., $\beta_i \sim \text{Normal}(\mu,\sigma^2)$, where $\mu$ and $\sigma^2$ are the mean and the variance of coefficient $\beta$ in the wider community of species from which the study species were drawn (alternatively, $\mu$ could be interpreted as the coefficient of the 'average species' in the modelled community).

The model described so far is simply a variant of the standard multi-species occupancy model (Dorazio and Royle 2005, Dorazio et al. 2006) with a probit regression for the occupancy component. In this paper, we extend the multi-species occupancy model described above by allowing for residual correlation in the occupancy probability that cannot be explained by the environmental covariates in the model.

*Including species correlations using a latent-variable model* – Our first extension uses a latent-variable approach (Hui et al. 2015). We introduce a set of $T$ latent variables $l_j = (l_{j1},\ldots, l_{jT})$ (also referred to as "factors" in ordination analysis) and a vector of $T$ corresponding species-specific latent variable coefficients $\theta_i = (\theta_{i1},\ldots, \theta_{iT})$ (also often referred to as "loadings" in ordination). The latent variables $l$ can be thought of as unmeasured site-level covariates; they are unknown, and specified in the model as random variables from a standard normal distribution. The coefficients $\theta$ are constrained to lie between -1 and 1 using a uniform prior distribution; this constraint is needed for parameter identifiability reasons with binary responses. Thus, the occupancy submodel becomes the following.

$$z_{ij} = I(u_{ij} > 0),$$

$$u_{ij} = Xocc_j \beta occ_i + l_j \theta_i + \varepsilon_{ij},$$

$$\sigma_i^2 = 1 - \sum_{t=1}^{T} \theta_{it}^2,$$

$\varepsilon_{ij} \sim \text{Normal}(0, \sigma_i^2)$.

With more than a single latent variable (i.e., when $T > 1$), we need to impose constraints on $\theta$ additional to those given above (Hui et al. 2015) to ensure parameter identifiability. In particular, if $\theta$ is an $n$ x $T$ matrix of coefficients for $T$ latent variables and $n$ species, the diagonal elements are constrained to lie between 0 and 1, while the upper diagonal elements are set to 0. To account for the variance absorbed by the latent variables, the variance of the residuals $\varepsilon_{ij}$ needs to be adjusted to ensure that the total variance is equal to one. We therefore calculate an adjusted variance $\sigma_i^2$ for each species $i$. Specifically, the formula for the variance of $\varepsilon_{ij}$ used above ensures that the overall variance of $u_{ij}$ remains at one, as in the probit version of the Dorazio-Royle multi-species occupancy model (alternatively, if this variance adjustment is not implemented in the model, a transformation is required on the estimated regression coefficients analog to the multivariate probit model below). After fitting the latent variable model, the full species correlation matrix $R$ can be derived from the correlation in the latent variables as $R = \theta \theta^T + diag(\sigma_1^2, \sigma_2^2, ...., \sigma_n^2)$. Hereafter, we refer to this multi-species occupancy model with residual correlation in occupancy specified via latent variables as 'the LV model.'

*Including species correlations with a multivariate probit model* – As a second variant of a JSDM with imperfect detection and species correlations, we extend the JSDM model proposed by Pollock et al. (2014) by adding a detection submodel. Here we follow the Bayesian implementation of the multivariate probit model proposed by McCulloch and Rossi (1994). We start with the same structure for the probit regression as above, but now we extend it to describe the residual correlations by means of a multivariate normal distribution:

$$z_{ij} = I(u_{ij} > 0),$$

$$u_{ij} = Xocc_j \beta occ^*_i + \varepsilon_{ij},$$

$$\varepsilon_{.j} \sim MVN(0, \sum^*),$$

where $\varepsilon_{.j} = (\varepsilon_{1j}, \ldots, \varepsilon_{nj})$. Here $\sum^*$ is a positive definite $n$ x $n$ covariance matrix with elements $\sigma = (\sigma_{11}, \sigma_{12}, \ldots, \sigma_{nn})$ defined by an inverse Wishart prior distribution with a $n$ x $n$ identity matrix as the

scale parameter and $n$+1 degrees of freedom. The detection model is identical to that in the LV model. In this model, the parameters $\beta occ^*$ and $\sum^*$ are not independently identifiable (McCulloch and Rossi 1994, Chib and Greenberg 1998). To obtain the correct correlation matrix $R$ and regression coefficients we need to calculate derived parameters as: $\beta occ = \beta occ^* C$ and $R = C \sum^* C^T$, where $C = diag(\sigma_{11}^{-1/2}, \sigma_{22}^{-1/2}, ...., \sigma_{nn}^{-1/2})$ (Chib and Greenberg 1998). Henceforward, we refer to the multi-species occupancy model with residual correlation in occupancy specified via a multivariate probit as 'the MP model.'

*Simulation studies*

To evaluate the performance of the LV and MP models with imperfect detection under a range of conditions, we conducted three simulation studies. For all simulations, the true occupancy status of each species ($z_{ij}$) was made a function of two random environmental covariates with values drawn from a Uniform(-1,1) distribution for each site. To simulate these environmental relationships, for each species we picked an intercept and values of the regression coefficients for each environmental covariate by sampling independently from a Normal(0,0.8) distribution. To induce the residual correlation in occupancy among species, in most simulations we generated a random, unstructured correlation matrix (see some exceptions under 'Simulation 1' below). We created the correlation matrix by selecting pairwise correlation coefficients from a Uniform(-1,1) distribution and then converting the resulting matrix to the nearest positive definite matrix using the *nearPD* function in the R-package Matrix (Bates and Maechler 2018). Based on this, we simulated correlated, binomial presence-absence data under the multivariate probit model as described above. To generate the observed detection/non-detection data $y_{ijk}$ we assumed three sampling occasions and a constant, species-specific detection probability (Simulations 1 and 2). We set these probabilities by randomly picking a value from a Uniform(0.1,0.7) distribution, representing the range from very elusive species to those that are

easy to detect. For Simulation 3, the detection probability was also made a function of covariates as outlined below.

We evaluated the estimation performance for each model by two means: (1) calculating the root mean square error (RMSE) between simulated (true) and estimated values (both for residual correlations and the regression coefficients in the occupancy sub-model) across all species; and (2) by using a linear regression of simulated vs estimated parameter values. The regression allows to detect a systematic bias (by inspecting the intercept) or an under- or overestimation of effect sizes (by inspecting the slope). For each simulation type, we considered a number of scenarios (e.g. with different number of species) and for each simulated scenario, we generated and analyzed 50 datasets.

We implemented all models in the BUGS language and fitted them in JAGS 4.3.0 (Plummer 2003) through R 3.4.2 (R Development Core Team 2015) (see code in Data S1). We ran the LV models drawing 15,000 MCMC samples with a burn-in of 10,000 samples and a thinning rate of 5 samples. We found the MP models to be computationally much more expensive; their convergence rates were much lower; hence, we ran them for 250,000 MCMC samples with a burn-in of 200,000 samples and a thinning rate of 50 samples. For all models, we ran three MCMC chains and assessed convergence visually and using the Brooks-Gelman-Rubin statistic (Gelman et al. 2014).

*Simulation 1: How many latent variables are required to estimate the correlation matrix?* – We asked how many latent variables are required for an accurate representation of the pairwise species correlation structure, given how we simulated the residual correlation structure. Previous studies (for models that ignore imperfect detection) suggested that as few as two to five latent variables might be sufficient (Warton et al. 2015). To address this question, we simulated data sets with communities of 10, 20 and 40 species and 1000 sites, and analyzed them with our new LV model with imperfect detection. For comparison, we also analyzed the data with 20 species using an LV model that did not account for imperfect detection. For each

dataset, we fitted models with an increasing number of latent variables (for 10 species: 2, 4, 6, 8, and 10; for 20 species: 2, 4, 8, 12, 16 and 20; for 40 species: 2, 5, 10, 15, 20, 25 and 30). We choose these simulation settings because they are informative, and a full factorial simulation design would have been prohibitively expensive in terms of computational demands.

While we chose to use unstructured, random correlation matrices for most of our simulations, there will be real-world cases where the residual correlations have a certain structure, be this due to missing environmental covariates, phylogeny, or guilds (Ovaskainen et al. 2017). We therefore also explored here how such a structure affects the required number of latent variables in the LV model. To simulate a structured correlation matrix, we drew random latent variables and derived from them a correlation matrix as indicated above in the LV model description. We simulated correlations with 2, 3 and 10 latent variables, the first two cases leading to highly structured correlation matrices and the last one resulting in an almost unstructured correlation matrix. We ran all of these additional simulations with 20 species and 1000 sites, again analyzing the data with an LV model with varying numbers of LVs.

*Simulation 2: Number of sites required* – We evaluated how the accuracy of the estimates of the occupancy parameters and the residual correlation matrix changed as we varied the size of the dataset (number of sites from 50 to 2,000 and the number of species from 10, 20 to 40). We analyzed all data sets with the LV and the MP models with imperfect detection, i.e., the two new models proposed in this paper. For comparison, we also analyzed these data sets with the corresponding LV model that did *not* account for imperfect detection, in order to gauge the bias incurred by ignoring unstructured imperfect detection. Based on the results from our first simulation study, we used 5 LVs for the simulations with 10 species, 10 LVs for those with 20 species, and 20 LVs for those with 40 species.

*Simulation 3: Can ignoring imperfect detection bias the correlation matrix estimates in traditional JSDMs?* It is sometimes assumed that ignoring imperfect detection in a JSDM only affects estimates of the occupancy intercept, but not those of coefficients of the environmental

variables nor, especially, of the residual correlation matrix. Our simulation 2 partly addresses this question. In simulation 3, we extend the assessment by simulating data where the detection probability was not only different across species but also affected by two spatial covariates that were independent of the occupancy covariates. We simulated a community of 20 species, and then fitted two JSDM with 10 LVs: one that did account for detection probability and modeled the detection covariates explicitly, and one that ignored detection probability (i.e. assumed perfect detection) and only modeled occupancy covariates. We compared the true and the estimated correlation matrices as well as occupancy parameter estimates between the two models.

*Case study: The Swiss passerine bird community*

We applied the LV model to the community of 79 passerine bird species detected in Switzerland during the surveys for the most recent Swiss breeding bird atlas (Knaus et al. 2018), where 2–3 surveys were conducted along irregular transects of typically 4–6 km length during one breeding season (15 April – 1 July) between 2012–2016 in a total of 2,318 randomly selected 1 $km^2$ quadrats. We expected species interactions to take place at the local scale of a territory, which for most passerines is on the order of one to a few hectares (see Kéry and Royle (2016, p. 279–282) for one group of passerines, the *Paridae* family). The comparatively large sampling area of 1 $km^2$ per site in the Swiss atlas might mask the consequences of species interactions on presence-absence patterns at the biologically relevant (local) scale. We therefore randomly picked one 1 ha quadrat within each 1 $km^2$ quadrat, provided it was covered by the survey transect. We excluded from the analysis 17 extremely rare species with detections in fewer than 10 quadrats, leaving 62 species in our analysis. Counts per surveyed hectare were reduced to binary detection/non-detection data prior to analysis, as our aim was to test our presence-absence models. To explain spatial variation in occupancy probability, we used linear and squared values of elevation, slope, northness (calculated as the cosine of aspect, which is equal

to 1 if the aspect is north and to -1 if the aspect is south) and forest cover. To explain spatiotemporal variation in detection probability, we used survey date and elevation and their interaction. As we modelled detectability as survey specific, we used a Bernoulli distribution formulation for the detection model instead of the Binomial that we used in the simulations. All covariates were standardized to a mean of zero and a standard deviation of one. We conducted 7 analyses of this dataset with the LV model (with 2, 5, 10, 15, 20, 25 and 30 latent variables) to determine the optimal number of latent variables for this dataset.

## Results

*Simulation 1: How many latent variables are required to estimate the correlation matrix?*

For the latent variable model with unstructured correlation matrices, we found that a low number of LVs resulted in poor estimates of the residual correlation matrix and that more LVs than usually recommended were required to obtain stable estimates. For 10 species, at least 5 LVs were necessary, while for communities of 20 species that number increased to 8–12 LVs and for 40 species to 15–20 LVs (Figure 1, Appendix S1: Figure S2 and S3). These findings held regardless of whether the model did or did not account for imperfect detection (Appendix S1: Figure S1). It appears that up to about $n/2$ LVs may be necessary to adequately approximate the residual correlation matrix when there are $n$ species in a community and the correlation matrix is unstructured. Increasing the number of LVs beyond $n/2$ yielded no improvement in the estimates and unnecessarily increased the complexity of the model, while extending run times considerably. In contrast to the correlation matrix, estimates of the occupancy parameters (regression intercept and coefficients) were accurate for all models and not affected by the number of LVs included in the model (Figure 1, Appendix S1: Figure S4).

For the simulations with structured correlation matrices, unsurprisingly, the best fitting model was the one where the number of LVs matched the number of LVs used to simulate the data (Figure 2, Appendix S1: Figure S5 and S6). Using a lower number of LVs resulted in loss of

accuracy of estimates and underestimation of correlation strength. Overfitting with a larger than necessary number of LVs also resulted in a reduction of accuracy, especially when the correlation matrix was highly structured (2 and 3 LVs). It is notable, though, that overfitting mainly resulted in an underestimation of correlation strength, but did not affect the correlation structure (high $R^2$ values, but slope smaller than one). The same effect can be seen for unstructured correlation matrices but is much less pronounced there (Appendix S1: Figure S2).

As the correlation structure is unknown for real-world datasets, we need a way to determine the optimal number of LVs to use in a model in order to avoid under or over-fitting. We found that the residual sum of squares RSS = $\sum_{i=1}^{n} \sigma_i^2$ across all species is a good indicator for accuracy. When plotting RSS against the number of LVs we can see that it rapidly declines until we reach the optimal number of LVs after which the decline is much slower (i.e., a so-called "elbow" in the trend; see Appendix S1: Figure S3 and S6). The above approach is very similar to the Cattell's scree test frequently used to determine the number of factors to retain in a principal components analysis (Cattell 1966).

*Simulation 2: How many sites do we need data from?*

For realistic ecological datasets simulated with imperfect detection, and analyzed with our LV model, we found that a large sample sizes were needed to accurately estimate species correlations and occupancy parameters. Patterns were fairly consistent across simulations, with the highest gains in accuracy observed up to 500 to 1000 sites, but still increasing with larger sample sizes (Figure 3, Appendix S1: Figure S7). Datasets with only 50 to 100 sites led to a drastic underestimation of species correlation strength and occupancy parameters (Appendix S1: Figure S7 and S8).

Ignoring detection probability in the LV model decreased the accuracy of parameters estimates and led to an underestimation of correlation strength as well as effect size of the occupancy parameters (Appendix S1: Figure S9, S10 and S11; for more results on the effect of ignoring imperfect detection see also Simulation 3 below).

Imperfect detection reduces the available data for each species, therefore increasing the number of sites required for accurate estimation. For datasets where detection is perfect or detection probabilities are high, accurate results can be obtained with a lower number of sites (Appendix S1: Figures S12 and S13). For example, the RMSE of the correlation matrix for 250 sites with perfect detection is comparable to the RMSE for 1000 sites with imperfect detection. Of course, these results depend on the specific detection probabilities.

For datasets with a low number of species ($n = 10$), the performance of the MP model was comparable to that of the LV model, although correlation strength was slightly underestimated (Figure 4, Appendix S1: Figure S14 and S15). For datasets with more than 10 species and a large number of sites, convergence of the model fitting MCMC sampling algorithm was hard to obtain even with long chains, resulting in inaccurate estimates of the correlation matrix and to a lesser degree the occupancy parameters (Figure 4, Appendix S1: Figure S14 and S15).

*Simulation 3: Can ignoring imperfect detection bias the correlation matrix estimates in traditional JSDMs?*

When detection probability was smaller than one and was affected by covariates, not accounting for detection probability (as in a traditional latent-variable model) led to reduced accuracy of the correlation estimates and an underestimation of correlation strength (Figure 5 left). It also led to poor estimates of occupancy parameters and in general to underestimation of occupancy (intercept) and the effect sizes of the covariates on occupancy (Figure 4, Appendix S1: Figure S16).

*Case study: Swiss passerine bird community*

The proportion of 1-ha-quadrats with observed occurrences among the 62 analyzed species ranged from 0.01 to 0.44 (mean = 0.07, median = 0.03). Graphing the sum of the residual

correlation against the number of LVs we determined that 20 LVs were adequate to describe the correlation structure in this dataset (Appendix S1: Figure S17). The residual correlation matrix for the entire community contained more positive than negative correlations (Figure 6). We inspected in more detail the estimates for one group of small, cavity-nesting species, the tits (family *Paridae*). The great tit (*Parus major*) was observed in 560 of the 2,318 sample quadrats, representing an apparent occupancy probability of 24.2%, followed by the coal tit (*Parus ater*; 18.9%), blue tit (*Cyanistes caeruleus*; 13.3 %), Crested tit (*Parus cristatus*; 4.7 %), Willow tit (*Parus montanus*; 4.6 %) and Marsh tit (*Parus palustris*; 4.2 %); see Appendix S1: Table S1.

Based on the limited nature of cavities suitable for nesting, we would have expected some negative residual correlations in their occurrence as previously found in temporal data for great and blue tits (Stenseth et al. 2015). However, quite to our surprise, we found only positive residual correlations in the occurrence probabilities among the six tit species, with values ranging from 0.03 to 0.54 under the LV model (Figure 7). The highest pairwise correlation was for great and blue tit, followed by coal and crested tit. Looking at the environmental correlation we found that habitat preferences for great, blue and marsh tits are similar, and so were habitat preferences for coal tit and crested tit. Habitat preferences of the willow tit were similar to coal and crested tit but very different from the other three species. Looking at the occupancy parameter estimates (Appendix S1: Table S2) and the correlation matrix (Appendix S1: Figure S18 and S19) it is clear that ignoring imperfect detection resulted in an underestimate of occupancy probability (i.e., the intercept) as well as correlation strength.

**Discussion**

Multi-species occupancy models were developed to describe species occurrence and community traits simultaneously by linking single-species models together in a hierarchical manner, but such models did not formally account for residual correlation between species (Dorazio and Royle 2005, Dorazio et al. 2006). In contrast, current joint species distribution models that include residual correlation (Warton et al. 2015) follow the same strategy of modeling species-

level regression coefficients as random effects, but they assume that all species are detected without error. With field data this latter assumption will rarely if ever be satisfied, not even for sessile organisms (Chen et al. 2013), as is often hoped, claimed or believed (Warton et al. 2015, Warton et al. 2016). Even in best-case scenarios of well-designed and highly standardized monitoring programs with surveys conducted by highly trained volunteers, as is the case with the Swiss breeding bird survey, detection for individual species varies between virtually 0 and 1 and is strongly dependent on the season and other factors (Kéry and Royle 2016: p. 706). Ignoring imperfect detection has the potential of biasing all inferences about species and communities in such community models.

In this paper, we extended two previously proposed models for analyzing correlated binary data that arise from multi-species presence-absence surveys: the multivariate probit model of Pollock et al. (2014) and the latent variable models of Hui et al. (2015), by adding a hierarchical level that describes the observation process. We tested and compared these two new models with simulated data of biological communities with species correlations. For small communities with fewer than about 10 species, we found that both models provided adequate estimates of species correlation and occupancy parameters, given a large enough sample size and an appropriate number of LVs in the LV model. For larger communities, however, the MP model showed poor convergence, very slow mixing and was often unable to accurately estimate parameters. Very long chains (e.g. 500,000-1,000,000 iterations) can estimates but result in extremely long run-times in the order of days or even weeks. This is not completely surprising given the large number of parameters in the correlation matrix in the MP model (e.g. in a community of size $n$=40, a total of $n(n\text{-}1)/2 = 780$ parameters would need to be estimated).

Interestingly, for the LV model we found that the number of LVs needed to adequately estimate the species correlation matrix was substantially larger than previously suggested (Letten et al. 2015, Warton et al. 2015), regardless of whether the model did or did not contain a detection component. As a rule of thumb, in our simulations with a completely unstructured correlation matrix, close to $n/2$ LVs appear to be needed until the estimates of the correlation

matrix stabilize, although the fact that such a large number of latent variables was needed is not overly surprising given the random nature in which we generated the residual correlation matrix. When the correlation matrix is highly structured, a much lower number of LVs adequately fits the data. The optimal number of LVs for a dataset can be found by running multiple models and plotting the sum or the residual variance against the number of LVs. In some cases, using a lower number of LVs could be useful, for example when the main goal is not to accurately estimating residual correlation, but simply to construct model-based ordinations (Hui et al. 2015, Warton et al. 2015).

While, based on the above results, the LV model did not necessarily reduce the number of parameters required compared to the MP model (for $n$=40 and LV=20 the number of parameters is 610), the MCMC sampling algorithm in JAGS was much more efficient for this model leading to quicker convergence and better mixing. We would therefore generally recommend the use of the LV model over that of the MP model, and encourage more research into choosing the number of LVs in situations where we expect the residual correlation matrix to exhibit more structure e.g., due to phylogeny.

A third formulation of a multi-species occupancy model with species correlations has been developed recently by Rota et al. (2016a). Their model is based on a multivariate Bernoulli model and can estimate and model the strength of species correlations as a function of covariates. However, this comes at the cost of an even larger number of parameters. It is unclear at present how well their approach would scale up to the large number of species found in many communities (Rota et al. used four species in their paper). The dependence of species correlations on the environment can also be evaluated with LVMs by modeling the latent variable coefficients as a function of environmental covariates (Tikhonov et al. 2017). It appears that comparative studies among these models would be valuable for practitioners, to help make a wise choice among these novel methods.

Our simulations provide a clear picture as to how ignoring imperfect detection biases inferences from a community model, and how this will depend on the magnitude of and on the patterns in detection errors. We saw in Simulation 2 that with constant detection probability ($p$) all the occupancy parameters as well as the estimates of the residual correlations were poorly estimated. Simulation 3 further emphasized the fact that the correlation matrix will be biased in a community model that is ignorant about detection if there are species-specific patterns in detection probability that are related to the habitat or other features of space (Appendix S1, Figure S16). Hence, accounting for imperfect detection and modeling the right covariates into the detection model may be and arguably often is required for unbiased inferences about species co-occurrence in biological communities.

Estimating correlations in species occurrence data is data hungry, requiring data from many sampling sites, even more so when detection is imperfect and detection probabilities and occupancy for some species are low. This is due to the large number of parameters that need to be estimated, but also due to the reduction in available information caused by the added uncertainty brought about by imperfect detection. As in any kind of capture-recapture type of model, the quality of parameter estimates increases with increasing detection probability (the "first rule of capture-recapture"; Kéry and Royle 2016). This suggests to us that when designing field studies to study species interactions it might be better to allocate survey effort to ensure relatively high detection probability than to increase the number of sites. This fits with design recommendations for single species models (Mackenzie and Royle 2005, Guillera-Arroita et al. 2010), but it would nevertheless be good to further investigate this with additional empirical research and simulations.

It is important to point out that while co-occurrence models can separate out environmental correlation that is explained by covariates that are included in the model from residual correlation, these models cannot tell us if the residual correlation is caused by missing environmental covariates or by true interactions among species, such as competition, predation or mutualism. As always, correlation is not the same as causation and the resulting correlations

need to be closely examined and interpreted in the light of what is known about species traits, phylogeny, trophic levels and other knowledge of the species and the ecosystem (Pollock et al. 2014, Morales-Castilla et al. 2015, Zurell et al. 2018). A good example for that is our Swiss passerine dataset: it is very unlikely that the high residual correlation in the occupancy probability between some species is caused by positive biological interactions (e.g., symbiosis or other forms of facilitation). On the contrary, we would have expected competition for some of these species and thus a negative correlation among closely related species like the tits, which use similar habitats and nest in a rare resource (cavities). It is much more likely that our model was missing important covariates that affect the distribution of several of the species simultaneously, leading to such positive residual correlations. For instance, our model did not contain the proportion of deciduous as opposed to coniferous trees in the forest. All six tits are basically woodland species, but great, blue, marsh and willow prefer deciduous, while crested and coal tit prefer mixed or coniferous woodland. The magnitude of estimated correlations is consistent with these effects and they may thus at least in part be explained by this missing covariate. Another possible explanation for the positive correlations is that they simply reflect spatial variability in the availability of cavities and all species are more abundant in areas with a higher density of cavities.

We highlight that this caveat, that a correlation does not equate to causation, must never be forgotten when modeling statistical correlations in the occurrence or abundance of a group of species. Put another way, we must to be very careful with assigning biological interactions to mere residual correlations found in an observational study. Another important point is that spatial scale affects co-occurrence patterns and the sampling scale needs to match the scale of the biological interactions (Zurell et al. 2018).

We implemented our models in the BUGS language and fitted them via JAGS software (Plummer 2003) due to its ease of implementation. While JAGS may not be quite as computationally efficient as other JSDM implementations (Wilkinson et al.), but it has the advantage of proven accessibility to very many ecologists. Furthermore, BUGS is essentially a

generic programming language for hierarchical models and it has a very large base of published code in ecology (e.g. McCarthy 2007, Royle and Dorazio 2008, Kéry and Schaub 2012, Kéry and Royle 2016). The BUGS language gives the ecologist user full flexibility to accommodate non-standard analyses or to integrate multiple and slightly different data sets in a single, integrated model. However, given the complexity of these models and the long run-times when the number of species increases, custom MCMC algorithms implemented in a fast programming language could significantly increase speed or possibly allow for better convergence of the MP model.

JSDMs with species correlations open up new possibilities to answer a wide range of questions that involve species interactions, e.g. how the distribution of predators is related to the distribution of prey species, which species are most affected by invasive species, and many others. The models could also be used to look at interaction among sexes or different age classes or could evaluate seasonal changes in co-occurrence and can potentially improve model predictions by accounting for unmeasured environmental variables or biotic interactions (Warton et al. 2015). We believe that our work will make these models even more useful now since accommodating for the universal fact that species are not detected perfectly will make their inferences more robust.

**Acknowledgments**

We thank J.Y. Barnagaud, A. Royle, B. Dorazio, R. Sollmann, E. Zipkin, D. Zurell, B. O'Hara and two anonymous reviewers for valuable comments that helped improve the manuscript. Mathias Tobler thanks J. Sheppard for the use of several workstations to run long simulations. Financial support of this study was provided by a grant from the Swiss National foundation (No31003A_1464125 to M. Kéry and M. Schaub). GGA was supported by the Australian Research Council (DE160100904). We thank the hundreds of dedicated volunteers who conducted the field work for the Swiss Breeding Bird Atlas.

**Literature cited**

Bates, D., and M. Maechler. 2018. Matrix: Sparse and Dense Matrix Classes and Methods. R package version 1.2-14.

Begon, M., C. R. Townsend, and J. L. Harper. 2006. Ecology from individuals to ecosystems. Fourth edition edition. Blackwell, Malden.

Beissinger, S. R., K. J. Iknayan, G. Guillera-Arroita, E. F. Zipkin, R. M. Dorazio, J. A. Royle, and M. Kéry. 2016. Incorporating Imperfect Detection into Joint Models of Communities: A response to Warton et al. Trends in Ecology & Evolution **31**:736-737.

Cattell, R. B. 1966. The scree test for the number of factors. Multivariate Behavioral Research **1**:245-276.

Chen, G., M. Kéry, M. Plattner, K. Ma, and B. Gardner. 2013. Imperfect detection is the rule rather than the exception in plant distribution studies. Journal of Ecology **101**:183-191.

Chib, S., and E. Greenberg. 1998. Analysis of multivariate probit models. Biometrika **85**:347-361.

Cody, M. L., and J. M. Diamond. 1975. Ecology and evolution of communities. Harvard University Press.

Dorazio, R. M., E. F. Connor, and R. A. Askins. 2015. Estimating the Effects of Habitat and Biological Interactions in an Avian Community. PLoS ONE **10**:e0135987.

Dorazio, R. M., M. Kéry, J. A. Royle, and M. Plattner. 2010. Models for inference in dynamic metacommunity systems. Ecology **91**:2466-2475.

Dorazio, R. M., and J. A. Royle. 2005. Estimating size and composition of biological communities by modeling the occurrence of species. Journal of the American Statistical Association **100**:389-398.

Dorazio, R. M., J. A. Royle, B. Soderstrom, and A. Glimskar. 2006. Estimating species richness and accumulation by modeling species occurrence and detectability. Ecology **87**:842-854.

Elith, J., and J. R. Leathwick. 2009. Species Distribution Models: Ecological Explanation and Prediction Across Space and Time. Annual review of ecology, evolution, and systematics **40**:677-697.

Gelfand, A. E., A. M. Schmidt, S. Wu, J. A. Silander, A. Latimer, and A. G. Rebelo. 2005. Modelling species diversity through species level hierarchical modelling. Journal of the Royal Statistical Society: Series C (Applied Statistics) **54**:1-20.

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin. 2014. Bayesian data analysis. Chapman & Hall/CRC Boca Raton, FL, USA.

Guillera-Arroita, G. 2017. Modelling of species distributions, range dynamics and communities under imperfect detection: advances, challenges and opportunities. Ecography **40**:281-295.

Guillera-Arroita, G., J. J. Lahoz-Monfort, D. I. MacKenzie, B. A. Wintle, and M. A. McCarthy. 2014. Ignoring Imperfect Detection in Biological Surveys Is Dangerous: A Response to 'Fitting and Interpreting Occupancy Models'. PLoS ONE **9**:e99571.

Guillera-Arroita, G., M. S. Ridout, and B. J. T. Morgan. 2010. Design of occupancy studies with imperfect detection. Methods in Ecology and Evolution **1**:131-139.

Hui, F. K. C., S. Taskinen, S. Pledger, S. D. Foster, and D. I. Warton. 2015. Model-based approaches to unconstrained ordination. Methods in Ecology and Evolution **6**:399-411.

Iknayan, K. J., M. W. Tingley, B. J. Furnas, and S. R. Beissinger. 2014. Detecting diversity: emerging methods to estimate species diversity. Trends in Ecology & Evolution **29**:97-106.

Kéry, M. 2011. Towards the modelling of true species distributions. Journal of Biogeography **38**:617-618.

Kéry, M., and J. A. Royle. 2008. Hierarchical Bayes estimation of species richness and occupancy in spatially replicated surveys. Journal of Applied Ecology **45**:589-598.

Kéry, M., and J. A. Royle. 2016. Applied Hierarchical Modeling in Ecology: Analysis of distribution, abundance and species richness in R and BUGS: Volume 1: Prelude and Static Models. Academic Press.

Kéry, M., and M. Schaub. 2012. Bayesian population analysis using WinBUGS a hierarchical perspective. Elsevier, Amsterdam.

Knaus, P., S. Antoniazza, S. Wechsler, J. Guélat, M. Kéry, N. Strebel, and T. Sattler. 2018. Swiss Breeding Bird Atlas 2013–2016. Distribution and population trends of birds in Switzerland and Liechtenstein. Swiss Ornithological Institute, Sempach, Switzerland.

Lahoz-Monfort, J. J., G. Guillera-Arroita, and B. A. Wintle. 2014. Imperfect detection impacts the performance of species distribution models. Global Ecology and Biogeography **23**:504-515.

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

Letten, A. D., D. A. Keith, M. G. Tozer, and F. K. C. Hui. 2015. Fine-scale hydrological niche differentiation through the lens of multi-species co-occurrence models. Journal of Ecology **103**:1264-1275.

MacKenzie, D. I. 2005. What are the issues with presence-absence data for wildlife managers? Journal of Wildlife Management **69**:849-860.

MacKenzie, D. I., L. L. Bailey, and J. D. Nichols. 2004. Investigating species co-occurrence patterns when species are detected imperfectly. Journal of Animal Ecology **73**:546-555.

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

MacKenzie, D. I., J. D. Nichols, J. A. Royle, K. H. Pollock, L. Bailey, and J. E. Hines. 2018. Occupancy estimation and modeling: inferring patterns and dynamics of species occurrence. 2nd edition. Elsevier.

Mackenzie, D. I., and J. A. Royle. 2005. Designing occupancy studies: general advice and allocating survey effort. Journal of Applied Ecology **42**:1105-1114.

McCarthy, M. A. 2007. Bayesian Methods for Ecology. Cambridge University Press, Cambridge.

McCulloch, R., and P. E. Rossi. 1994. An exact likelihood analysis of the multinomial probit model. Journal of Econometrics **64**:207-240.

Morales-Castilla, I., M. G. Matias, D. Gravel, and M. B. Araújo. 2015. Inferring biotic interactions from proxies. Trends in Ecology & Evolution **30**:347-356.

Morin, P. J. 2009. Community ecology. John Wiley & Sons.

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

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

Ovaskainen, O., G. Tikhonov, A. Norberg, F. Guillaume Blanchet, L. Duan, D. Dunson, T. Roslin, and N. Abrego. 2017. How to make more out of community data? A conceptual framework and its implementation as models and software. Ecology Letters **20**:561-576.

Plummer, M. 2003. JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling.*in* Proceedings of the 3rd International Workshop on Distributed Statistical Computing, Vienna, Austria.

Pollock, L. J., R. Tingley, W. K. Morris, N. Golding, R. B. O'Hara, K. M. Parris, P. A. Vesk, and M. A. McCarthy. 2014. Understanding co-occurrence by modelling species simultaneously with a Joint Species Distribution Model (JSDM). Methods in Ecology and Evolution **5**:397-406.

R Development Core Team. 2015. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Richmond, O. M. W., J. E. Hines, and S. R. Beissinger. 2010. Two-species occupancy models: a new parameterization applied to co-occurrence of secretive rails. Ecological Applications **20**:2036-2046.

Rota, C. T., M. A. R. Ferreira, R. W. Kays, T. D. Forrester, E. L. Kalies, W. J. McShea, A. W. Parsons, and J. J. Millspaugh. 2016a. A multispecies occupancy model for two or more interacting species. Methods in Ecology and Evolution **7**:1164-1173.

Rota, C. T., C. K. Wikle, R. W. Kays, T. D. Forrester, W. J. McShea, A. W. Parsons, and J. J. Millspaugh. 2016b. A two-species occupancy model accommodating simultaneous spatial and interspecific dependence. Ecology **97**:48-53.

Royle, J. A., and R. M. Dorazio. 2008. Hierarchical modeling and inference in ecology : the analysis of data from populations, metapopulations and communities. 1st edition. Academic Press, London.

Ruiz-Gutiérrez, V., and E. F. Zipkin. 2011. Detection biases yield misleading patterns of species persistence and colonization in fragmented landscapes. Ecosphere **2**:1-14.

Sollmann, R., M. M. Furtado, H. Hofer, A. T. A. Jácomo, N. M. Tôrres, and L. Silveira. 2012. Using occupancy models to investigate space partitioning between two sympatric large predators, the jaguar and puma in central Brazil. Mammalian Biology - Zeitschrift fur Saugetierkunde **77**:41-46.

Sollmann, R., B. Gardner, K. A. Williams, A. T. Gilbert, and R. R. Veit. 2015. A hierarchical distance sampling model to estimate abundance and covariate associations of species and communities. Methods in Ecology and Evolution:n/a-n/a.

Stenseth, N. C., J. M. Durant, M. S. Fowler, E. Matthysen, F. Adriaensen, N. Jonzén, K.-S. Chan, H. Liu, J. De Laet, B. C. Sheldon, M. E. Visser, and A. A. Dhondt. 2015. Testing for effects of climate change on competitive relationships and coexistence between two bird species. Proceedings of the Royal Society B: Biological Sciences **282**.

Tikhonov, G., N. Abrego, D. Dunson, and O. Ovaskainen. 2017. Using joint species distribution models for evaluating how species-to-species associations depend on the environmental context. Methods in Ecology and Evolution **8**:443-452.

Waddle, J. H., R. M. Dorazio, S. C. Walls, K. G. Rice, J. Beauchamp, M. J. Schuman, and F. J. Mazzotti. 2010. A new parameterization for estimating co-occurrence of interacting species. Ecological Applications **20**:1467-1475.

Warton, D. I., F. G. Blanchet, R. O'Hara, O. Ovaskainen, S. Taskinen, S. C. Walker, and F. K. C. Hui. 2016. Extending Joint Models in Community Ecology: A Response to Beissinger *et al.* Trends in Ecology & Evolution **31**:737-738.

Warton, D. I., F. G. Blanchet, R. B. O'Hara, O. Ovaskainen, S. Taskinen, S. C. Walker, and F. K. C. Hui. 2015. So Many Variables: Joint Modeling in Community Ecology. Trends in Ecology & Evolution **30**:766-779.

Wilkinson, D. P., N. Golding, G. Guillera-Arroita, R. Tingley, and M. A. McCarthy. A comparison of joint species distribution models for presence-absence data. Methods in Ecology and Evolution **0**.

Yamaura, Y., J. Andrew Royle, K. Kuboi, T. Tada, S. Ikeno, and S. i. Makino. 2011. Modelling community dynamics based on species-level abundance models from detection/nondetection data. Journal of Applied Ecology **48**:67-75.

Yamaura, Y., J. Royle, N. Shimada, S. Asanuma, T. Sato, H. Taki, and S. i. Makino. 2012. Biodiversity of man-made open habitats in an underused country: a class of multispecies abundance models for count data. Biodiversity and Conservation **21**:1365-1380.

Zipkin, E. F., A. DeWan, and A. J. Royle. 2009. Impacts of forest fragmentation on species richness: a hierarchical approach to community modelling. Journal of Applied Ecology **46**:815-822.

Zurell, D., J. Pollock Laura, and W. Thuiller. 2018. Do joint species distribution models reliably detect interspecific interactions from co-occurrence data in homogenous environments? Ecography **0**.

**Figure legends**

Figure 1: Effects of the number of species and latent variables on the accuracy of the species residual correlation matrix and occupancy parameters when estimated with our new latent variable (LV) multi-species co-occurrence model with imperfect detection (simulated data with an unstructured correlation matrix). Figures show the root mean square error (RMSE) and the grey polygon indicates the 95% confidence interval across 50 simulations.

Figure 2: Effects of number of latent variables on the accuracy of the species residual correlation matrix when estimated with a latent variable (LV) multi-species co-occurrence model without imperfect detection (simulated data for 20 species and 1000 sites). Data were simulated with different numbers of latent variables resulting in varying degrees of structure in the correlation matrix. The black line shows the mean and the and the grey polygons indicate the 95% confidence interval across 50 simulations.

Figure 3: Effects of the number of species and sites on the accuracy of the species residual correlation matrix and occupancy parameters when estimated with our new latent variable (LV) multi-species co-occurrence model with imperfect detection (simulated data with an unstructured correlation matrix). Figures show the root mean square error (RMSE) and the grey polygon indicates the 95% confidence interval across 50 simulations.

Figure 4: Effects of the number of species and sites on the accuracy of the species residual correlation matrix and occupancy parameters when estimated with our new multivariate probit (MP) multi-species co-occurrence model with imperfect detection (simulated data with an unstructured correlation matrix). Figures show the root mean square error (RMSE) and the grey polygon indicates the 95% confidence interval across 50 simulations.

Figure 5: Effects of ignoring imperfect detection on correlation and occupancy estimates by a latent variable (LV) multi-species co-occurrence model. We simulated data for 20 species and 1000 sites with two occupancy covariates and two detection covariates. For each sub-plot, the left side shows results from a LV model that accounts for imperfect detection and the right side shows results from a model ignoring imperfect detection.

Figure 6: Residual correlation matrix of occupancy probability in a community of 62 passerine species in Switzerland estimated under our new latent variable multi-species co-occurrence model with imperfect detection and 20 latent variables.

Figure 7: Residual correlation matrix of occupancy probability for the six tit species (family Paridae) in the community under our new latent variable multi-species co-occurrence model with imperfect detection and 20 latent variables (detail from Figure 6).