Received: 29 June 2016 | Revised: 6 February 2017 | Accepted: 8 February 2017

DOI 10.1002/env.2440

RESEARCH ARTICLE

Modeling joint abundance of multiple species using Dirichlet process mixtures

Devin S. Johnson | **Elizabeth H. Sinclair**

Alaska Fisheries Science Center, NOAA Fisheries, Seattle, WA 98115, USA

Correspondence
Devin S. Johnson, Alaska Fisheries Science Center, NOAA Fisheries, Seattle, WA 98115, USA.
Email: devin.johnson@noaa.gov

We present a method for modeling the distributions of multiple species simultaneously using Dirichlet process random effects to cluster species into guilds. Guilds are ecological groups of species that behave or react similarly to some environmental conditions. By modeling latent guild structure, we capture the cross-correlations in abundance or occurrence of species over surveys. In addition, ecological information about the community structure is obtained as a by-product of the model. By clustering species into similar functional groups, prediction uncertainty of community structure at additional sites is reduced over treating each species separately. The proposed model also presents an improvement over previously proposed joint species distribution models by reducing the number of parameters necessary to capture interspecies correlations and eliminating the need to have a priori information on the number of groups or a distance metric over species traits. The method is illustrated with a small simulation demonstration, as well as an analysis of a mesopelagic fish survey from the eastern Bering Sea near Alaska. The simulation data analysis shows that guild membership can be extracted as the differences between groups become larger and if guild differences are small, the model naturally collapses all the species into a small number of guilds, which increases predictive efficiency by reducing the number of parameters to that which is supported by the data.

KEYWORDS

abundance, Dirichlet process, joint species distribution model, multivariate, occurrence

1 | INTRODUCTION

In recent years, there has been considerable development of methodology for modeling and predicting abundance and occurrence of species of interest. Much of this development uses a hierarchical framework for developing models to fit the complexities of the observed data or natural abundance processes (Cressie, Calder, Clark, Hoef, & Wikle, 2009; Royle & Dorazio, 2008; Hobbs & Hooten, 2015). Some of these complexities may include spatial and temporal dependence (Carroll, Johnson, Dunk, & Zielinski, 2010; Latimer, Banerjee, Sang, Mosher, & Silander, 2009; Johnson, Ream, Towell, Williams, & Guerrero, 2013b; Thorson et al., (2015); Ward et al., (2010); Thorson et al., 2016), nondetection of individuals at sampled sites (Dorazio & Connor, 2014; Royle, 2004), and zero inflation (Johnson & Fritz, 2014; Thorson et al., 2016). Many of these species distribution models (SDMs) were used to make inference to a single species or one-at-a-time modeling if community inference was desired. However, by not recognizing the fact that species interact, use of single-species models for making inference for community abundance and structure can produce misleading results (Clark, Gelfand, Woodall, & Zhu, 2014). Hence, new joint species distribution models (JSDMs), which explicitly model species interactions (or cross-correlation), have recently been developed (e.g., Dorazio & Connor, 2014; Latimer et al., 2009; Thorson et al., 2015, 2016). Herein, we propose a novel JSDM approach, which models species interactions through membership in a latent ecological guild (Simberloff & Dayan, 1991) or functional group within the sampled range of habitats.

Typically, description of an abundance model begins with a generalized linear model (GLM) structure for the abundance process using a discrete value distribution such as

Poisson or negative binomial. For example, one might model the abundance as a Poisson observation with log mean that is a function of covariates. Those covariates might include habitat variables or variables related to the sampling procedure, which are thought to be related to the observed abundance. Alternatively, one might log transform the abundance and use Gaussian linear models (Johnson et al., 2013b; Johnson & Fritz, 2014; Ward et al., 2010), but the general mean structure is usually the same. Herein, we will focus on the GLM versions. The focus of the abundance modeling is related to either establishing an ecological relationship between (joint) abundance and the environmental covariates or predicting abundance at unsampled locations.

To extend the single-species GLM-oriented model to account for interactions of multiple species and improve prediction and inference of community structure and abundance, there have been several approaches that differ in the details of interaction modeling. Most were placed in the GLM framework by adding random effects, which are either directly correlated between species (Clark et al., 2014; Dorazio & Connor, 2014; Latimer et al., 2009) or when marginalized from the model (on the log scale) create a cross-species correlation structure (Thorson et al., 2015, 2016). The direct approach of using a free parameter for every pair of species when modeling the species-level correlation has been successfully implemented (Clark et al., 2014; Latimer et al., 2009); however, in those studies, there were a high number of sampled sites or a low number of species considered. In other studies, unstructured covariance did not produce reliable results (Dorazio & Connor, 2014). Thus, recent efforts to contribute novel methodology for JSDMs have focused on reducing the number of parameters used to model species interactions. Dorazio and Connor (2014) used a known species trait proximity matrix to model the species-level covariance matrix using a spatial correlation function. By using the known information on species similarity, there are only two parameters necessary to model the cross-correlation. Another low complexity approach has been proposed using linear combinations of latent random effects (Thorson et al., 2015, 2016). Specifically, the latent effects are spatial fields, but the same methodology could be applied using independent random effects. If the number of random effects is small relative to the number of species modeled, the number of parameters necessary for modeling species cross-correlation can be significantly reduced from the unstructured scenario.

As a novel alternative, we propose a fully Bayesian JSDM that uses latent ecological guilds to model interactions among species and obtain joint abundance inference. We also consider joint species occurrence as well, where occurrence is defined as the binary presence (i.e., abundance > 0) or absence (abundance = 0) of a species. Dorazio and Connor (2014) use known guild membership of different species to model independence of some species in a cross-correlated JSDM. Simberloff and Dayan (1991) defines an ecological guild to be “a group of species that exploit the same class of environmental resources in a similar way.” With this definition in mind, we seek to build a model where species are cross-correlated in abundance because there are unknown group effects for some set of covariates, that is, if the guild structure was known, they could be represented by (guild × covariate) interaction terms in the GLM abundance models. To accomplish this task, we format the model as a latent class or mixture model (see McLachlan & Peel, 2004). Mixture models or latent class models are often used to model dependance between variables in a nonparametric fashion because for a sufficiently large number of groups, marginalizing over the random latent classes can approximate any dependence structure to whatever degree desired (McLachlan & Peel, 2004; Vermunt, Van Ginkel, Der Ark, Andries, & Sijtsma, 2008). It has been shown that this holds even when the conditional models are independent given group membership (Dunson & Xing, 2009). In an ecological abundance context, finite mixture models have been used in the past to model spatial heterogeneity and correlation in a nonparametric fashion (Dorazio et al., 2008; Johnson et al., 2013b).

Dunstan, Foster, and Darnell (2011) and Dunstan, Foster, Hui, and Warton (2013) have proposed a finite mixture of GLMs approach (using the term “archetypes” as opposed to guilds) for JSDM. We build on their work by extending the finite mixture to an infinite mixture using a nonparametric Dirichlet process (DP) mixture model to account for environmental effects on guild abundance. By using the DP to model guild response to the environment, the number of guilds is a derived parameter in the model, which can be estimated. Dunstan et al. (2011) use information criteria (Bayesian information criterion [BIC]) to make inference to the number of guilds (archetypes). In a fully Bayesian framework, the use of BIC implies a uniform prior distribution for the number of guilds and this has been shown to overestimate the number of groups (Casella, Moreno, & Girón, 2014). By using an informative prior for the DP parameter, we can have control over the prior distribution of the number of guilds to avoid overfitting and make more accurate inference for guild membership.

In the following section, we propose a DP mixture model JSDM (DP-JSDM). The DP-JSDM is motivated using the Chinese restaurant process (CRP) for partitioning species into guilds. The CRP provides a method to construct a DP mixture model that, can, serve as a description of the DP process. There are several choices for modeling guild partitioning, but we utilize the DP/CRP due to its long history and good clustering properties (Casella et al. 2014). Parameter estimation in the DP-JSDM is challenging due to the latent guild process. We provide a reversible-jump Markov chain Monte Carlo (RJMCMC; Green, 2003) algorithm for making fully Bayesian inference, as opposed the maximum likelihood approach of Dunstan et al. (2011). Finally, we apply the method to few simulated data sets, as well as a real data set on mesopelagic fish communities in the eastern Bering Sea, Alaska.

2 | METHODS

2.1 | General model framework

We begin the description of the proposed methods with some notation. First, we assume that there are J surveys, for which abundance (or count index; hereafter, we use the term "counts") of I different species is measured. Let n_{ij} be the observed count for ith species in survey j. We also use the vector notation $\mathbf{n}_i = (n_{i1}, \ldots, n_{iJ})'$ and $\mathbf{n} = (\mathbf{n}_1', \ldots, \mathbf{n}_I')'$. For occurrence modeling, we denote occurrence as $y_{ij} = 1$ if $n_{ij} > 0$ otherwise $y_{ij} = 0$. In practice, n_{ij} need not necessarily be observed for occurrence modeling. The notations $\mathbf{y}_i$ and $\mathbf{y}$ are similar to the abundance counterparts.

2.1.1 | A JSDM for known guild membership

For abundance modeling, there are several possible distributions that could be used to model the observed discrete counts, Poisson, negative binomial, zero-inflated Poisson (ZIP), etc., so we will generically denote this observation model as $[n_{ij}|z_{ij}, \boldsymbol{\gamma}]$ where z_{ij} is a latent Gaussian variable controlling the level of expected abundance and $\boldsymbol{\gamma}$ is a vector of parameters. The notation "$[A|B]$" refers to the conditional probability distribution (density) function (PDF) of A given B. For example, if a Poisson distribution is considered, where $E(n_{ij}) = e^{z_{ij}}$, then $\boldsymbol{\gamma}$ is not necessary. In the example analysis of mesopelagic fish surveys, we utilize a ZIP model, so $\boldsymbol{\gamma}$ would represent the mixture probabilities for the extra zeros. For occurrence modeling $[y_{ij}|z_{ij}, \boldsymbol{\gamma}]$ would be a Bernoulli PDF. We use a probit link for computational ease, that is, probability of occurrence is $\Phi^{-1}(z_{ij})$, where Φ is a standard normal cumulative density function. Therefore, again, $\boldsymbol{\gamma}$ is not necessary.

To account for unknown interspecies correlations, we take a clustering approach inspired by the analysis of Johnson et al. (2013b) for incorporating spatial structure when there are no reasonable distance metrics or neighborhood groupings are unknown. First, if the species are unrelated in their environmental response, we might model the $\mathbf{z}_i = (z_{i1}, \ldots, z_{iJ})'$ vectors with the linear model

$$[\mathbf{z}_i|\boldsymbol{\delta}_i^*, \boldsymbol{\beta}, \sigma] = \mathrm{N}(\mathbf{X}\boldsymbol{\beta} + \mathbf{H}\boldsymbol{\delta}_i^*, \Sigma_i), \tag{1}$$

where

- $\mathbf{X}$ is a design matrix of covariates for which there are no species-level effects,
- $\boldsymbol{\beta}$ is a vector of regression coefficients (common to all species),
- $\mathbf{H}$ is a $J \times q$ matrix of q habitat or environmental covariates recorded at the jth survey.
- $\boldsymbol{\delta}_i^*$ are species-specific response (in terms of abundance or occurrence) coefficients to the environmental variables measured in $\mathbf{H}$, and
- Σ_i is a diagonal matrix with entries σ_{ij}^2 (for occurrence modeling $\sigma_{ij} = 1$).

Nonmixture JSDMs have proposed including species interactions by modeling $\boldsymbol{\delta}_i^*$ as random effects where $\mathrm{Cov}(\boldsymbol{\delta}_i^*, \boldsymbol{\delta}_{i'}^*) \neq \mathbf{0}$ to induce association between species abundance. The mixture JSDMs of Dunstan et al. (2011) and Dunstan et al. (2013), however, follow the view that if two species respond to environmental conditions in a similar way (i.e., belong to the same guild), then we would expect that $\boldsymbol{\delta}_i^* \approx \boldsymbol{\delta}_{i'}^*$. Thus, there are unique responses at the guild level, not the species level. We can fold any species-specific effects, which do not cluster into guilds into the $\mathbf{X}\boldsymbol{\beta}$ term.

Mixture formulations of JSDMs are constructed by envisioning an unknown partition, indexed by p, of the species into κ_p guilds (or archetypes) such that species within groups behave similarly with respect to the abundance process. That is, $\boldsymbol{\delta}_i^* = \boldsymbol{\delta}_{pk}$ for all species i belonging to guild k of partition p. To reduce notational burden, we will also use "p" to refer to the partition itself depending on the situation. For a given p, the joint model can be written as

$$[\mathbf{z}|p, \Delta_p, \boldsymbol{\beta}, \sigma] = \mathrm{N}(\mathbf{X}\boldsymbol{\beta} + \mathbf{K}_p\Delta_p, \Sigma), \tag{2}$$

where

- $\mathbf{K}_p = \mathbf{C}_p \otimes \mathbf{H}$, $\mathbf{C}_p$ is an $I \times \kappa_p$ binary matrix indicating which species belong to each guild in p ($\otimes$ is the Kronecker product),
- $\Delta_p = (\boldsymbol{\delta}_{p1}', \ldots, \boldsymbol{\delta}_{p\kappa_p}')'$ is a concatenated vector of unique guild coefficient vectors, here, we will assume that they are independent random effects such that $[\boldsymbol{\delta}_{pk}|\Omega] = N(\mathbf{0}, \Omega)$, for $k = 1, \ldots, \kappa_p$, and
- $\Sigma = \mathrm{blockDiag}(\Sigma_i)$.

To reduce the complexity of the proposed model, we suggest the following simplifications for general practice:

(i) setting $\boldsymbol{\sigma} = \mathrm{diag}(\Sigma^{1/2}) = \exp\{\mathbf{L}\boldsymbol{\theta}\}$, where $\mathbf{L}$ is a matrix of design covariates and
(ii) setting $\Omega = \omega^2(\mathbf{H}'\mathbf{H})^{-1}$, where $\omega = \exp(\xi)$.

With respect to (i), there are some useful special cases, namely, $\mathbf{L} = \mathbf{1}$ gives $\sigma_{ij} = \sigma$ and $\mathbf{L} = \mathbf{I}_I \otimes \mathbf{1}_J$ gives $\sigma_{ij} = \sigma_i$. However, the overdispersion parameters could also be modeled based on covariates associated with sampling methods, etc. Suggestion (ii) was formulated from the covariances of the g-prior (Tiao & Zellner, 1964). The g-prior, $N(\mathbf{0}, \omega^2(\mathbf{H}'\mathbf{H})^{-1})$, is an often used prior for regression coefficient parameters. It has the nice benefit that, with a single parameter, it automatically controls the scale of variance and covariance for each coefficient based on the scale of the covariates and their cross-correlation. The exponential reparameterization is used for ease of inference because ξ can be unconstrained as opposed to ω.

2.1.2 | Partitioning species into guilds

The previous description assumed that the correct partitioning of the species is known; however, for most real data sets, the correct partition is unknown. Thus, we must also provide a probability model for the number and membership of guilds. A commonly used distribution over partitions is the CRP. A construction definition of the CRP is described as follows, for a given parameter $\alpha > 0$:

1. A customer enters the restaurant and sits at one of an infinite number of tables.
2. The next customer enters and chooses to sit next to the previous customer with probability $1/(1+\alpha)$ or a new table with probability $\alpha/(1+\alpha)$.
3. In general, the $i+1$ customer chooses to sit by (link with) one of the previous customers, each with probability $\propto 1$, or by themselves with probability $\propto \alpha$.
4. Groups are constructed by collecting the cliques of the mathematical graph formed by the links between customers. That is, groups are defined by all customers that are linked, possibly through other customers.

This formulation of the CRP may be slightly different than the traditional description. Namely, it is usually described by the ith customer choosing to sit at an occupied table with probability proportional to the number of occupants or a new table with probability proportional to α. Blei and Frazier (2011) have shown that these two definitions are equivalent, but computations used for inference can be more efficient under this formulation. The "individual links" version was also used by Johnson et al. (2013b) for clustering spatial abundance trends. The density function for the CRP cluster model is given by

$$[p|\alpha] = \text{CRP}(\alpha) \propto \frac{\Gamma(\alpha)}{\Gamma(\alpha+I)}\alpha^{\kappa_p}\prod_{k=1}^{\kappa_p}(g_{pk}-1)!, \qquad (3)$$

where g_{pk} is the size of the kth group in partition p. Note that the PDF of partition p is only a function of the number and sizes of the groups. Partitions with the same number of groups and group sizes have the same probability regardless of which individuals fall in each cluster.

2.1.3 | An infinite mixture model

Given that we have added a CRP partition model where the number of guilds is not fixed and known, we should investigate what implications this has for the marginal species-specific effects, $\boldsymbol{\delta}_i^* = \sum_p\sum_k C_{pik}\boldsymbol{\delta}_{pk}[p|\alpha]$, where C_{pik} is the (i,k) entry of the $\mathbf{C}_p$ matrix, that is, an indicator that species i belongs to guild k. Using the well-known relationship between the CRP and the DP (Sethuraman, 1994), we can write the $\mathbf{z}$ portion of the model as the DP mixture model,

$$\begin{aligned} [\mathbf{z}_i|\boldsymbol{\delta}_i^*, \boldsymbol{\beta}, \sigma] &= \text{N}(\mathbf{X}\boldsymbol{\beta} + \mathbf{H}\boldsymbol{\delta}_i^*, \Sigma_i), \\ [\boldsymbol{\delta}_i^*|\text{G}] = \text{G} &= \sum_{k=1}^{\infty}\pi_k\mathcal{I}(\boldsymbol{\delta}_i^* = \boldsymbol{\delta}_k), \\ \text{G} \sim \text{DP}(\alpha, \text{G}_0); &\quad \text{G}_0 = \text{N}(\mathbf{0}, \Omega), \end{aligned} \qquad (4)$$

where $DP(\alpha, G_0)$ represents a DP with parameter α (same as the CRP) and base distribution, G_0, which in this case is $N(\mathbf{0}, \Omega)$. Drawing a realization of the random PDF, G, from $DP(\alpha, G_0)$ is accomplished by selecting the random sequence of probabilities, π_k by drawing $v_k \sim \text{beta}(1, \alpha)$ and setting $\pi_k = v_k\prod_{k'=1}^{k-1}(1 - v_{k'})$, then drawing $\boldsymbol{\delta}_k \sim G_0 = N(\mathbf{0}, \Omega)$. As it happens, selecting guilds via the CRP is equivalent to assigning groups according to the probability distribution $\{\pi_k\}_{k=1}^{\infty}$ for each of the I species (Sethuraman, 1994). So, the species-specific effects, $\boldsymbol{\delta}_i^*$, arise from the infinite mixture distribution in Equation 4.

Using the infinite mixture representation in Equation 4, we can make some comparisons to the previous mixture models of Dunstan et al. (2011). In the Dunstan et al. (2011) framework, the number of guilds is a known component of the model, then model selection is performed using BIC to select the appropriate number of guilds. This is equivalent to specifying a flat prior distribution over the number guilds (Casella et al. 2014). This procedure has been known to positively bias the estimated number of groups in a mixture model, that is, BIC tends to select a model with too many groups (Casella et al. 2014). From the infinite mixture perspective, κ_p can be thought of as a derived quantity that is the random number of unique guilds obtained from I draws from the distribution $\{\pi_k\}_{k=1}^{\infty}$. Using a prior distribution for α, the implied distribution on the number of guilds can be adjusted as desired. In the following sections, we demonstrate selecting a gamma prior for α such that, approximately, $[\kappa_p] \propto 1/\kappa_p$.

2.1.4 | Species effects cross-correlation structure

Like the spatial covariance model use by Dorazio and Connor (2014), the DP-JSDM also marginally possesses generally positive cross-covariance structure. This makes intuitive sense as one is grouping similar species together or, if species are dissimilar, allowing them to be independent. The covariance structure of the DP-JSDM can be derived by forming an intercept random effect, $\boldsymbol{\eta} = \mathbf{K}_p\Delta_p$, such that $\mathbf{z} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\eta} + \boldsymbol{\epsilon}$, where $[\boldsymbol{\epsilon}] = N(\mathbf{0}, \Sigma)$. Then, conditioning on the cluster assignment, the covariance matrix of the random effect $\boldsymbol{\eta}$ is

$$\text{Var}(\boldsymbol{\eta}|p) = \mathbf{C}_p\mathbf{C}_p' \otimes \mathbf{H}\Omega\mathbf{H}', \qquad (5)$$

and the marginal variance is given by the mixture,

$$\text{Var}(\boldsymbol{\eta}) = \left\{\sum_p \mathbf{C}_p\mathbf{C}_p'[p|\alpha]\right\} \otimes \mathbf{H}\Omega\mathbf{H}' = \Psi \otimes \mathbf{H}\Omega\mathbf{H}', \quad (6)$$

where Ψ is a matrix with (i, i') entries equal to the probabilities that species i shares a guild with species i'. We term the Ψ

matrix to be the species proximity matrix due to the fact that it forms a distance, of sorts, in the guild space of the species. Although the covariance is never negative between any two species, it can be zero; thus, those species that occupy different guilds will have uncorrelated η random effects, that is, if $\psi_{ii'} \approx 0$, then $\text{Cov}(\eta_{ij}, \eta_{i'j}) \approx 0$.

It should be noted, however, that although the covariance of the $\boldsymbol{\eta}$ random effect is generally positive, that does not mean that there are only "positive" (or zero) relationships between species. The clustering is based on the relationship each species has with the chosen covariates. For example, one species may react positively along a covariate gradient ($\delta_i > 0$) and another reacts negatively along that same gradient ($\delta_{i'} < 0$); therefore, if a new site has a high level of this covariate, the first species will be predicted to be relatively abundant, whereas the other species abundance will be lower.

2.2 | Bayesian inference

Because of the hierarchical and variable dimensional nature of the parameter space of the DP-JSDM model, we employ a Bayesian approach via MCMC for model fitting and inference. The posterior distribution of interest is given by

$$\begin{aligned}\left[\mathbf{z}, p, \Delta_p, \boldsymbol{\beta}, \omega, \boldsymbol{\sigma}|\mathbf{n}\right] &\propto [\mathbf{n}|\mathbf{z}]\left[\mathbf{z}|\boldsymbol{\beta}, \Delta_p, \boldsymbol{\sigma}\right] \\ &\times \left[\Delta_p|\omega, p\right]\left[p|\alpha\right][\omega][\boldsymbol{\sigma}][\boldsymbol{\beta}][\alpha],\end{aligned} \tag{7}$$

where $[\omega]$, $[\boldsymbol{\sigma}]$, $[\boldsymbol{\beta}]$, and $[\alpha]$ are the prior distributions for the parameters.

2.2.1 | An RJMCMC algorithm

The most direct way to make inferences on the proposed hierarchical clustering model is through an RJMCMC algorithm (Green, 2003) to sample the posterior distribution of the parameters, number of guilds, and guild assignment. Here, we provide an overview of the RJMCMC, additional details of the sampler are given in Supplementary Material A.

In our description, we will assume the following prior distributions for the parameters:

$$\begin{aligned}&[\boldsymbol{\beta}] = \text{N}\left(\boldsymbol{\mu}_\beta, \Sigma_\beta\right), \left[\Delta_p|\omega, p\right] = \text{N}\left(\mathbf{0}, \mathbf{I}_{\kappa_p} \otimes \omega^2\mathbf{Q}\right), \\ &[\omega] = \text{HT}(\phi_\omega, d_\omega), [\sigma] = \text{HT}(\phi_\sigma, d_\sigma) \\ &[p|\alpha] = \text{CRP}(\alpha), \quad \text{and} \quad [\alpha] = \text{gamma}(a, b),\end{aligned}$$

where $\mathbf{I}_{\kappa_p}$ is an identity matrix of size κ_p, $\mathbf{Q}$ is a known positive-definite matrix, and $HT(\phi, d)$ represents a scaled half-t (t density truncated to $[0, \infty)$) PDF with scale parameter ϕ and d degrees of freedom. For most of these parameters, the priors can be adjusted to whatever distribution the user would like, the trade-off being a Metropolis-Hastings (MH) update instead of a Gibbs step (e.g., for $\boldsymbol{\beta}$) or no difference at all if the parameter has to be updated with an MH step to begin with (ω, σ, and α). However, the normal $[\Delta_p|\omega, p]$ prior is necessary to the proposed RJMCMC algorithm. Although, the known $\mathbf{Q}$ is not necessary. This is not as critical as it sounds as the marginal distribution is still a nonparametric DP process; we just require that the base distribution be a multivariate normal.

The majority of the proposed RJMCMC algorithm is a standard Metropolis-within-Gibbs sampler for a GLM-like model. Conditioned on a realization of p, all the other parameters can be updated with a traditional MH step or a Gibbs step. Hence, we do not focus on their updates here (see Supplementary Material A). However, to update p, the dimension of the Δ_p vector will potentially change, necessitating the transdimensional aspect of the RJMCMC. Naively, the transdimensional moves require a joint (p, Δ_p) proposal, which can be rejected often if one of those quantities is a bad fit for the current state of the remaining parameters even though the other is acceptable. Second, proposing new p such that the MCMC chain will mix well over the space of partitions is itself challenging. Because we are assuming that $[\mathbf{z}|\boldsymbol{\beta}, \Delta_p, \boldsymbol{\sigma}]$ and $[\Delta_p|\omega, p]$ are multivariate normal, the first problem can be handled with the partial analytic RJMCMC method proposed by Godsill (2001) and utilized by Johnson and Hoeting (2011) and Johnson et al. (2013b) in similar transdimensional MCMC applications. The partial analytic method allows proposal of a new model (p in this case) without jointly proposing the associated model-specific parameters (Δ_p) because they can be analytically marginalized. This is a special case of a collapsed Gibbs sampler (Van Dyk & Park, 2008).

To update p, we denote $\ell_i \in \{1, \ldots, i\}$, to be the link between the ith customer and the person with which they choose to sit. If $\ell_i = i$, then the customer has chosen to sit by themselves (i.e., at a new table). Now, p is updated by sequential sampling of all ℓ_i. The reason that an MCMC based on the links version of the CRP moves through partition space faster is that changing one ℓ_i might split a group apart or join two groups together. This provides a bigger jump in partition space than simply moving one individual to another group as ids the case with the traditional CRP MCMC (e.g., Neal, 1991). The marginal PDF used to sample each ℓ_i in the partial analytic RJMCMC is

$$\begin{aligned}[\ell_i|\mathbf{z}, \boldsymbol{\beta}, \boldsymbol{\sigma}, \omega, \alpha] &= \int [\mathbf{z}|\boldsymbol{\beta}, \Delta_p, \boldsymbol{\sigma}][\Delta_p|\omega, p][\ell_i|\alpha] \; d\Delta_p \\ &= [\mathbf{z}|\boldsymbol{\beta}, \boldsymbol{\sigma}, \omega, p][\ell_i|\alpha],\end{aligned} \tag{8}$$

where $[\ell_i|\alpha] \propto \alpha\mathcal{I}(\ell_i = i) + \mathcal{I}(\ell_i < i)$. We found the direct method for calculating $[\ell_i|\mathbf{z}, \boldsymbol{\beta}, \boldsymbol{\sigma}, \omega, \alpha]$ used by Johnson and Hoeting (2011) and Johnson et al. (2013b) too computationally intensive, so we used an alternative formulation based on the Laplace approximation (see Supplementary Material A) for Equation 8, which in this case is exact because the integrand is proportional to a normal PDF for Δ_p (Goutis & Casella, 1999).

2.2.2 | Derived parameters

There are several derived parameters that may be of interest for making desired ecological inference. First are predictions of community abundance at new locations or times. Second, one may be interested in the overall effect of the environmental covariates for a particular species, that is, $\boldsymbol{\delta}_i^*$. Finally, the associations between species may be of interest, so an estimate of the $\boldsymbol{\Psi}$ matrix would be desired. All of these desires are easily fulfilled because we are drawing a sample of the base parameters via an RJMCMC algorithm; therefore, all that is necessary is to calculate these derived parameters within the RJMCMC.

First, abundance can be predicted at additional sites (or times) by drawing $\mathbf{z}^a \sim \text{N}(\mathbf{X}^a\boldsymbol{\beta} + \mathbf{H}^a\Delta_p, \Sigma)$, where the "$a$" superscript denotes additional environmental conditions, site, etc. So, the design matrices $\mathbf{X}^a$ and $\mathbf{H}^a$ are populated with covariate values for the additional conditions. Then draw $\mathbf{n}^a \sim [\mathbf{n}^a|\mathbf{z}^a, \boldsymbol{\gamma}]$. After each update of p, we can calculate $\mathbf{C}_p$ based on the links between individuals. Then κ_p is just the number columns of $\mathbf{C}_p$ and one can monitor that quantity to obtain a sample from the posterior distribution of the number of guilds. The species-specific environmental effects can be calculated at each iteration as $\boldsymbol{\delta}_i^* = \sum_{k=1}^{\kappa_p} C_{pik}\boldsymbol{\delta}_k$. The posterior sample mean of $\mathbf{C}_p\mathbf{C}p'$ provides an estimate of of the guild proximity matrix $\boldsymbol{\Psi}$.

3 | A SIMULATION PROOF-OF-CONCEPT

To examine the ability of the DP-JSDM model to make inference to species interaction, as well as to make community abundance predictions, we tested the model and RJMCMC sampler with a small group of simulated data sets. In analyzing the simulated data, our objective was to assess whether the DP-JDSM model would, in practice, produce generally correct estimates of the guild structure. Second, would the DP-JSDM exhibit the expected behavior that as ω becomes small, the number of guilds (groups) estimated will go to one as the functional differences between the guilds (with respect to the variables in $\mathbf{H}$) becomes insignificant.

3.1 | Simulation and analysis

Data were simulated for $I = 20$ species, $J = 35$ samples, and $\kappa_p = 5$ groups. Six data sets were simulated corresponding to ω equal to 0.25, 0.5, 0.75, 1, 1.5, and 2. Although the true number of groups is always technically equal to 5, the practical differences between the groups tends to zero as ω becomes smaller. The group sizes were $g_{pk} = 7$, 5, 4, 3, and 1. Three environmental variables composing the guild design matrix $\mathbf{H}$ were generated from a standard normal distribution. In addition, a single survey effort variable, $\mathbf{x}$, was generated to adjust overall abundance. The global design matrix was set to $\mathbf{X} = [\mathbf{1}, \mathbf{x}, \mathbf{H}_x]$, where $\mathbf{H}_x = [\mathbf{H}'| \ldots |\mathbf{H}']'$, that is, $\mathbf{H}$ matrix is concatenated I times over species. Thus, Δ_p denotes guild differences from the overall global effect of the environmental variables, $\mathbf{H}$. In order to maintain identifiability, we imposed the constraint that $\sum_{k=1}^{\kappa_p} \boldsymbol{\delta}_k = \mathbf{0}$. The global coefficient was set to $\boldsymbol{\beta} = (2, 1, 0, -1, 0.5)'$, and each $\boldsymbol{\delta}_k$; $k = 1, \ldots, 5$, was drawn from $N(\mathbf{0}, \omega^2\mathbf{H}'\mathbf{H})$. In these simulations, all $\sigma_{ij} = 0$; therefore, $\mathbf{z} \equiv \mathbf{X}\boldsymbol{\beta} + \mathbf{K}_p\Delta_p$. However, a common σ was estimated in each analysis using a Poisson observation model, that is, $[n_{ij}|z_{ij}] = \text{Poisson}(e^{z_{ij}})$.

The prior distributions used were the same as specified in subsection 2.2, specifically,

- $[\boldsymbol{\beta}]$: $\boldsymbol{\mu}_{\beta} = (\hat{\mu}_0, 0, 0, 0)'$, and $\hat{\mu}_0$ is the log of the mean observed count and $\Sigma_{\beta} = 100(\mathbf{X}'\mathbf{X})^{-1}$.
- $[\omega]$: $\phi_{\omega} = 1$, and $d_{\omega} = 1$, which implies a half-Cauchy prior distribution.
- $[\sigma]$: $\phi_{\sigma} = 1$, and $d_{\sigma} \to \infty$, which implies a half-normal prior distribution.
- $[\alpha]$: $a = 0.258$, and $b = 0.038$.

The prior distribution parameters for the gamma PDF $[\alpha]$ were chosen based upon the method of Dorazio (2009) with one alteration. Dorazio (2009) used the method to choose a and b such that the prior distribution over the number of groups was approximately uniform, that is, $[\kappa_p] \approx 1/I$, $\kappa_p = 1, \ldots, I$. However, we agree with the philosophy of Casella et al. (2014) that a priori, we should prefer fewer groups; therefore, using the same optimization approach as Dorazio (2009), we chose a and b such that, approximately, $[\kappa_p] \propto 1/\kappa_p$. So, all else being equal, a smaller number of groups is a priori preferred.

For each of the six simulated data sets, we sampled the posterior distribution (Equation 7) using the RJMCMC algorithm detailed in Supplementary Material A. Each sample consisted of 50,000 iterations following a burn-in of 10,000 iterations. Convergence of the MCMC was informally assessed through repeated runs from different starting values with group size ranging from one to 10. All of the runs produced very similar results (accounting for Monte Carlo error) so, we felt confident that run length was sufficient. Code to run the RJMCMC for the DP-JSDM can be found in the `multAbund`[1] package for the `R` statistical environment (R Development Core Team, 2015), which contains the code to run the RJMCMC algorithm described in Supplementary Material A.

3.2 | Simulation results

As expected, when ω became small, the DP-JSDM model was not able to distinguish guild differences between the species and essentially estimated one single group (Figure 1; $\omega = 0.25$).

[1] Available from github at: https://github.com/dsjohnson/multAbund. The package can be installed from within an R session using the `devtools` package, but users need to be able to compile source code on their platform as the `multAbund` package uses `C++` code in its routines.

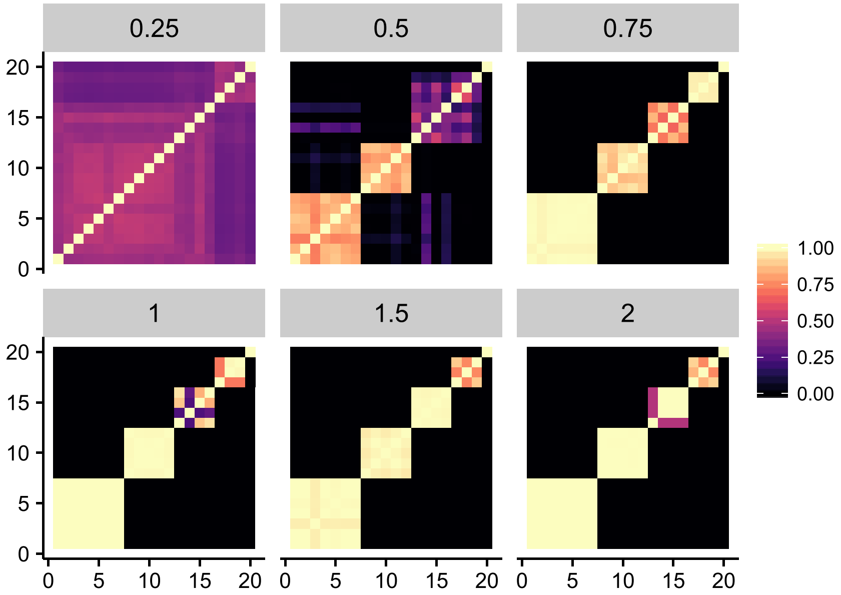

FIGURE 1 Estimated probabilities of joint guild membership between each species. For each panel, the value of ω used to simulated the data is provided in the bar above the plot

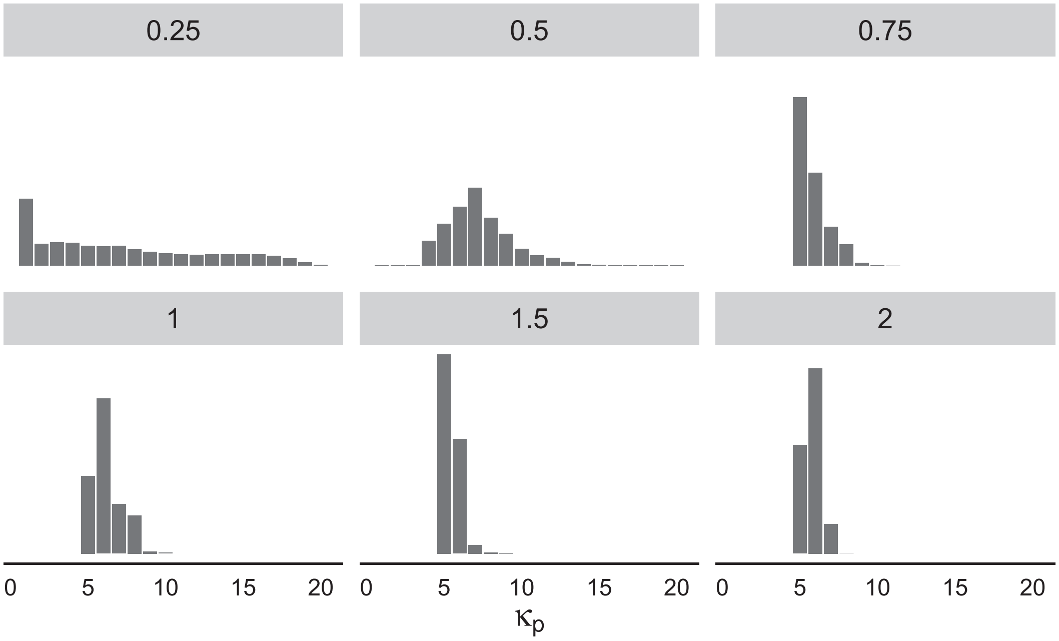

FIGURE 2 Estimated number of guilds, κ_p, for simulated Poisson data sets with ω ranging from 0.25 to 2. For each panel, the value of ω used to simulate the data is provided in the bar above the plot

As ω increased and guild differences became apparent, the model was able to separate the species into their respective guilds reasonably well (Figure 1). In addition, as ω became large, the precision with which the number of guilds was estimated increased as well (Figure 2).

There may be some bias as a few of the simulation runs produced $\hat{\kappa}_p = 6$ (Figure 2; $\omega = 1$ and 2); however, a full simulation experiment would be necessary to assess that fact. Even though we strived to create an efficient RJMCMC algorithm, it is still somewhat computationally intensive.

4 | EXAMPLE: MESOPELAGIC FISH ABUNDANCE

4.1 | Data

In our next demonstration of the DP-JSDM, we analyze community structure and abundance of fishes that migrate diurnally between three mesopelagic depths in the eastern Bering Sea near Alaska. Foster, Dunstan, Althaus, and Williams (2015) provide a similar analysis of fisheries trawl data from Australia using the finite-mixture MLE approach of Dunstan et al. (2011). In the Bering Sea, the tendency for

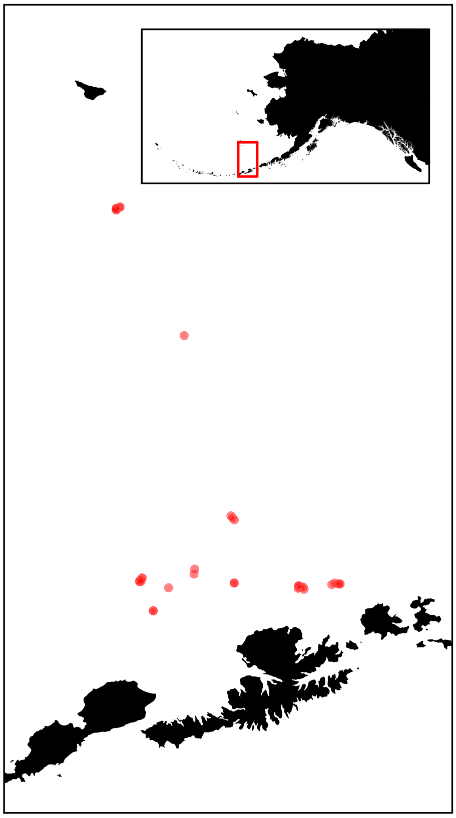

FIGURE 3 Locations of the mesopelagic trawl surveys. There were $J = 41$ separate trawl surveys used in the analysis of Section 4; however, some surveys were attempted geographically near other surveys, so they are somewhat obscured in the figure

most mesopelagic species to vertically migrate makes them an important trophic link between the deep scattering layer and upper surface waters (Sinclair, Walker, & Thomason, 2015), yet fundamental aspects of multispecies distributions and relative abundances have not been previously described.

The field effort identified highly productive areas of the eastern Bering Sea pelagic (Figure 3) for trawl sampling.

In the summers of 1999 and 2000, 29 daytime and 16 nighttime trawls were conducted at three depths (250, 500, and 1,000 m) during a narrow sampling period. Four of these trawls were not analyzed due to technical difficulties in the field, and we discarded them, resulting in $J = 41$ samples. Trawls were run at depth for 15–90 min resulting in collections of over 50,000 individuals representing 55 species of fish and squid. Essentially, each individual trawl sample represents a community as influenced by depth and time of day. Here, we will demonstrate the DP-JSDM using $I = 20$ of the relatively most common fish species (as opposed to squids, etc.). Many of the species were extremely rare in the survey effort (i.e., one individual observed over the entire study) and were removed. As opposed to many fisheries trawl surveys, all individuals caught in the trawl were classified to the species level. There was no subsampling of the individuals caught in the trawl.

The variables we put in the **H** design matrix reflect the belief that the species segregate into guilds based on diurnal vertical migration characteristics. So the guild covariates recorded for each trawl are daylight cycle (day or night) and depth category (250, 500, or 1,000 m). Here, we used the full interaction model to define the **H** design matrix (i.e., `'~ cycle*depth'` in `R` language model syntax). Because the duration of the trawl varied from survey to survey, the duration was included in the **X** matrix to model the overall abundance of fish caught in the trawl.

4.2 | Model and analysis

Initial attempts at fitting a DP-JSDM proceeded in the same manner as the analysis of the simulation data in the previous section. Namely, we used the same Poisson model for the observed abundance counts. However, after initial fittings, it became evident that the trawl data set possessed a significant level of zero inflation relative to the Poisson distribution. This is likely due to the spatial patchiness of pelagic fish occurrence distributions (Benoit-Bird & Au, 2003). In addition, there may also be detection issues in the survey such that a zero count in the trawl does not necessarily mean absence of the species. However, unlike Dorazio and Connor (2014), we do not have replicated surveys at the same site and time in which to separate detection and absence. Therefore, we utilized a ZIP model in place of a Poisson GLM. The ZIP model used for this analysis is

$$[n_{ij}|z_{ij},\gamma_i] = \gamma_i \mathcal{I}(n_{ij}=0) + (1-\gamma_i)\text{Poisson}(n_{ij}|e^{z_{ij}}), \qquad (9)$$

where γ_i is a species-specific zero inflation mixture that models the probability that a species is not present or not caught in the trawl if it is present. We used a *t* prior distribution on logit γ_i,

$$[\text{logit}\gamma_i] = \text{T}(\phi_\gamma, d_\gamma), \qquad (10)$$

with scale parameter $\phi_\gamma = 1.5$ and degrees of freedom $d_\gamma = 6$. This *t* distributed prior implies a prior distribution for γ_i that is approximately uniform over (0,1). For the remaining parameters, we used the same prior specification as the simulated data analysis of subsection 3.1.

To assess if there is any improvement gained by using the DP-JSDM, we also fitted the "independent species" JSDM, that is, $\kappa_p = I$, to the data. This independent JSDM did not truly treat each species independently because there are shared terms in the **X** design matrix (i.e., trawl duration), but it allows us to assess improvement in classifying animals into functional guilds relative to cycle and depth over treating them separately. To ascertain the magnitude of improvement, we would have liked to be able to use the "leave one out" Bayesian predictive information criterion (BPIC) given by

$$\begin{aligned} -2 \quad \text{BPIC} &= -2\sum_{i,j} E\{\log[n_{ij}|\mathbf{n}_{-(i,j)}, \mathbf{z}_{-(ij)}, \boldsymbol{\gamma}, \boldsymbol{\beta}, \Delta_p, p, \boldsymbol{\sigma}, \omega, \alpha]\} \\ &= -2\sum_{i,j} E\{\log[n_{ij}|\mathbf{n}_{-(i,j)}, \mathbf{z}_{-(ij)}, \boldsymbol{\gamma}]\}, \end{aligned} \qquad (11)$$

where $\mathbf{n}_{-(i,j)}$ is a vector of all observed data except n_{ij} and $\log[n_{ij}|\mathbf{n}_{-(i,j)}, \boldsymbol{\gamma}, \boldsymbol{\beta}, \Delta_p, p, \boldsymbol{\sigma}, \omega, \alpha]$ is the log posterior predictive density for the (i,j)th observation. However, it would be computationally infeasible to rerun the RJMCMC for every left out (i,j) entry. So we used the widely applicable information criterion (WAIC; Watanabe, 2013) as an approximation (Watanabe, 2010; Link & Sauer, 2016) to -2 BPIC, where

$$\begin{aligned} \text{WAIC} = &-2\sum_{i,j} E\{\log[n_{ij}|\mathbf{n},\mathbf{z},\boldsymbol{\gamma}]\} \\ &+2\sum_{i,j} Var\{\log[n_{ij}|\mathbf{n},\mathbf{z},\boldsymbol{\gamma}]\}. \end{aligned} \tag{12}$$

The WAIC requires only one run of the RJMCMC with the full data set. There are also other selection methods applicable, (Hooten & Hobbs, 2015); however, we found WAIC straightforward to implement for the DP-JSDM.

The model was fitted using the `R` package `multAbund`. The RJMCMC algorithm was run for 100,000 iterations following a burn-in of 10,000 iterations. Several trial runs at different starting values resulted in similar posterior distributions, so we felt that this run length was sufficient for posterior inference. The package contains code to fit the Poisson abundance data model as well as the ZIP and Bernoulli probit model for occurrence. In addition to the joint analysis of abundance, we also analyzed the trawl survey data as an occurrence data set where $y_{ij} = 1$ if $n_{ij} > 0$, else $y_{ij} = 0$. The occurrence analysis results are presented in Supplementary Material C.

4.3 | Results

After fitting the ZIP version of the DP-JSDM and the independent species JSDM, we noted there was a substantial improvement in WAIC under the DP-JSDM. WAIC for the DP-JSDM model was 3,052.071 and WAIC = 3,078.992 for the independence model. The posterior mode of the number

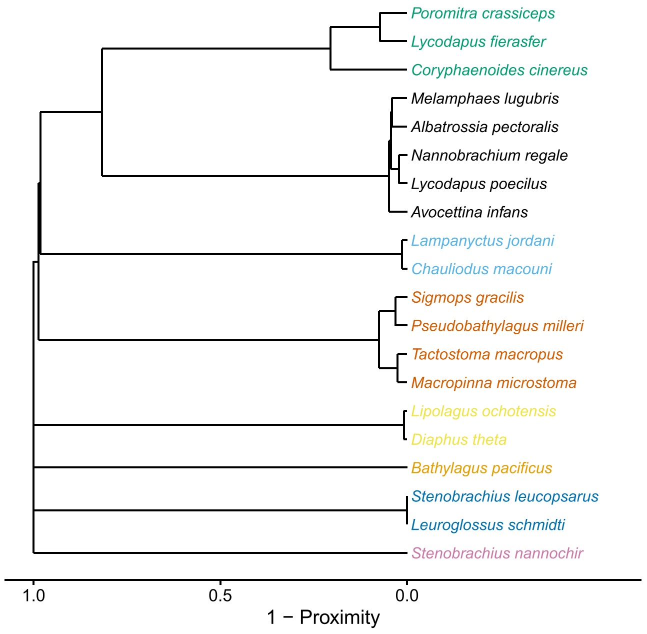

FIGURE 5 Clustering of trawl survey fish species based on the estimated probability of joint guild membership. The matrix $1 - \widehat{\Psi}$ was used as a distance matrix for forming the dendogram. The colored labels reflect guild groupings based on the posterior mode number of guilds, $\hat{\kappa}_p = 8$

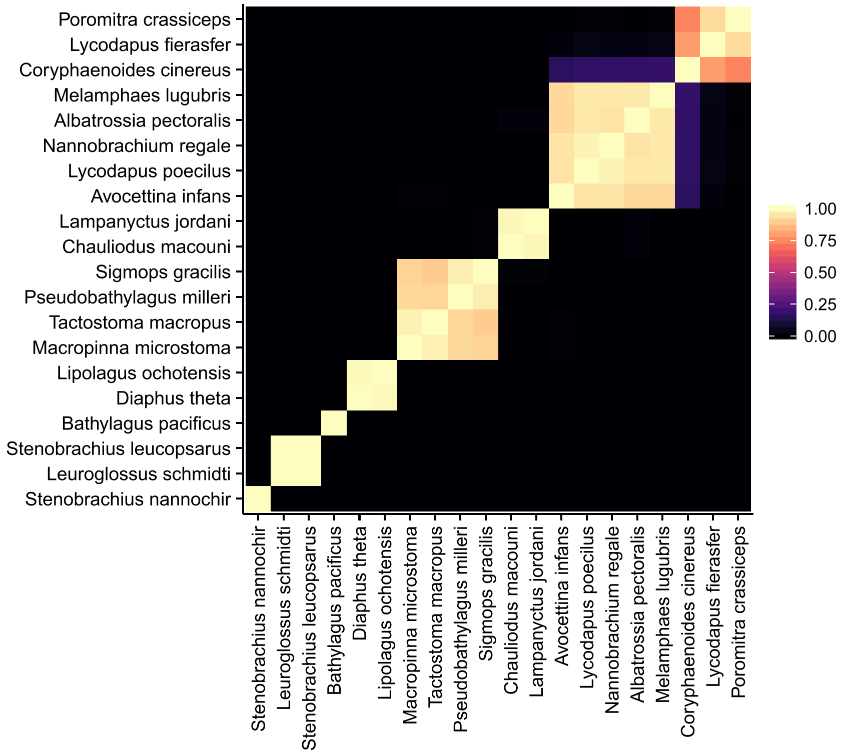

FIGURE 4 Estimated probability of joint guild membership (proximity), Ψ, for 20 of the fish species in the trawl survey with respect to abundance

of guilds was $\hat{\kappa}_p = 8$ with 95% of the posterior probability mass falling on $\kappa_p = 8$ or nine guilds. Figure 4 depicts the estimated posterior matrix, $\widehat{\Psi} = E[\mathbf{C}_p\mathbf{C}_p'|\mathbf{n}]$, which defines the probability that any two species share the same vertical migration guild.

Using $1 - \widehat{\Psi}$ as a measure of distance between species, we plotted the species according to the associated dendogram (Figure 5), which gives a better visualization of the groupings.

The predicted abundance for each species was calculated as $\hat{\mathbf{n}}^* = E[\mathbf{n}^*|\mathbf{n}]$ where $\mathbf{n}^* = (n_1^*, \dots, n_I^*)'$ is an observation under the various environmental conditions (Figure 6).

Results for the $\boldsymbol{\gamma}$ parameters are presented in Table B.1 of Supplementary Material B along with estimates of the $\boldsymbol{\delta}_i^*$ values (Figure B.1). Supplementary Material C provides similar figures and results for the DP-JSDM model using binary occurrence data instead of the observed abundance.

The model profiled a wide range in behavior among species from the two dominant mesopelagic fish families in the Bering Sea, Myctophidae and Bathylagidae. All but one of the eight guilds described by the model (Figures 5 and C.2) include a single species from one or both of these families, implying that they partition the water column based on a characteristic response to physical factors and foraging requirements.

The accuracy and predictive capability of the model were confirmed by the correct guild assignment of individual

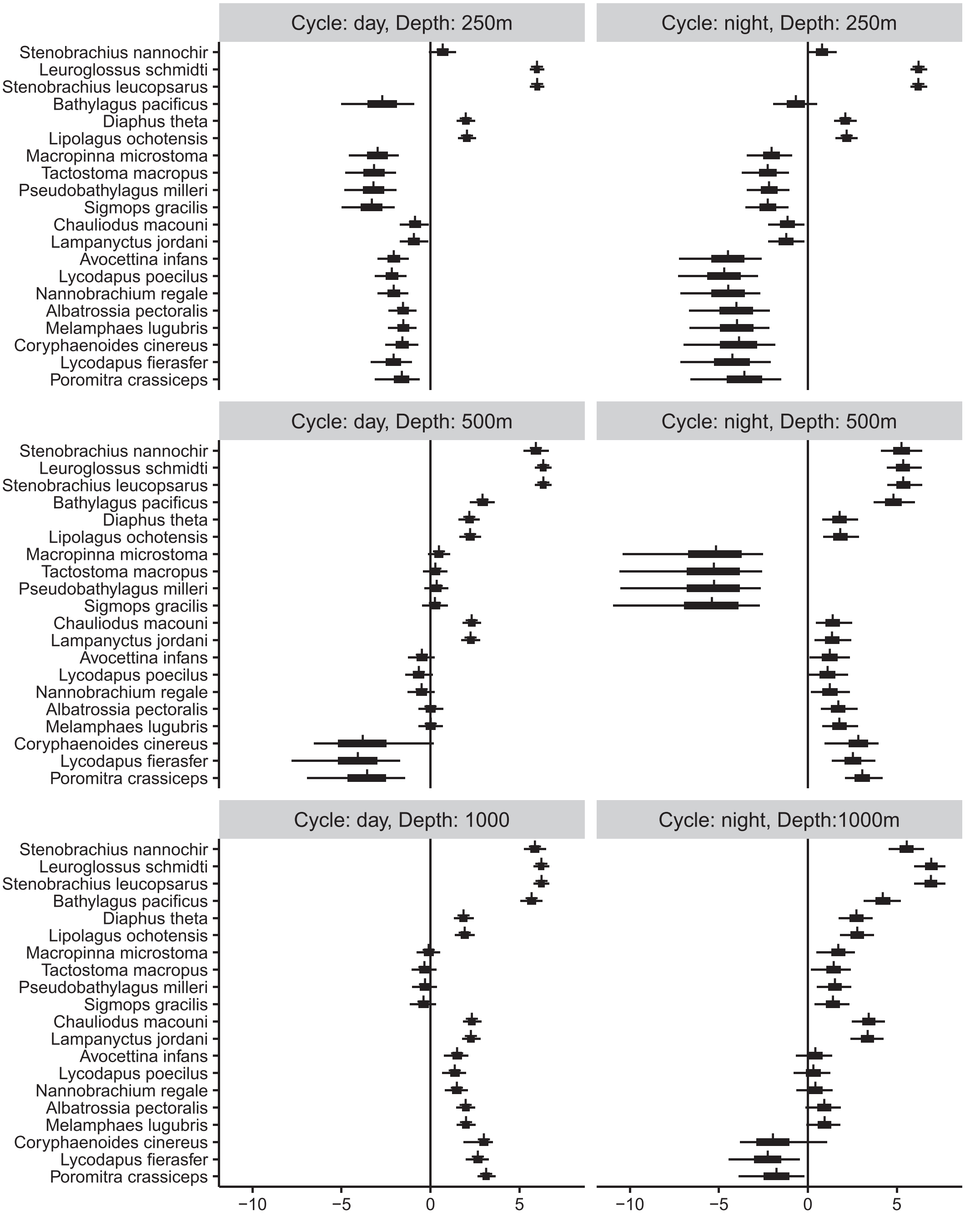

FIGURE 6 Species-specific predictions of log abundance for each level of cycle (day or night) and depth (250, 500, or 1,000 m)

species with previously known abundance and depth distribution profiles in the Bering Sea (e.g., bathylagids, *Leuroglossus schmidti* and *Lipolagus ochotensis*). Then by virtue of guild membership, the model described distribution patterns of species for which there is little reported data (e.g., myctophids, *Stenobrachias leucopsarus* and *Diaphus theta*). For instance, *L. schmidti* and *S. leucopsarus* formed the tightest cluster in both abundance and occurrence dendograms (Figures 5 and C.2). Each is the most abundant species within their respective families in the Bering Sea (Brodeur, Wilson, Walters, & Melnikov, 1999; Sinclair, Balanov, Kubodera, Radchenko, & Fedorets, 1999), and both were highly represented throughout the water column day and night in this study. Guild membership with *L. schmidti* suggests that *S. leucopsarus* shares a similar life history and foraging strategy wherein juveniles and adults have indistinct vertical migration and are stratified in the water column according to age (size) with adults remaining below 240 m (Beamish et al., 1999; Mecklenburg, Mecklenburg, & Thorsteinson, 2002).

The bathylagid *L. ochotensis* and myctophid *D. theta* also form a guild in abundance (Figure 5) along with *Stenobrachias nannochir* in occurrence guilds (Figure C.2). *L. ochotensis* and *S. nannochir* are among the most abundant mesopelagic species in the Bering Sea (Sinclair et al., 1999; Mecklenburg et al., 2002). Both are size stratified by depth with adults residing in the deepest layers and especially present between 500 and 1,000 m (Mecklenburg et al. 2002). As a strong vertical migrator, *L. ochotensis* is also abundant between 200 and 500 m (Sinclair et al., 1999; Mecklenburg et al., 2002). Little is known about *D. theta* from directed catch in the Bering Sea; however, guild identity with *S. nannochir* and especially with *L. ochotensis* suggests that they share similar patterns of behavior. The model implication that *D. theta* is an age stratified strong vertical migrator available at upper mesopelagic depths (Figure 6, B.1, and C.3) is supported by observations that it is a primary prey item of Dall's porpoise (*Phocoenoides dalli*) in the top 250 m of water column (Crawford, 1981).

The best example of the degree of fine detail captured by the model was demonstrated by *Bathylagus pacificus*, a common and abundant species of Bathylagidae that formed its own cluster (Figure 5). Like other members of its family, *B. pacificus* demonstrates a bimodal pattern in body size at depth (Peden, Ostermann, & Pozar, 1985; Mecklenburg et al., 2002). In our study, juvenile fish were concentrated at midlayer levels during the day (500 m) rising to 250 m at night, whereas adults concentrate at deeper daytime layers (1,000 m) rising to 500 m at night (Sinclair & Stabeno, 2002). This vertical migratory movement is apparent in the log abundance plots (Figure 6; and $\boldsymbol{\delta}_i^*$ values in Figure B.1) that together with known age distribution suggest *B. pacificus* may form its own guild based on abundances at depth driven by varying foraging requirements of juvenile and adults.

For some depth and cycle combinations, there are some guilds for which $\boldsymbol{\delta}_k$ values contain more variation that others, for example, the guild containing *Macropinna microstoma* contained more variability than others for the (night, 500 m) samples. This is due to the fact that these species are less abundant at those depths and times. This is also reflected in higher uncertainty in their γ_i estimates as well (Table B.1)

5 | DISCUSSION

We present a new methodology for modeling joint species distributions based on DP mixture random effects to model species associations through a latent guild structure. Instead of trying to directly parameterize cross-correlation in a species-specific random effect, we used latent membership in an ecological guild. Species belonging to the same guild followed the same response to environmental conditions through random coefficient effects in a GLM-like setting. Unlike simple cross-correlated species random intercepts, the DP-JSDM provides some valuable information on which species belong to guilds together and for the species within a guild, how they respond to the selected environmental conditions together.

A fundamental aspect of mesopelagic ecology in the Bering Sea is diel vertical migration. The DP-JSDM successfully identified community structure among 20 species of fish from the eastern Bering Sea within this framework. The selected model parameters of depth and light describe real-time clusters of species that move together similarly through the water column on a 24-hr cycle, presumably in relation to foraging. Based on studies conducted in the North Pacific Ocean, the diets of many of these same species collected from different depths match vertical distribution patterns of the variety of copepods and euphausiids that they consume (Beamish et al. 1999).

Although the DP-JSDM model was initially designed to model species association, it has the added benefit that it automatically adjusts to the necessary complexity because the number of guilds is also simultaneously being estimated as well. In the simulation experiment, it was demonstrated that if there is little difference between the species in their response to the recorded environmental conditions, the DP-JSDM will collapse to one guild, that is, no statistical difference between the species. This reduction in model complexity was noted by Johnson et al. (2013b) in reference to spatially clustering abundance trends. Here, we used only one model for guild membership, the CRP; however, as Casella et al. (2014) noted, there are other models for random set partitions. The CRP and similar models have rather straightforward requirements that each individual belongs to only one group and relationships are symmetric, that is, if individual i is linked with i', then i' is linked with i. However, there are models where group membership can be more complicated, mixed membership block models (MMBM; Airoldi, Blei, Fienberg, & Xing, 2008; Mohamed, Heller, & Ghahramani, 2014). MMBM have the ability to model asymmetric relationships and membership in multiple groups. Application of MMBM to JSDM is an area of future research that may help model more complex ecological patterns with respect to species associations.

In our description of the model and our examples, we have provided a relatively straightforward demonstration of the model and associated RJMCMC algorithm. However, there are several extensions that would be useful in other ecological settings. Here, we did not have repeated observations at each site, so we could not add an identifiable detection model to the observation process, although we illustrated that covariates (i.e., trawl duration) could be added as a quasi-detection model (Ver Hoef & Frost, 2003). However, if multiple observations are available for each site, then a detection process could be added to the observation model. Dorazio and Connor (2014) made use of an *N*-mixture model, and the DP-JSDM could use that as well. Instead of the ZIP model, one could add a another observation model,

$$[\tilde{n}_{ijk}, n_{ij}|...] = \text{Binomial}(\tilde{n}_{ijk}|n_{ij}, \gamma_{ijk})\text{Poisson}(n_{ij}|z_{ij}), \quad (13)$$

as the observation portion of the model, where $\tilde{n}_{ij}$ is the observed abundance of species i at site j during survey k and γ_{ijk} is the probability of each of the n_{ij} individuals being observed. If one marginalizes over the true abundances, the Poisson observation model results

$$[\tilde{n}_{ijk}|\gamma_{ijk}, z_{ij}] = \text{Poisson}(\tilde{n}_{ijk}|\log\gamma_{ijk} + z_{ij}), \quad (14)$$

where $E[n_{ijk}] = \exp\{\log\gamma_{ijk} + z_{ij}\}$. The same approach could also be used for occurrence modeling, in which case, it becomes occupancy modeling, that is, for the observed presence $\tilde{y}_{ijk}$, we use the hierarchical observation model,

$$[\tilde{y}_{ijk}, y_{ij}|...] = \text{Bernoulli}(\tilde{y}_{ijk}|y_{ij}\gamma_{ijk})\text{Bernoulli}(y_{ij}|z_{ij}), \quad (15)$$

where the probability that $\tilde{y}_{ijk} = 1$ is $y_{ij}\gamma_{ijk}$. The main point being that the process model does not change in either of these two settings, so the DP-JSDM can easily be adapted to these situations.

There is also an alteration that can be made when many sites are visited and spatial correlation between sites might also be a consideration. We are not calling this an extension, because spatial correlation can be added without making additions to the basic structure presented. All that needs to be changed to add random spatial effects is to use the basis function approach of Ver Hoef and Jansen (2014), Johnson, Conn, Hooten, Ray, and Pond (2013a), or Hefley et al. (2016). In a spatial basis function model, the random spatial field is modeled as $\boldsymbol{\eta} = \mathbf{H}\boldsymbol{\delta}$ where the columns of the matrix $\mathbf{H}$ contain the spatial basis functions evaluated at each of the modeled sites (rows). Each basis column represents a different frequency. In the notation just presented, it should be fairly obvious how the DP-JSDM can be changed to contain spatial correlation, one simply needs to use a basis function matrix for the environmental design matrix. In that case, it might be appropriate to use $[\boldsymbol{\delta}|\omega] = N(\mathbf{0}, \omega^2\mathbf{I})$ for the DP baseline distribution to match prior specifications that are usually used in spatial analysis. And, of course, one could combine the spatial model with the previously mentioned detection model extensions to form mutivariate spatial models for occupancy and abundance modeling.

ACKNOWLEDGMENTS

The authors thank M. B. Hooten and J. L. Laake for a critical reading of the original version of the paper. The findings and conclusions in the paper are those of the authors and do not necessarily represent the views of the National Marine Fisheries Service, NOAA. Reference to trade names does not imply endorsement by the National Marine Fisheries Service, NOAA.

REFERENCES

Airoldi, E. M., Blei, D. M., Fienberg, S. E., & Xing, E. P. (2008). Mixed membership stochastic blockmodels. *Journal of Machine Learning Research*, *9*(Sep), 1981–2014.

Beamish, R., Leask, K., Ivanov, O., Balanov, A., Orlov, A., & Sinclair, B. (1999). The ecology, distribution, and abundance of midwater fishes of the subarctic pacific gyres. *Progress in Oceanography*, *43*(2), 399–442.

Benoit-Bird, K. J., & Au, W. W. (2003). Spatial dynamics of a nearshore, micronekton sound-scattering layer. *ICES Journal of Marine Science: Journal du Conseil*, *60*(4), 899–913.

Blei, D. M., & Frazier, P. I. (2011). Distance dependent Chinese restaurant processes. *Journal of Machine Learning Research*, *12*, 2461–2488.

Brodeur, R. D., Wilson, M. T., Walters, G. E., & Melnikov, I. V. (1999). Forage fishes in the Bering Sea: Distribution, species associations, and biomass trends. In T. R. Loughlin & K. Ohtani (Eds.), *Dynamics of the Bering Sea*, (pp. 509–536). Fairbanks, Alaska: University of Alaska Sea Grant.

Carroll, C., Johnson, D. S., Dunk, J. R., & Zielinski, W. J. (2010). Hierarchical Bayesian spatial models for multispecies conservation planning and monitoring. *Conservation Biology*, *24*(6), 1538–1548.

Casella, G., Moreno, E., & Girón, F. J. (2014). Cluster analysis, model selection, and prior distributions on models. *Bayesian Analysis*, *9*, 613–658.

Clark, J. S., Gelfand, A. E., Woodall, C. W., & Zhu, K. (2014). More than the sum of the parts: Forest climate response from joint species distribution models. *Ecological Applications*, *24*(5), 990–999.

Crawford, T. W. (1981). *Vertebrate prey of Phocoenoides dalli,(Dall's porpoise): Associated with the Japanese high seas salmon fishery in the North Pacific Ocean*. Master's thesis: University of Washington.

Cressie, N., Calder, C. A., Clark, J. S., Hoef, J. M. V., & Wikle, C. K. (2009). Accounting for uncertainty in ecological analysis: The strengths and limitations of hierarchical statistical modeling. *Ecological Applications*, *19*(3), 553–570.

Dorazio, R. M. (2009). On selecting a prior for the precision parameter of Dirichlet process mixture models. *Journal of Statistical Planning and Inference*, *139*(9), 3384–3390.

Dorazio, R. M., & Connor, E. F. (2014). Estimating abundances of interacting species using morphological traits, foraging guilds, and habitat. *PloS one*, *9*(4), e94323.

Dorazio, R. M., Mukherjee, B., Zhang, L., Ghosh, M., Jelks, H. L., & Jordan, F. (2008). Modeling unobserved sources of heterogeneity in animal abundance using a Dirichlet process prior. *Biometrics*, *64*(2), 635–644.

Dunson, D. B., & Xing, C. (2009). Nonparametric Bayes modeling of multivariate categorical data. *Journal of the American Statistical Association*, *104*(487), 1042–1051.

Dunstan, P. K., Foster, S. D., & Darnell, R. (2011). Model based grouping of species across environmental gradients. *Ecological Modelling*, *222*(4), 955–963.

Dunstan, P. K., Foster, S. D., Hui, F. K., & Warton, D. I. (2013). Finite mixture of regression modeling for high-dimensional count and biomass data in ecology. *Journal of Agricultural Biological, and Environmental Statistics*, *18*(3), 357–375.

Foster, S. D., Dunstan, P. K., Althaus, F., & Williams, A. (2015). The cumulative effect of trawl fishing on a multispecies fish assemblage in south-eastern Australia. *Journal of Applied Ecology*, *52*(1), 129–139.

Godsill, S. (2001). On the relationship between Markov Chain Monte Carlo methods for model uncertainty. *Journal of Computational and Graphical Statistics*, *10*, 230–248.

Goutis, C., & Casella, G. (1999). Explaining the saddlepoint approximation. *The American Statistician*, *53*(3), 216–224.

Green, P. J. (2003). Trans-dimensional Markov Chain Monte Carlo. In Green, P. J., Hjort, N. L., & Richardson, S. (Eds.), *Highly Structured Stochastic Systems*. New York: Oxford University Press, Inc., pp. 179–196.

Hefley, T. J., Broms, K. M., Brost, B. M., Buderman, F. E., Kay, S. L., Scharf, H. R., ..., & Hooten, M. B. (2016). The basis function approach for modeling autocorrelation in ecological data. *Ecology*, *In press*.

Hobbs, N. T., & Hooten, M. B. (2015). *Bayesian models: A statistical primer for ecologists*. Princeton: Princeton University Press.

Hooten, M. B., & Hobbs, N. T. (2015). A guide to Bayesian model selection for ecologists. *Ecological Monographs*, *85*(1), 3–28.

Johnson, D. S., Conn, P. B., Hooten, M. B., Ray, J. C., & Pond, B. A. (2013a). Spatial occupancy models for large data sets. *Ecology*, *94*(4), 801–808.

Johnson, D. S., & Fritz, L. (2014). agtrend: A Bayesian approach for estimating trends of aggregated abundance. *Methods in Ecology and Evolution*, *5*, 1110–1115.

Johnson, D. S., & Hoeting, J. A. (2011). Bayesian multimodel inference for geostatistical regression models. *Plos One*, *6*(11), e25677.

Johnson, D. S., Ream, R. R., Towell, R. G., Williams, M. T., & Guerrero, J. D. L. (2013b). Bayesian clustering of animal abundance trends for inference and dimension reduction. *Journal of Agricultural Biological and Environmental Statistics*, *18*(3), 299–313.

Latimer, A., Banerjee, S., Sang, Jr, Mosher, E., & 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*(2), 144–154.

Link, W. A., & Sauer, J. R. (2016). Bayesian cross-validation for model evaluation and selection, with application to the North American Breeding Bird Survey. *Ecology*, *97*, 1746–1758.

McLachlan, G., & Peel, D. (2004). *Finite mixture models*. New York, NY: John Wiley & Sons.

Mecklenburg, C. W., Mecklenburg, T. A., & Thorsteinson, L. K. (2002). *Fishes of Alaska*. Bethesda, Maryland: American Fisheries Society.

Mohamed, S., Heller, K, & Ghahramani, Z (2014). Handbook of mixed membership models and their applications.

Neal, R. M. (1991). Bayesian mixture modeling by Monte Carlo simulation. (*Technical Report CRG-TR-91-2*): Department of Computer Science, University of Toronto.

Peden, A. E., Ostermann, W., & Pozar, L. J. (1985). *Fishes observed at Canadian Weathership Station PAPA (50° N, 145° W): with notes on the transpacific cruise of the CSS Endeavor*, Vol. *18*. British Columbia Provincial Museum.

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

Royle, J. A. (2004). N-mixture models for estimating population size from spatially replicated counts. *Biometrics*, *60*(1), 108–115.

Royle, J. A., & Dorazio, R. M. (2008). *Hierarchical modeling and inference in Ecology*. London, UK: Academic Press- Elsevier Ltd.

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica*, *4*, 639–650.

Simberloff, D., & Dayan, T. (1991). The guild concept and the structure of ecological communities. *Annual Review of Ecology and Systematics*, *22*, 115–143.

Sinclair, E., Balanov, A., Kubodera, T., Radchenko, V., & Fedorets, Y. A. (1999). Distribution and ecology of mesopelagic fishes and cephalopods. In Loughlin, T., & Ohtani, K. (Eds.), *Dynamics of the Bering Sea*. University of Alaska Fairbanks: Alaska Sea Grant College Program AK-SG-99-03, pp. 485–508.

Sinclair, E. H., & Stabeno, P. J. (2002). Mesopelagic nekton and associated physics of the southeastern Bering Sea. *Deep Sea Research Part II: Topical Studies in Oceanography*, *49*(26), 6127–6145.

Sinclair, E. H., Walker, W. A., & Thomason, J. R. (2015). Body size regression formulae, proximate composition and energy density of eastern Bering Sea mesopelagic fish and squid. *PloS one*, *10*(8), e0132289.

Thorson, J. T., Ianelli, J. N., Larsen, E. A., Ries, L., Scheuerell, M. D., Szuwalski, C., & Zipkin, E. F. (2016). Joint dynamic species distribution models: A tool for community ordination and spatio-temporal monitoring. *Global Ecology and Biogeography*, *25*, 1144–1158.

Thorson, J. T., Scheuerell, M. D., Shelton, A. O., See, K. E., Skaug, H. J., & Kristensen, K. (2015). Spatial factor analysis: A new tool for estimating joint species distributions and correlations in species range. *Methods in Ecology and Evolution*, *6*(6), 627–637.

Tiao, G. C., & Zellner, A. (1964). Bayes's theorem and the use of prior knowledge in regression analysis. *Biometrika*, *51*, 219–230.

Van Dyk, D. A., & Park, T. (2008). Partially collapsed Gibbs samplers: Theory and methods. *Journal of the American Statistical Association*, *103*(482), 790–796.

Ver Hoef, J. M., & Frost, K. J. (2003). A Bayesian hierarchical model for monitoring harbor seal changes in Prince William Sound, Alaska. *Environmental and Ecological Statistics*, *10*, 201–219.

Ver Hoef, J. M., & Jansen, J. K. (2014). Estimating abundance from counts in large data sets of irregularly-spaced plots using spatial basis functions. *Journal of Agricultural, Biological, and Environmental Statistics*, *20*, 1–27.

Vermunt, J. K., Van Ginkel, J. R., Der Ark, V., Andries, L., & Sijtsma, K. (2008). Multiple imputation of incomplete categorical data using latent class analysis. *Sociological Methodology*, *38*(1), 369–397.

Ward, E. J., Chirakkal, H., Gonzalez-Suarez, M., Aurioles-Gamboa, D., Holmes, E. E., & Gerber, L. (2010). Inferring spatial structure from time-series data: Using multivariate state-space models to detect metapopulation structure of California sea lions in the Gulf of California, Mexico. *Journal of Applied Ecology*, *47*(1), 47–56.

Watanabe, S. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *The Journal of Machine Learning Research*, *11*, 3571–3594.

Watanabe, S. (2013). A widely applicable Bayesian information criterion. *The Journal of Machine Learning Research*, *14*(1), 867–897.

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

How to cite this article: Johnson DS, Sinclair EH. Modeling joint abundance of multiple species using Dirichlet process mixtures. *Environmetrics*. 2017;28:e2440. https://doi.org/10.1002/env.2440