Joint dynamic species distribution models: a tool for community ordination and spatio-temporal monitoring: Joint...
Joint dynamic species distribution models: a tool for community ordination and spatio-temporal monitoring

James T. Thorson1*, James N. Ianelli2, Elise A. Larsen3, Leslie Ries4, Mark D. Scheuerell5, Cody Szuwalski6,7 and Elise F. Zipkin8,9

1Fisheries Resource Assessment and Monitoring Division, Northwest Fisheries Science Center, National Marine Fisheries Service, NOAA, Seattle, WA, USA, 2Resource Ecology and Fisheries Management Division, Alaska Fisheries Science Center, National Marine Fisheries Service, NOAA, Seattle, WA, USA, 3National Socio-environmental Synthesis Center, Annapolis, MD, USA, 4Department of Biology, Georgetown University, Washington, DC, USA, 5Fish Ecology Division, Northwest Fisheries Science Center, National Marine Fisheries Service, NOAA, Seattle, WA, USA, 6Sustainable Fisheries Group, Bren School of Environmental and Resource Management, University of California, Santa Barbara, CA, USA, 7Marine Science Institute, University of California, Santa Barbara, CA, USA, 8Department of Integrative Biology, Michigan State University, East Lansing, MI, USA, 9Ecology, Evolutionary Biology, and Behavior Program, Michigan State University, East Lansing, MI, USA

ABSTRACT

Aim Spatial analysis of the distribution and density of species is of continuing interest within theoretical and applied ecology. Species distribution models (SDMs) are being increasingly used to analyse count, presence–absence and presence-only data sets. There is a growing literature on dynamic SDMs (which incorporate temporal variation in species distribution), joint SDMs (which simultaneously analyse the correlated distribution of multiple species) and geostatistical models (which account for similarity between nearby sites caused by unobserved covariates). However, no previous study has combined all three attributes within a single framework.

Innovation We develop spatial dynamic factor analysis for use as a ‘joint, dynamic SDM’ (JDSDM), which uses geostatistical methods to account for spatial similarity when estimating one or more ‘factors’. Each factor evolves over time following a density-dependent (Gompertz) process, and the log-density of each species is approximated as a linear combination of different factors. We demonstrate a JDSDM using two multispecies case studies (an annual survey of bottom-associated species in the Bering Sea and a seasonal survey of butterfly density in the continental USA), and also provide our code publicly as an R package.

Main conclusions Case study applications show that JDSDMs can be used for species ordination, i.e. showing that dynamics for butterfly species within the same genus are significantly more correlated than for species from different genera. We also demonstrate how JDSDMs can rapidly identify dominant patterns in community dynamics, including the decline and recovery of several Bering Sea fishes since 2008, and the ‘flight curves’ typical of early or late-emerging butterflies. We conclude by suggesting future research that could incorporate phylogenetic relatedness or functional similarity, and propose that our approach could be used to monitor community dynamics at large spatial and temporal scales.

Keywords Bering Sea, dynamic factor analysis, flight curve, geostatistics, spatio-temporal model, species co-occurrence, species distribution model, species ordination.

INTRODUCTION

Understanding the spatial distribution of species is a central concern in ecology, and is necessary for evaluating spatial protections for threatened species, interpreting the impacts of abiotic and biotic factors on species distribution, tracking the spread of invasive species and interpreting the impacts of climate on species and communities (Hooten & Wikle, 2010; Rassweiler et al., 2014). Species distribution models (SDMs)
are increasingly used to approximate the distribution of a given species as a function of measured environmental variables and unobserved spatial covariation (Elith & Leathwick, 2009; Harris, 2015).

SDMs typically model either the distribution (presence–absence) or density of a given species as a function of measured environmental variables, while sometimes accounting for spatial covariation caused by geographical proximity. However, the relationship between abiotic habitat and species distribution may vary over time due to species invasions, climate shifts and many other factors (Pearman et al., 2008). Biotic interactions and physical access to new habitats (e.g., changes in dispersal caused by changing climate or the removal of dams on rivers) may also cause temporal changes in distribution (Soberón, 2007). This concern has increased interest in modelling temporal variation in species distribution via dynamic SDMs (DeVisser et al., 2010; Merow et al., 2011), where environmental linkages or residual spatial variation is allowed to vary over time.

By estimating the presence or density of species across space, SDMs are also a natural building block for understanding community dynamics. However, stacking estimates of the distribution or density for individual species from multiple SDMs and subsequently using these as summaries of community distribution is likely to yield biased estimates of community-level patterns (Clark et al., 2014, and references therein). Joint SDMs are an alternative procedure for simultaneously modelling the distribution of multiple species within a community, and were originally developed to estimate total species richness when accounting for imperfect species detection (Dorazio & Royle, 2005). Joint SDMs may also yield more accurate estimates of occupancy or density for rare or otherwise infrequently encountered species (Dorazio et al., 2006; Ovaskainen & Soininen, 2011; Rota et al., 2015; Thorson et al., 2015a). For these reasons, interest has increased in developing joint SDMs that include covariation in occurrence or density among species, either at the landscape or the sample level, while also accounting for standard habitat predictors (Latimer et al., 2009; Ovaskainen et al., 2010; Clark et al., 2014; Pollock et al., 2014; Thorson et al., 2015a).

Finally, species density is likely to be more similar on average for nearby locations than for geographically distant sites (termed ‘spatial autocorrelation’). Spatial autocorrelation in species density is likely to arise whenever an important but unobserved driver of species distribution is itself spatially autocorrelated (Dormann, 2007), or when dispersal patterns cause synchronous variation in density for nearby components of a population (Earn et al., 2000). Spatial autocorrelation can bias estimates of habitat associations, inflate Type-I errors for statistical tests of habitat associations (Bahn & McGill, 2007) and decrease explanatory power for SDMs (Dormann, 2007). Ecologists have a growing ability to account for spatial autocorrelation in species density via geo-statistical models and spatial statistics (Conn et al., 2014; Thorson et al., 2015b). In many cases, a simple linear or quadratic relationship between a habitat predictor variable and population density may be a poor approximation to the impact of habitat on density (Harris, 2015). In these cases, a spatio-temporal model can be thought of as being ‘semi-parametric’, where measured habitat variables are used to account for simple habitat relationships and the spatio-temporal component accounts for remaining nonlinear and unmeasured habitat associations (Shelton et al., 2014).

We identify a need for a joint dynamic species distribution model (JDSDM) that accounts for spatial autocorrelation in density for multiple species, as well as changes in spatial distributions over time. JDSDMs have previously been developed for site-level dynamics (e.g. Dorazio et al., 2010), but these models have not generally included spatial autocorrelation. We envision that a JDSDM could be used for two major projects in community ecology. First, a JDSDM could be used to assess the degree to which species have similar or different spatio-temporal patterns (‘species ordination’). For example, classifying species based on similar spatio-temporal dynamics could be used to determine whether high-density species are useful as indicators of abundance trends for rare or poorly surveyed species. Second, a JDSDM could be used for community monitoring to estimate dominant trends in community abundance and whether trends are similar across spatial areas (‘community monitoring’; see Dorazio et al., 2010). Community monitoring has increasing importance given that species distributions are likely to shift in unexpected ways due to human-caused climate change. Previous research suggests that jointly modelling density for multiple species is likely to be a more statistically efficient use of limited sampling data than single-species models (Dorazio et al., 2006; Latimer et al., 2009; Ovaskainen & Soininen, 2011; Ovaskainen et al., 2015; Thorson et al., 2015a), so community monitoring will likely be more efficient for detecting climate impacts than monitoring individual species.

We therefore develop a new spatial dynamic factor analysis (S DFA) model and propose its use as a JDSDM. S DFA estimates one or more factors, where each factor represents a variable that is not directly observed but which contributes to the distribution of each species in the community. S DFA then approximates the log-density of each species at any given site and time as a linear combination of those factors. Spatial variation in each factor is estimated as a Gaussian random field, and each factor evolves over time following a simple density-dependent process. In the following, we use two case studies to demonstrate the use of S DFA in species ordination and to identify large-scale patterns in community dynamics. The first involves 33 years of data for 10 numerically dominant, bottom-associated species in the Bering Sea. This case study shows that a JDSDM is able to distinguish broad-scale spatial partitioning in the fish and invertebrate community, and also captures recent distributional changes caused by an intrusion of cold waters (termed the ‘cold pool’). The second case-study involves 16 biweekly samples of 63 butterfly species in Ohio during 2010. This case study demonstrates that seasonal patterns in density can be
reconstructed for species with as few as 10 observations, while also showing that species within the same genus are significantly more correlated than otherwise unrelated species. We conclude with research recommendations that can lead to rapid, generic monitoring of community spatio-temporal dynamics and species similarity.

METHODS

A primer on dynamic factor analysis

Before describing SDFA, we first present a short overview of dynamic factor analysis (DFA). DFA is a dimension-reduction technique designed specifically for sequential observations of a variable (e.g. abundance) for different units of analysis (e.g. subpopulations of a species). Although ecologists may be more familiar with rank-reduction using principal components analysis or other multivariate techniques, these typically do not incorporate information about the time-series ordering of observations from each unit. There is a long history of using DFA in the analyses of multivariate time-series data in economics (Harvey, 1990) and medicine (Molenaar, 1985) but it is relatively new to ecology (Zuur et al., 2003).

The general motivation for DFA is as follows. If we had only a few time series (e.g. abundance for one of several subpopulations for a species), we might choose to model each of them as if they follow entirely independent dynamics over time. However, as the number of time series gets much larger, some of them will be likely show similar temporal patterns due to shared features of the environment (e.g. temperature) that tend to synchronize dynamics (e.g. similar population responses to an environmental driver). Therefore, we can model N time series as a linear combination of M latent variables using DFA, where N << M. These latent variables can be thought of as unmeasured or unknown environmental drivers of ecological processes. Recent applications of DFA have included identifying three dominant patterns in the abundance of 34 Pacific salmon stocks from North America and Asia (Stachura et al., 2014) and detecting important differences in the sensitivity of 80 boreal streams to variation in summer air temperature (List et al., 2013).

Introducing SDFA

We next develop a new SDFA model that provides a parsimonious approach to JDSDM. Here and throughout, we use JDSDM to refer to any model that includes spatial and temporal variation in occupancy or density of multiple species, and SDFA to refer to our specific hierarchical model for this JDSDM task. We also use the terms dynamic SDM and joint SDM, respectively, to refer to SDMs with temporal dynamics for a single species or multispecies models without temporal dynamics.

SDFA simultaneously fits to data for multiple species at different sites and time intervals using one or more latent spatial variables (termed ‘factors’ in our model, in analogy to DFA). Latent variable models are seeing increased use in community ecology as a sparse method for analysing correlations among multiple species (Ovaskainen et al., 2015; Thorson et al., 2015a; Warton et al., 2015). Each factor, \( \psi_j \), is a function that returns a value (associated with positive or negative density for one or more species) at any location \( s \) within the spatial domain \( s \in D \in \mathbb{R}^2 \), where \( D \) is the domain, indexed by latitude/longitude or any other appropriate two-dimensional measure), and any time interval \( t \) within the period of interest \( t \in [1, 2, \ldots, T] \), where \( T \) is the maximum time interval, see Table 1 for a definition of all parameters.

Data for each species \( p \) are assumed to arise as a function of log-density \( \theta_p \) for that species:

\[
c_p(s, t) \sim g(\exp(\theta_p(s, t)))
\]

where \( c_p(s, t) \) is data for species \( p \) at location \( s \) and time \( t \), and \( g \) is a measurement process (i.e. a Bernoulli distribution for presence–absence data, or a Poisson distribution for count data, etc.). The log-density \( \theta_p(s, t) \) at site \( s \) and time \( t \) of each species \( p \) is modelled as a linear combination of \( n_j \) factors:

\[
\theta_p(s, t) = \sum_{j=1}^{n_j} L_{pj, j} \psi_j(s, t) + \sum_{j=1}^{n_j} \gamma_{k, p} x_k(s, t)
\]

where the loadings matrix \( L \) represents the association \( L_{pj, j} \) between factor \( j \) and species \( p \), and \( \gamma_{k, p} \) is the linear effect of the \( k \)th of \( n_k \) covariates, \( x_k(s, t) \), for each site \( s \) and time \( t \) on log-expected counts for species \( p \). For the following case studies, we specify that \( x_k(s, t) = 1 \) for all observations and times (i.e. it represents an intercept governing differences in expected counts among species), but future studies could explore the impact of additional covariates on model performance. Given that we do not include any measured covariates, dynamic factors \( \psi_j(s, t) \) are used to capture all spatial and temporal variation in each species.

The \( j \)th dynamic factor follows a simple autocorrelated process over time:

\[
\psi_j(s, t+1) = \rho_j \psi_j(s, t) + \omega_j(s) + \xi_j(s, t)
\]

where \( \rho_j \) is the strength of autocorrelation for that factor (ranging from \(-1\) to \(1\), where \( \rho = 0 \) implies independent fluctuations around the expected log-density, and \( \rho = 1 \) implies a random walk with no tendency to revert to its long-term mean), \( \omega_j(s) \) represents spatial variation in density at site \( s \), and \( \xi_j(s, t) \) represents otherwise unexplained spatio-temporal variation in dynamics for factor \( j \). Each factor therefore follows a spatial Gompertz process (Thorson et al., 2015b), where \( \rho \) can be interpreted as the strength of density dependence. Spatial variation \( \omega_j \) is represented as a Gaussian random field:

\[
\omega_j \sim MVN(\mathbf{0}, \Sigma_w)
\]
where the covariance in average density $\Sigma_x$ is assumed to be higher for nearby locations than for geographically distant locations. We specifically model the relationship between geographical distance between locations $s$ and $s + h$ (i.e. where $|h|$ is the distance between locations) using a Matérn function:

$$\text{Var}(\omega_j(s), \omega_j(s + h)) = \frac{1}{2^{\nu+1}} \Gamma(\nu) \tau_{\nu}^2 (\kappa_\nu |h|^\nu) \exp(-\kappa_\nu |h|) = f(|h|, \kappa_\nu, \tau_{\nu}^2)$$

(5)

where $\kappa_\nu$ governs the range over which covariance declines as a function of distance $|h|$, $\nu$ governs the smoothness of the covariance matrix (we assume that $\nu = 1$ in the following) and $\tau_{\nu}^2$ governs the marginal variance of spatial variation. The covariance in spatio-temporal variation $\xi_j$ is defined similarly:

$$\xi_j \sim \text{MVN}(0, \Sigma_{\xi} \otimes I)$$

(6)

where covariance $\Sigma_{\xi}$ is calculated given range parameter $\kappa_\xi$, and precision $\tau_\xi$ and $I$ is an $n_t$ by $n_t$ identity matrix (where $n_t$ is the number of modelled time intervals).

Each factor $\psi_j$ is initialized given the assumption that $\psi_j$ starts away from its long-term stationary distribution:

$$\psi_j(s, 1) = \phi_j + \frac{\omega_j(s)}{1 - \rho_j} + \xi_j(s, t)$$

(7)

where $\omega_j(s)/(1 - \rho_j)$ is the median of the stationary distribution for this factor, and $\phi_j$ is the difference between the initial value and the median of its stationary distribution. This initial condition allows us to reparameterize dynamics for each factor in a way that is more computationally efficient.
Joint dynamic species distribution models

Pr(\(C=cp(s, t)\)) = \begin{cases} 
\exp \left( -v_{p,2}\lambda_p(s, t) \right) 
& \text{if } C=0 \\
\left( 1 - \exp \left( -v_{p,2}\lambda_p(s, t) \right) \right) \text{Gamma} \left( C ; v_{p,1}^2 \left[ 1 - \exp \left( -v_{p,2}\lambda_p(s, t) \right) \right] \varphi_{p,1}^2 \right) 
& \text{if } C > 0 
\end{cases} \tag{8a}

where Gamma(\(C, x, y\)) is the PDF of a gamma distribution with shape \(x\) and scale \(y\), \(v_{p,1}\) is the coefficient of variation for positive catches for species \(p\) and \(v_{p,2}\) controls the relationship between the probability of encountering zero individuals and predicted density for species \(p\), such that probability of not encountering a species \((C=0)\) is identical to a Poisson distribution with intensity \(v_{p,2}\lambda_p(s, t)\). We here use the same linear prediction \(\lambda_p(s, t)\) to control positive catch rates and the probability of encounter, but future research could explore more detailed models for the latter (Martin et al., 2005). We also envision future research exploring alternative sampling data including presence-absence or repeated-measures sampling (e.g. Yamaura et al., 2012).

Estimation and interpretation

Parameters for the generic spatial dynamic factor analysis model are not uniquely identifiable without further conditions. We therefore impose constraints on the form of the loadings matrix \(L\), and specify that the marginal variance of each factor \(\text{Var}(\theta_j)\) is one (Harvey, 1990; Zuur et al., 2003; Thorson et al., 2015a), as explained in Appendix S2. Given these two conditions, the covariance between log-density of two species (labelled \(p1\) and \(p2\)) at a given site and time can be calculated as:

\[
\text{Var}(\theta_{p1}(s, t), \theta_{p2}(s, t)) = \sum_j L_{p1,j} L_{p2,j}. \tag{9}
\]

This relationship has been noted previously for joint SDMs (Pollock et al., 2014; Thorson et al., 2015a; Warton et al., 2015), and can be generalized to calculate the covariance between species at different times and/or sites (see, e.g., Ovaskainen et al., 2016).

Given this model and constraints, all fixed effects can in theory be uniquely identified. In the following, the set of fixed effects includes density dependence \(\rho\) (we assume that \(\rho_j = \rho\)), initial condition \(\phi_p\) variance ratio \(\chi\) (used to calculate \(\tau_\phi\) and \(\tau_\chi\), and assumed to be constant among factors), range for spatial variation \(\kappa_{x0}\), range for spatio-temporal variation \(\kappa_{t0}\), measurement error parameters \(\nu\), and loadings matrix \(L\). We treat spatial variation \(\omega_j\) and spatio-temporal variation \(\xi_j\) (used to calculate \(\xi_j\)) as random effects for each factor \(j\). Finally, the lognormal Poisson distribution (used for applications involving count data) involves treating overdispersion \(\eta\) as a random effect.

We estimate fixed effects by identifying the value that maximizes the marginal likelihood function. This is accomplished using the following three steps:

1. We compute the joint likelihood of the data and random effects. The joint likelihood is calculated as the product of the probability of the data (equations 7a & 7b) and the probability of the dynamic factors (equation 3 and 5).

2. We then use the Laplace approximation (Skau et al., 2006) to approximate the marginal likelihood, obtained when integrating the joint likelihood with respect to random effects \((\omega_j, \xi_j, \eta)\).

3. Finally, we use a nonlinear optimizer to identify the values of fixed effects that maximizes the marginal likelihood function.

Steps 1 and 2 are implemented using Template Model Builder (TMB, Kristensen et al., in press; Kristensen, 2014), which computes gradients of the joint and marginal likelihoods using automatic differentiation techniques. Step 3 is done in the R statistical environment (R Core Team, 2014) using gradients provided by TMB in Step 2. After identifying the maximum likelihood estimates of fixed effects, we predict random effects (e.g. the value of factors \(\psi(s, t)\)) via empirical Bayes. To aid computational efficiency, we also use a stochastic partial differential equation approach to approximate the probability of Gaussian random fields (i.e. when computing the joint likelihood in Step 1 using equations 3 and 5; Lindgren et al., 2011; Thorson et al., 2015b). Parameter estimation is feasible on a laptop using one to six factors in a matter of hours. The time required to estimate parameters follows a close to linear increase with increasing sites, years and species, and a faster than linear increase with increasing number of factors. We hypothesize that future developments (i.e. regarding spatial approximations or parallel processing) will lead to increased computational speed.
Table 2 List of species (with common and scientific name) used in each case study, along with 'species number' used to reference species in later figures.

<table>
<thead>
<tr>
<th>Case study 1: interannual dynamics of Bering Sea demersal community</th>
<th>Case study 2: within-season dynamics of Ohio butterfly assemblage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Species number</td>
<td>Scientific name</td>
</tr>
<tr>
<td>1</td>
<td>Chionoecetes bairdi</td>
</tr>
<tr>
<td>2</td>
<td>Chionoecetes opilio</td>
</tr>
<tr>
<td>3</td>
<td>Gadus chalcogrammus</td>
</tr>
<tr>
<td>4</td>
<td>Gadus macrocephalus</td>
</tr>
<tr>
<td>5</td>
<td>Hippoglossoides elassodon</td>
</tr>
<tr>
<td>6</td>
<td>Hippoglossus stenolepis</td>
</tr>
<tr>
<td>7</td>
<td>Limanda aspera</td>
</tr>
<tr>
<td>8</td>
<td>Platichthys stellatus</td>
</tr>
<tr>
<td>9</td>
<td>Pleuronectes quadrirrucularus</td>
</tr>
<tr>
<td>10</td>
<td>Podotrichus accipenserinus</td>
</tr>
<tr>
<td>11</td>
<td>Colias philodice</td>
</tr>
<tr>
<td>12</td>
<td>Cyllipsis gemma</td>
</tr>
<tr>
<td>13</td>
<td>Danausplexippus</td>
</tr>
<tr>
<td>14</td>
<td>Enodia anthedon</td>
</tr>
<tr>
<td>15</td>
<td>Erynginus clarus</td>
</tr>
<tr>
<td>16</td>
<td>Erynginus bapiciniae</td>
</tr>
<tr>
<td>17</td>
<td>Erynginus brizo</td>
</tr>
<tr>
<td>18</td>
<td>Erynginus horatius</td>
</tr>
<tr>
<td>19</td>
<td>Erynginus juvenalis</td>
</tr>
<tr>
<td>20</td>
<td>Euphydryas phaeton</td>
</tr>
<tr>
<td>21</td>
<td>Euphyes vestris</td>
</tr>
<tr>
<td>22</td>
<td>Eurema lisa</td>
</tr>
<tr>
<td>23</td>
<td>Eurema nicippe</td>
</tr>
<tr>
<td>24</td>
<td>Eurytides marcellus</td>
</tr>
<tr>
<td>25</td>
<td>Evers comyns</td>
</tr>
<tr>
<td>26</td>
<td>Hermeuptychia sosoebius</td>
</tr>
<tr>
<td>27</td>
<td>Hylephila phyleus</td>
</tr>
<tr>
<td>28</td>
<td>Junonia coenia</td>
</tr>
<tr>
<td>29</td>
<td>Libytheana carinenta</td>
</tr>
<tr>
<td>30</td>
<td>Linenitis archippus</td>
</tr>
<tr>
<td>31</td>
<td>Linenitis arthemis astyanax</td>
</tr>
<tr>
<td>32</td>
<td>Lycaena hyllus</td>
</tr>
<tr>
<td>33</td>
<td>Lycaena phlaeas</td>
</tr>
<tr>
<td>34</td>
<td>Megisto cymela</td>
</tr>
<tr>
<td>35</td>
<td>Nymphalis antiopa</td>
</tr>
<tr>
<td>36</td>
<td>Papilio glaucus</td>
</tr>
<tr>
<td>37</td>
<td>Papilio polychrus</td>
</tr>
<tr>
<td>38</td>
<td>Papilio troilus</td>
</tr>
<tr>
<td>39</td>
<td>Phoebis sennae</td>
</tr>
<tr>
<td>40</td>
<td>Phoebis catullus</td>
</tr>
<tr>
<td>41</td>
<td>Phaidon sterh</td>
</tr>
<tr>
<td>42</td>
<td>Pieris rapae</td>
</tr>
<tr>
<td>43</td>
<td>Poanes hobomok</td>
</tr>
<tr>
<td>44</td>
<td>Poanes viator</td>
</tr>
<tr>
<td>45</td>
<td>Poanes zubalon</td>
</tr>
<tr>
<td>46</td>
<td>Polites mystic</td>
</tr>
</tbody>
</table>
Similar to conventional factor analysis, the resulting estimates of loadings \( L \) and factors \( w_j(n, t) \) can be rotated to ease interpretation. We therefore calculate a linear transformation matrix \( H \), and then interpret \( L^* = LH \) and \( \Psi^* = \Psi H^{-1} \). We specifically propose a new transformation matrix \( H \), which is designed to ensure that rotated Factor 1 explains the maximum proportion of total variance, Factor 2 explains the next highest proportion of total variance, and so forth (see Appendix S3). Hereafter, we refer to the loadings matrix and dynamic factors after transformation unless otherwise noted.

One advantage of SDFA is that it is estimated using maximum likelihood techniques, and therefore the number of factors (and other modelling decisions) can be informed using the Akaike information criterion (AIC) or other model selection techniques. In practice, however, we find that AIC favours the selection of a large number of factors (i.e. as many factors as species in some cases), where many of these factors explain a very small proportion of total variance. Therefore, model selection using AIC becomes impractical for a model involving many species. We use an alternative strategy for selecting the number of factors in the following case studies. Specifically, we estimate the SDFA model using one and two factors, and then successively increase the number of factors until the final factor explains less than 5% of total explained variance. This strategy is intended to allow

### TABLE 2. Continued

<table>
<thead>
<tr>
<th>Species number</th>
<th>Scientific name</th>
<th>Common name</th>
<th>No. of encounters ((n = 12,210) samples)</th>
<th>Scientific name</th>
<th>Common name</th>
<th>No. of encounters ((n = 1132) samples)</th>
</tr>
</thead>
<tbody>
<tr>
<td>47</td>
<td>Polites orogenes</td>
<td>Crossline skipper</td>
<td>25</td>
<td>Polites peckius</td>
<td>Peck’s skipper</td>
<td>263</td>
</tr>
<tr>
<td>48</td>
<td>Polites peckius</td>
<td>Peck’s skipper</td>
<td>263</td>
<td>Polites themistocles</td>
<td>Tawny-edged skipper</td>
<td>57</td>
</tr>
<tr>
<td>49</td>
<td>Polygonia comma</td>
<td>Eastern comma</td>
<td>201</td>
<td>Polygonia interrogationis</td>
<td>Question mark</td>
<td>255</td>
</tr>
<tr>
<td>50</td>
<td>Pompeius verna</td>
<td>Little glasswing</td>
<td>73</td>
<td>Pyrgus communis</td>
<td>Common checkerred-skipper</td>
<td>51</td>
</tr>
<tr>
<td>51</td>
<td>Satyroides appalachia</td>
<td>Appalachian brown</td>
<td>28</td>
<td>Satyroides eurydice</td>
<td>Eyed brown</td>
<td>11</td>
</tr>
<tr>
<td>52</td>
<td>Speyeria cybele</td>
<td>Great spangled fritillary</td>
<td>357</td>
<td>Strymon melinus</td>
<td>Gray hairstreak</td>
<td>101</td>
</tr>
<tr>
<td>53</td>
<td>Thymelicus lineola</td>
<td>European skipper</td>
<td>64</td>
<td>Vanessa atalanta</td>
<td>Red admiral</td>
<td>358</td>
</tr>
<tr>
<td>54</td>
<td>Vanessa cardui</td>
<td>Painted lady</td>
<td>56</td>
<td>Vanessa virginiensis</td>
<td>American lady</td>
<td>51</td>
</tr>
<tr>
<td>55</td>
<td>Wallengrenia egeremet</td>
<td>Northern broken dash</td>
<td>55</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
analysis of very large communities, while only adding factors that explain a biologically significant portion of community variance.

**Case study examples**

To demonstrate the usefulness of SDFA for estimating species similarity and identifying broad patterns in community spatio-temporal dynamics, we present results for two case study applications that differ substantially in taxa, number of species, spatial scale and temporal scale. The first case study demonstrates a smaller subset of important species at larger spatial and temporal scales, while the second demonstrates the analysis of communities with a large number of species at fine spatial and temporal scales. In each case study, we use a spherical projection based on latitude and longitude to compute distances among sites (Lindgren *et al.*, 2011).

**Example 1: Bering Sea demersal community**

Our first case study analyses data from annual bottom trawl surveys of the eastern Bering Sea during summer from 1982 to 2014. This survey was conducted by the Alaska Fisheries Science Center with minimal changes in survey design over this period, and consisted of about 375 bottom trawl tows per year. Each tow was on the bottom for about 30 min and data collection involved enumerating each species of finfish and invertebrate in the catch. This survey provides a synoptic picture of the demersal community of the Bering Sea. The Bering Sea is often classified into three biogeographical regions: the inner (0–50 m depth), middle (50–100 m depth) and outer (100–200 m depth) domains (Schumacher & Stabeno, 1998). Some species are distributed broadly throughout the Bering Sea (e.g. *Gadus chalcogrammus*), while others have a more restricted spatial distribution (e.g. *Chionoecetes opilio* in the middle and outer domains). Species distribution is additionally influenced by the intrusion of cold waters south from the northern border of the Bering Sea (termed the ‘cold pool’; Wyllie-Echeverria & Wooster, 1998). For computational speed, we aggregate all surveyed locations to 100 ‘sites’ distributed throughout the Bering Sea, where the locations of these sites are determined by a k-means algorithm. Model exploration suggests that increasing spatial resolution (i.e. by aggregating locations to a greater number of ‘sites’) has relatively little impact on parameter estimates. We also restrict analysis to the 10 most abundant species (by weight) observed in the demersal community (see Table 1) and, given that we analyse catches by weight, we use the delta-gamma distribution for measurement errors (equation 7b).

**Example 2: Ohio butterfly assemblage**

As a second case study, we analyse count data from 2010 regarding the distribution of butterflies at 58 sampled locations in Ohio. This state-wide butterfly monitoring programme, conducted by the Ohio Lepidopterist Society, consisted of a local volunteer at each location conducting repeated transect surveys using protocols based on those developed by Pollard (1977) and utilized across North American and European butterfly monitoring programmes. All adult butterflies present at each site were counted approximately weekly during the period from 1 April to 31 October (Julian dates 91–304), although not every location was surveyed every week. These counts have been previously analysed to demonstrate the ‘flight curve’ for each species, i.e. the pattern of increase and subsequent decrease in counts for butterflies during a species’ flight period (Cayton *et al.*, 2015). Changes in the flight curve among years have been used in a range of ecological studies, for example to identify phenological changes caused by climate change (Roy & Sparks, 2000). We analyse counts for all 63 butterfly species that were encountered 10 or more times during 2010 (Table 1) and use a lognormal-Poisson distribution for measurement errors (equation 7a). SDFA analyses observations within discrete time periods, $t \in \{1, 2, \ldots, T\}$, so we aggregate observations into 16 2-week ‘intervals’ from 1 April to 31 October.

**Simulation experiment**

We also conduct a simulation experiment intended to explore and validate the estimation properties of SDFA when confronted with small sample sizes. To do so, we simulate dynamics at 20 ‘sites’ over 20 years, where sites are randomly distributed within a $1 \times 1$ grid. Dynamics evolve following five factors, where each factor has a spatial scale of 0.25, a marginal standard deviation for spatial variation and spatio-temporal process error of 0.5, and where each factor is weakly density-dependent (i.e. $\rho = 0.8$) and starts in the first year with a small deviation away from its equilibrium (i.e. $\varphi_j$ is drawn from a standard normal distribution). Similarly, each element of the 5-by-5 loadings matrix $L$ is drawn from a standard normal distribution (i.e. each factor contributes variance drawn from a chi-squared distribution with five degrees of freedom). Each site is surveyed once per year, yielding one count of every species, site and year (2000 counts total). Counts arise from a lognormal Poisson distribution with a log-standard deviation of 0.1 (to account for overdispersion in the measurement process), and each species has an expected log-count of 1.0 (i.e. $\gamma_p = 1$). Code to replicate this simulation experiment or to implement SDFA for a different data set is provided as an R package on the first author’s website (https://github.com/James-Thorson/spatial-DFA).

**RESULTS**

**Case study 1: among-year dynamics of Bering Sea demersal community**

Analysing data from the Bering Sea demersal community, we find that the sixth factor adds less than 5% of total variance, so we proceed by interpreting the SDFA model with six factors. Given this model, SDFA illustrates three main groups (Fig. 1). The first involves *Gadus chalcogrammus* and *Gadus macrocephalus*, which have a high correlation (0.91). *Chionoecetes bairdi* and *Chionoecetes opilio* also have
highest correlation with one another (0.46), and generally have a negative or close to zero correlation with all other species. Finally, *Limanda aspera* and *Pleuronectes quadrituberculatus* have a high correlation (0.65). These ten species represent eight different genera, and the average within-genus correlation (0.68) is higher than the among-genus correlation (0.15).

The first three factors capture 79% of total explained variance, and we present different summaries for these factors including the average spatial distribution, the average temporal trends and the loadings of each species on the dominant factors (Fig. 2). The spatial range (defined as the distance at which locations have a correlation of 10%, as calculated from estimates of $\kappa_{xy}$ and $\kappa_{x}$) is larger for spatial (1987 km) than spatio-temporal variation (343 km). This suggests that the Bering Sea has large differences in community structure from northern to southern boundaries, but that annual variation at a given site is only predictive of variation at nearby sites. Factor 1 (top row of Fig. 2) is highly associated with *G. chalcogrammus* and *G. macrocephalus*, and has weaker positive associations with *Hippoglossoides elassodon*, *Hippoglossus stenolepis*, *L. aspera* and *Podotherecus accipenserinus*. This factor is highest at intermediate depths and the south-eastern shelf of the Bering Sea, and is lowest in the northern portion of the Bering Sea. This factor shows a drop in 2008 followed by an increase through 2013, and this trend is consistent with the abundance estimates from population dynamics models for these

Figure 2 Summary of results for the dominant three dynamic factors in the first case study ('among-year dynamics of Bering Sea demersal community'), including a depiction of average spatial effect (the median across years of the value of the factor at a given site, left column, where blue is a low value, grey is an intermediate value and red is a high value), the average temporal effect (middle column, where each grey line corresponds to a site, and the solid black line is the median value across sites for a given year) and estimated loadings for each species on each factor, where the proportion of total explained variance that is explained by each factor is listed in the upper left corner of each panel. We display the first three factors, which collectively explain c. 80% of the variance explained by all six estimated factors. When plotting spatial estimates, we extrapolate from knots to all locations within the sampling domain of the survey.
species, where *G. chalcogrammus* and *G. macrocephalus* had distributions that were restricted by a large cold pool starting in 2008 (Aydin et al., 2014). Factor 2 is positively associated with *C. bairdi* and *C. opilio* and negatively associated with *H. stenolepis* and *Platichthys stellatus*. It is highest in deep waters of the northern portion, and captures a previously documented ‘environmental ratchet’ (Orensanz et al., 2004) in which the distribution of *C. opilio* contracted north during the early 1990s and subsequently did not fully recolonize the southern area of the survey. Factor 3 is positively associated with *C. bairdi*, *Limanda aspera*, *P. stellatus* and *Podothecus accipenserinus*, and is positive in the portion of the Bering Sea adjacent to the Aleutian Islands.

**Case study 2: within-year dynamics of the butterfly community in Ohio**

Analysing data regarding within-year dynamics of butterflies in Ohio, we find that the sixth factor adds less than 5% of total variance, and we therefore proceed by interpreting a model with six factors. This model estimates a positive correlation among most species (see Appendix S4). These 63 species represent 44 genera, and the average within-genus correlation (0.52) is higher than the average among-genus correlation (0.45). As a post hoc test, we estimated the difference and its standard error (0.072, SE = 0.036). A one-sided Wald test therefore indicates that the increased correlation for species in the same genus relative to other species pairs is statistically significant (*P* = 0.023).

The first two factors explain 67% of total explained variance (Fig. 3). The spatial range is smaller for spatial (14.7 km) than spatio-temporal variation (333.1 km), suggesting that factors account for fine-scale spatial variation (e.g. there are high- and low-density sites near Cleveland in north-east Ohio) but that changes over time are generally synchronous among sites. Factor 1 is positively associated with almost all species (the major exception being *Poanes viator*), and captures an increase and subsequent decrease in butterfly densities (with a broad peak between 150 and 250). Factor 2 is also strongly associated with many species but has a mix of positive and negative associations. It captures a declining trend in abundance during the summer, such that species with a positive association (e.g. species no. 34, *Megisto cymela*) have a peak in the late spring, while species with a negative association (e.g. species no. 53, *Pyrgus communis*) are predicted to peak in the late summer. For model validation, we also compare predictions of peak density from our analysis with flight curves reported by Belth (2012) in Illinois (see Appendix S5). This comparison illustrates that SDFA is able to correctly predict narrow peaks in density (e.g. *Satyrodes eurydice* with 11 occurrences, or *Euphydryas phaeton* with 10 occurrences) and broad peaks for other data-poor species (e.g. *Euphyes vestris*, with 43 occurrences). However, SDFA also provides poor estimates in some cases (e.g. *Hermeuptchia sosybius*, with 11 occurrences, where it misses the earliest of three peaks reported in nearby Illinois).

**Simulation experiment**

Estimated parameters from the simulation experiment (Fig. 4) illustrate that the SDFA model is able to accurately estimate parameters governing spatial and temporal variation.
In particular, the model can accurately estimate the degree of autocorrelation (0.8), the ratio of spatial and temporal variance (0.12), the spatial range of spatial variation ($\Theta$ (0.25, derived from $\kappa_{\Theta}$)) and the spatial range of spatio-temporal variation ($\Gamma$ (0.25, derived from $\kappa_{\Gamma}$)). However, we note that one replicate resulted in an estimate of a spatial scale for spatial variation that approached infinity, indicating that the model in this replicate estimated essential zero spatial variation ($\Theta = 0$).

Comparison of the estimated and true correlation in density among species, Corr($\psi_i$, $\psi_j$), also illustrates that the SDFA is able to precisely identify similarities and differences among species, given the magnitude of data available in the simulation experiment (Fig. 5). In particular, a plot of the estimated and true correlation is generally on the 1:1 line, and this 1:1 line explains nearly 90% of variation in the correlation among species.

**DISCUSSION**

In this study we have developed a SDFA model for use as a JDSDM, and used two case studies (involving different taxa, numbers of species and spatial and temporal scales) to show that SDFA is useful for species ordination. In particular, phylogenetically related species had greater similarity in spatio-temporal dynamics than unrelated species in both case studies (analogous to the conclusion using spatial factor analysis in Thorson et al., 2015a). This result suggests that well-sampled species could be used as indicators for related species that are poorly sampled in each community. If this result holds for other communities (and all else being equal), we also expect that phylogenetically diverse communities will be more stable (i.e. have a lower variance in total abundance spatially and over time due to the portfolio effect; see Doak et al., 1998) than communities composed of phylogenetically related species.

We have also demonstrated that SDFA can be used for community monitoring. In particular, our first case study identified a northward shift in bottom-associated fishes in the Bering Sea, where this recent shift coincides with a receding cold pool after its expansion in 2008. The second case study showed that seasonal dynamics could be predicted with reasonable accuracy even for species with as few as 10 observations. Many other taxa have shown responses to climate change through phenological shifts, leading to changes in their spatio-temporal distributions and correlations among species (Root et al., 2003). Recent research suggests that bees (Kerr et al., 2015), butterflies (Parmesan et al., 1999) and marine...
Figure 5 A scatterplot showing the estimated correlation in density among species, \( \text{Cor}(\psi_i, \psi_j) \) for species \( i \) and \( j \), and the true correlation, where each is calculated from the estimated and true loadings matrices \( L_{i,j} \), respectively. The proportion of variance explained when regressing the true correlation on the estimated correlation is shown at the top.

Fishes (Pinsky et al., 2013), among other taxa, are already exhibit shifting ranges or phenology. We hypothesize that SDFA and other JDSDMs will have a growing role in identifying spatial and phenological shifts in ecological communities.

We envision several fruitful lines of future research regarding spatio-temporal community monitoring. Most obviously, future research could explore the benefits in precision or interpretation from the inclusion of measured covariates (as can be easily done in SDFA). Covariates may be particularly important for extrapolation, while our use of Gaussian random fields may be sufficient when interpolating among densely located sampling data (Bahn & McGill, 2007). We also recommend exploring the use of additional data types (e.g. presence–absence or repeated-measures sampling) when fitting JDSDMs to spatio-temporal community data. Finally, there is great need for determining whether ‘hotspots’ in species richness or density change substantially over time. Hotspots are frequently used in conservation efforts when planning spatial protection of harvested or imperilled species, and changes over time in the location of hotspots have substantial implications for the usefulness of these spatial protections (Piacenza et al., 2015).

Recent multispecies occurrence models represent an alternative approach to estimation of community dynamics relative to covariates (Zipkin et al., 2010; Iknayan et al., 2014). These models generally focus on estimating species distributions using presence–absence data while explicitly accounting for detection biases that occur during sampling. Detection probability could similarly be estimated in SDFA by incorporating an alternative distribution for repeated sampling at a given site and time (Royle & Dorazio, 2008). While these community modelling approaches allow for joint estimation of species occurrence/abundance across a landscape, they typically assume that unexplained variations in dynamics (‘process errors’) are independent among species. Our multivariate approach for estimating spatial correlation among species may allow JDSDMs to more efficiently use available data to model spatio-temporal patterns.

We envision a complementary role between phenomenological and mechanistic models for community dynamics, and an increased emphasis on mechanism could be incorporated in many ways. In particular, phylogenetic relatedness or similarity in functional traits could help inform estimates of species similarity. Covariance among species, \( \text{Cov}(\psi_i, \psi_j) \), could be estimated as an explicit function of phylogenetic distance, in addition to residual covariance modelled via the loadings matrix \( (LL)^T \). Increased emphasis on mechanism could also be incorporated into future JDSDMs by estimating a ‘species interaction matrix’ that is used to approximate density-dependent interactions among species (Ives et al., 2003). However, we envision the need for substantial dimension reduction when estimating a species interaction matrix within a large community (Kissling et al., 2012).

We conclude by offering guidance for future researchers interested in using SDFA as a JDSDM. Our software is publicly available, and can be applied to other data sets involving count (discrete) or density (continuous) data for multiple species at multiple sites and time periods. We recommend that parameters are estimated while sequentially increasing the number of factors (i.e. estimating one factor, then estimating two factors while starting parameters at their previous estimates, etc.) and stopping based on decreasing variance explained or model-selection criteria. We tested this process using communities with up to 100 species, sites and time intervals, and it is feasible to estimate parameters for these data sets on a scale of hours. We hypothesize that JDSDMs such as SDFA will continue to grow in importance for applied and community ecologists interested in either species ordination or spatio-temporal community monitoring.

ACKNOWLEDGEMENTS

We gratefully acknowledge the Ohio Lepidopterist Society (including Jerry Weidmann and Rick Ruggles) for providing access to the butterfly survey data, the many volunteers who conducted butterfly surveys, and the scientists who sampled demersal fish dynamics in the Bering Sea. We also thank Adrian Stier, Paul Conn, and two anonymous referees for comments on an earlier draft.

REFERENCES

Aydin, K., Barbeaux, S., Barnard, D. et al. (2014) Stock assessment and fishery evaluation report for the groundfish resources of the Bering Sea/Aleutian Islands Regions, Alaska Fisheries Science Center, Anchorage, AK.


SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article at the publisher’s web-site.

Appendix S1 Linear parameterization for each dynamic factor.
Appendix S2 Identifiability conditions.
Appendix S3 Transformation of loadings matrix and factors to aid interpretation.
Appendix S4 Estimated correlation among species in case study 2.
Appendix S5 Validation of estimated flight curves by comparison with independent reports.

BIOSKETCH

James Thorson is a statistical ecologist who is interested in estimating density dependence and multispecies interactions at large spatial scales using information from multiple data sources. He works with a wide range of taxa and biological systems, but always seeks to apply new methods to improve management of marine fishes.

Editor: Alberto Jiménez-Valverde