Joint dynamic species distribution models: a tool for community ordination and spatiotemporal monitoring

James T. Thorson[1*], James N. Ianelli[2], Elise A. Larsen[3], Leslie Ries[4], Mark D. Scheuerell[5], Cody Szuwalski[6,7], Elise F. Zipkin[8,9]

[1] Fisheries Resource Assessment and Monitoring Division, Northwest Fisheries Science Center, National Marine Fisheries Service, NOAA, Seattle, WA, USA

[2] Resource Ecology and Fisheries Management Division, Alaska Fisheries Science Center, National Marine Fisheries Service, NOAA, Seattle, WA

[3] National Socio-environmental Synthesis Center, Annapolis, MD, USA

[4] Department of Biology, Georgetown University, Washington, DC, USA

[5] Fish Ecology Division, Northwest Fisheries Science Center, National Marine Fisheries Service, NOAA, Seattle, WA, USA

[6] Bren School of Environmental and Resource Management, University of California, Santa Barbara, CA, USA

[7] Marine Science Institute, University of California, Santa Barbara, CA, USA

[8] Department of Integrative Biology, Michigan State University, East Lansing, MI, USA

[9] Ecology, Evolutionary Biology, and Behavior Program, Michigan State University, East Lansing, MI, USA

[*] Corresponding author: James.Thorson@noaa.gov

This is the author manuscript accepted for publication and has undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version record. Please cite this article as doi:10.1111/geb.12464.

Keywords: Species distribution model; geostatistics; flight curve; Bering Sea; spatiotemporal model; species co-occurrence; dynamic factor analysis; species ordination

Running title: joint dynamic species distribution models

Number of words in abstract: 284

Number of words in body: 5677

Number of references: 51

Abstract

Aim: Spatial analysis of the distribution and density of species is of continued interest within theoretical and applied ecology. Species distribution models (SDM) are increasingly used to analyze count, presence/absence, and presence-only data sets. There is a growing literature regarding dynamic SDM (which incorporate temporal variation in species distribution), joint SDM (which simultaneously analyze 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 therefore 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 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 that JDSDM can be used for species ordination, i.e., showing that dynamics for butterfly species within the same genus is significantly more correlated that species from different genera. We also demonstrate how JDSDM can rapidly identify dominant patterns in community dynamics, including the decline and recovery of several Bering Sea fishes following 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.

Introduction

Understanding the spatial distribution of species is a central concern in ecology, and is necessary to evaluate spatial protections for threatened species, interpret abiotic and biotic impacts on species distribution, track the spread of invasive species, and interpret climate impacts on species and communities (Hooten & Wikle, 2010; Rassweiler *et al.*, 2014). Species distribution models (SDM) 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).

Species distribution models 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 geographic 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 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, species distribution models 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 this as a summary of community distribution is likely to yield biased estimates of community-level patterns (Clark *et al.*, 2013, and references therein). Joint species

distribution models are an alternative procedure for simultaneously modeling the distribution of multiple species within a community, and were originally developed to estimate total species richness when accounting for imperfect species distribution (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 sample level while also accounting for standard habitat predictors (Latimer *et al.*, 2009; Ovaskainen *et al.*, 2010; Clark *et al.*, 2013; 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 species distribution models (Dormann, 2007). Ecologists have a growing ability to account for spatial autocorrelation in species density via geostatistical 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. JDSMD has 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, JDSDM could be used for to access the degree to which species have similar or different spatiotemporal patterns ("species ordination"). For example, classifying species based on similar spatiotemporal dynamics could be used to determine whether high-density species are useful as indicators of abundance trends for rare or poorly-surveyed species. Second, 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.*, 2015a; 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 (SDFA) model and propose its use as a JDSDM. SDFA 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. SDFA 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 SDFA in species ordination and to identify large-scale patterns in community dynamics. The first involves 33 years of data for ten 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 spatiotemporal dynamics and species similarity.

Methods

A primer on dynamic factor analysis (DFA)

Before describing spatial dynamic factor analysis (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., sub-populations of a species). Although ecologists may be more familiar with rank-reduction using principle-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, 1989) 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 sub-populations 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 the time series will 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 $M << N$. These latent variables can be thought of as unmeasured or unknown environmental drivers of ecological processes. Recent applications of DFA including (1) identifying three dominant patterns in the abundance of 34 Pacific salmon stocks from North America and Asia (Stachura *et al.*, 2014), and important differences in the sensitivity of 80 boreal streams to variation in summer air temperature (Lisi *et al.*, 2015).

Introducing Spatial Dynamic Factor Analysis (SDFA)

We next develop a new spatial dynamic factor analysis (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 use 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 variable (termed "factors" in our model, in analogy to DFA).

Latent variable models are seeing increased use in community ecology as a sparse method for analyzing correlations among multiple species (Ovaskainen *et al.*, 2015a; Thorson *et al.*, 2015a; Warton *et al.*, 2015). Each factor, ψ, 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 ($s \in D \in \mathcal{R}^2$, where D is the domain, indexed by latitude/longitude or any other appropriate 2-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).

Data for each species p are assumed to arise as a function of log-density θ_p for that species:

$$c_p(s,t) \sim g\left(\exp(\theta_p(s,t))\right) \quad (1)$$

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 modeled as a linear combination of n_j factors:

$$\theta_p(s,t) = \sum_{j=1}^{n_j} L_{p,j}\psi_j(s,t) + \sum_{k=1}^{n_k} \gamma_{k,p} x_k(s,t) \quad (2)$$

where the loadings matrix $\mathbf{L}$ represents the association $L_{p,j}$ 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:

$$\psi_j(s,t+1) = \rho_j \psi_j(s,t) + \omega_j(s) + \xi_j(s,t) \quad (3)$$

where ρ_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 spatiotemporal variation in dynamics for factor j. Each factor therefore follows a spatial Gompertz process (Thorson *et al.*, 2015b), where ρ can be interpreted as the strength of density dependence. Spatial variation ω_j is represented as a Gaussian random field:

$$\boldsymbol{\omega}_j \sim \mathrm{MVN}(\mathbf{0}, \boldsymbol{\Sigma}_\omega) \quad (4)$$

where the covariance in average density $\boldsymbol{\Sigma}_\omega$ is assumed to be higher for nearby locations than for geographically distant locations. We specifically model the relationship between geographic distance between location s and $s+h$ (i.e., where $|h|$ is the distance between locations) using a Matérn function:

$$\mathrm{Var}\left(\omega_j(s), \omega_j(s+h)\right) = \frac{1}{2^{\nu-1}\Gamma(\nu)\tau_\omega^2}(\kappa_\omega |h|)^\nu K_\nu(\kappa_\omega |h|) = f(|h|, \kappa_\omega, \tau_\omega^2) \quad (5)$$

where κ_ω governs the range over which covariance declines as a function of distance $|h|$, ν governs the smoothness of the covariance matrix (we assume that $\nu=1$ in the following) and τ_ω governs the marginal variance of spatial variation. The covariance in spatiotemporal variation $\boldsymbol{\xi}_j$ is defined similarly:

$$\boldsymbol{\xi}_j \sim \mathrm{MVN}(\mathbf{0}, \boldsymbol{\Sigma}_\xi \otimes \mathbf{I}) \quad (6)$$

where covariance $\boldsymbol{\Sigma}_\xi$ is calculated given range parameter κ_ξ and precision τ_ξ and $\mathbf{I}$ is an n_t by n_t identity matrix (where n_t is the number of modeled time intervals).

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

$$\psi_j(s,1) = \phi_j + \frac{\omega_j(s)}{1-\rho_j} + \xi_j(s,t) \quad (7)$$

where $\omega_j(s)/(1-\rho_j)$ is the median of the stationary distribution for this factor, and ϕ_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 (see Thorson et al. (2015b) or Appendix S1 for details).

Finally, observation of density $c_p(s,t)$ for species p at site n in year t is assumed to follow a simple measurement process. In the following case-studies, we use both counts (i.e., integer-valued data) and catch-per-unit-effort (i.e., non-negative real-valued data). For count data, we propose an overdispersed lognormal-Poisson process:

$$\Pr\left(C = c_p(s,t)\right) = \text{Poisson}\left(C; \lambda_p(s,t)\exp(\eta_p(s,t))\right) \quad (8a)$$

where mean count $\lambda_p(s,t) = \exp\left(\theta_p(s,t)\right)$ and $\eta_p(s,t) \sim N(0,\sigma_p^2)$ represents lognormal overdispersion with variance σ_p^2 for species p. For catch-per-unit-effort data, we use a zero-inflated gamma distribution, which describes the probability of the probability density function (PDF) for positive catches with an additional probability mass at zero (i.e., the probability of not encountering species p):

$$\Pr\left(C = c_p(s,t)\right) = \begin{cases} \exp(-\nu_{p,2}\lambda_p(s,t)) & \text{if } C = 0 \\ \left(1-\exp(-\nu_{p,2}\lambda_p(s,t))\right)\text{Gamma}\left(C; \nu_{p,1}^{-2}, \frac{\lambda_p(s,t)}{1-\exp(-\nu_{p,2}\lambda_p(s,t))}\nu_{p,1}^2\right) & \text{if } C > 0 \end{cases} \quad (8b)$$

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 $\nu_{p,2}\lambda_p(s,t)$. 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 $Var(\psi_j)$ 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:

$$\mathrm{Var}(\theta_{p1}(s,t),\theta_{p2}(s,t)) = \sum_j L_{p1,j}L_{p2,j} \quad (9)$$

This relationship has been noted previously for joint species distribution models (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.*, 2015b)).

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 ρ (we assume that $\rho_j = \rho$ for all factors), initial condition φ_j, variance ratio χ (used to calculate τ_ω and τ_ξ, and assumed to be constant among factors), range for spatial variation κ_ω, range for spatiotemporal variation κ_ξ, measurement error parameters $\boldsymbol{\nu}$, and loadings matrix **L**. We treat spatial variation $\boldsymbol{\omega}_j$ and

spatiotemporal variation $\boldsymbol{\varepsilon}_j$ (used to calculate $\boldsymbol{\xi}_j$) as random effects for each factor j. Finally, the lognormal-Poisson distribution (used for applications involving count data) involves treating overdispersion $\boldsymbol{\eta}$ 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 (Eq. 7a/7b) and the probability of the dynamic factors (Eq. 3 and 5).
2. We then use the Laplace approximation (Skaug & Fournier, 2006) to approximate the marginal likelihood, obtained when integrating the joint likelihood with respect to random effects ($\boldsymbol{\omega}_j$, $\boldsymbol{\varepsilon}_j$, and $\boldsymbol{\eta}$).
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 $\psi_j(s,t)$) 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 Eq. 3 and 5 (Lindgren *et al.*, 2011; Thorson *et al.*, 2015b)). Parameter estimation is feasible on a laptop using 1-6 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.

Similar to conventional factor analysis, the resulting estimates of loadings **L** and factors $\psi_j(n,t)$ can be rotated to ease interpretation. We therefore calculate a linear transformation matrix **H**, and then interpret $\mathbf{L}^*=\mathbf{LH}$ and $\mathbf{\Psi}^*=\mathbf{\Psi H}^{-1}$. 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 favors selecting 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 1 and 2 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 analysis of very large communities, while only adding factors that explain a biological significant portion of community variance.

Case study examples

To demonstrate the usefulness of SDFA to estimate species similarity and identify broad patterns in community spatiotemporal dynamics, we present results for two case-study applications that differ substantially in taxa, number of species, spatial scale, and temporal scale:

1. *Bering Sea demersal community*: Our first case study analyzes data from annual bottom trawl surveys of the eastern Bering Sea during summers 1982-2014. This survey has been conducted by the Alaska Fisheries Science Center with minimal changes in survey design over this period, and consists of about 375 bottom trawl tows per year. Each tow is on bottom for about 30 minutes and data collection involves 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 biogeographic regions: the inner (0-50 m. depth), middle (50-100 m. depth) and outer domain (100-200 m. depth (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 location 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 ten most abundant species (by weight) observed in the

demersal community (see Table 2) and, given that we analyze catches in weight, we use the delta-gamma distribution for measurement errors (Eq. 7b).

2. *Ohio butterfly assemblage*: As a second case study, we analyze count data from 2010 regarding the distribution of butterflies at 58 sampled locations in Ohio. This state-wide butterfly monitoring program, conducted by the Ohio Lepidopterist Society, consists 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 programs. All adult butterflies present at each site are counted approximately weekly during the period from April 1 to Oct. 31 (Julian dates 91-304), although not every location is surveyed every week. These counts have previously been analyzed 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, e.g. to identify phenological changes caused by climate change (Roy & Sparks, 2000). We analyze counts for all 63 butterfly species that are encountered 10 or more times during 2010 (Table 2), and use a lognormal-Poisson distribution for measurement errors (Eq. 7a). SDFA analyzes observations within discrete time periods, $t \in \{1,2,\ldots,T\}$, so we aggregate observations into 16 two-week "intervals" from April 1 to Oct. 31.

The first case study demonstrates a smaller subset of important species at large 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).

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 1x1 grid. Dynamics evolve following five factors, where each factor has a spatial scale of 0.25, a marginal standard deviation for spatial variation and spatiotemporal process error of 0.5, and where each factor is weakly density-dependence (i.e., ρ=0.8) and starts in the first year with a small deviation away from its equilibrium (i.e., φ_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 5 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., γ_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

Analyzing data from the Bering Sea demersal community, we find that the 6th factor adds less than 5% of total variance, so we proceed by interpreting the SDFA model with 6 factors. Given this model, SDFA illustrates three main groups (Fig. 1). The first involves *Gadus chalcogrammus* and *G. macrocephalus*, which have a high correlation (0.91). *Chionoecetes bairdi* and *C. 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 κ_ω and κ_ξ) is larger for spatial (1987 km) than spatiotemporal 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 *Gadus chalcogrammus* and *G. macrocephalus*, and has weaker positive associations with *Hippoglossoides elassodon*, *Hippoglossus stenolepis*, *Limanda aspera*, and *Podothecus accipenserinus*. This factor is highest at intermediate depths and the Southeastern 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 species, where *Gadus 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 *Chionoecetes bairdi* and *C. opilio* and negatively associated with *Hippoglossus 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 *Chionoecetes opilio* contracted north during the early 1990s and subsequently did not fully recolonise the southern area of the survey. Factor 3 is positively

associated with *Chionoecetes bairdi*, *Limanda aspera*, *Platichthys 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

Analyzing data regarding within-year dynamics of butterflies in Ohio, we find that the 6th 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 in Supporting Information). 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 a 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 spatiotemporal variation (333.1 km), suggesting that both factors account for fine-scale spatial variation (e.g., there are high and low-density sites near Cleveland in Northeast 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., #34, *Megisto cymela*) have a peak in the late spring, while species with a negative association (e.g., #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 in Supporting Information). This comparison illustrates that SDFA is able to corrected 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 $\boldsymbol{\omega}$ (0.25, derived from κ_ω), and the spatial range of spatiotemporal variation $\boldsymbol{\varepsilon}$ (0.25, derived from κ_ε). 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 ($\boldsymbol{\omega}$=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 spatial dynamic factor analysis (SDFA) model for use as a joint dynamic species distribution model (JDSDM), and used two case studies (involving different taxa, number 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 spatiotemporal 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 spatiotemporal 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 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 spatiotemporal community monitoring. Most obviously, future research could explore the benefits in precision or interpretation from including 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 JDSDM to spatiotemporal 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 protections of harvested or imperiled species, and changes over time in the location of hotspots has substantial implications for the usefulness of these spatial protections (Piacenza *et al.*, 2015).

Recent multi-species occurrence models represent an alternative approach to estimate 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 variation 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 spatiotemporal 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 ($Cov(\psi_i, \psi_j)$) could be estimated as an explicit function of phylogenetic distance, in addition to residual covariance modeled via the loadings matrix ($\mathbf{L}\mathbf{L}^T$). Increased 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 1 factor, then estimate 2 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 spatiotemporal 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 reviewers for comments on an earlier draft.

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.

Bibliography

Aydin, K., Barbeaux, S., Barnard, D., Chilton, L., Clark, B., Conners, M.E., Conrath, C., Dalton, M., Echave, K., Fritz, L., Furuness, M., Hanselman, D., Haynie, A., Hoff, J., Honkalehto, T., Hulson, P.J., Ianelli, J., Kotwicki, S., Lauth, S., Lowe, S., Lunsford, C., McGilliard, C., McKelvey, D., Nichol., D., Norcross, B., Ormseth, O.A., Palsson, W., Rodgveller, C.J., Rooper, C.N., Sigler, M., Spencer, P., Spies, I., Stockhausen, W., Stram, D., TenBrink, T., Thompson, G., Tribuzio, C., Wilderbuer, T. & Williamson, N. (2014) *Stock Assessment and Fishery Evaluation Report for the Groundfish Resources of the Bering Sea/Aleutian Islands Regions*, Anchorage, AK.

Bahn, V. & McGill, B.J. (2007) Can niche-based distribution models outperform spatial interpolation? *Global Ecology and Biogeography*, **16**, 733–742.

Belth, J.E. (2012) *Butterflies of Indiana: a field guide*, Indiana University Press, Bloomington, Indiana.

Cayton, H.L., Haddad, N.M., Gross, K., Diamond, S.E. & Ries, L. (2015) Do growing degree days predict phenology across butterfly species? *Ecology*.

Clark, J.S., Gelfand, A.E., Woodall, C.W. & Zhu, K. (2013) More than the sum of the parts: Forest climate response from Joint Species Distribution Models. *Ecological Applications*.

Conn, P.B., Johnson, D.S., Ver Hoef, J.M., Hooten, M.B., London, J.M. & Boveng, P.L. (2014) Using spatiotemporal statistical models to estimate animal abundance and infer ecological dynamics from survey counts. *Ecological Monographs*, **85**, 235–252.

DeVisser, M.H., Messina, J.P., Moore, N.J., Lusch, D.P. & Maitima, J. (2010) A dynamic species distribution model of Glossina subgenus Morsitans: The identification of tsetse reservoirs and refugia. *Ecosphere*, **1**, art6.

Doak, D.F., Bigger, D., Harding, E.K., Marvier, M.A., O'malley, R.E. & Thomson, D. (1998) The statistical inevitability of stability-diversity relationships in community ecology. *The American Naturalist*, **151**, 264–276.

Dorazio, R.M., Kéry, M., Royle, J.A. & Plattner, M. (2010) Models for inference in dynamic metacommunity systems. *Ecology*, **91**, 2466–2475.

Dorazio, R.M. & Royle, J.A. (2005) Estimating Size and Composition of Biological Communities by Modeling the Occurrence of Species. *Journal of the American Statistical Association*, **100**, 389–398.

Dorazio, R.M., Royle, J.A., Söderström, B. & Glimskär, A. (2006) Estimating species richness and accumulation by modeling species occurrence and detectability. *Ecology*, **87**, 842–854.

Dormann, C.F. (2007) Effects of incorporating spatial autocorrelation into the analysis of species distribution data. *Global Ecology and Biogeography*, **16**, 129–138.

Earn, D.J.D., Levin, S.A. & Rohani, P. (2000) Coherence and Conservation. *Science*, **290**, 1360–1364.

Elith, J. & Leathwick, J.R. (2009) Species distribution models: ecological explanation and prediction across space and time. *Annual Review of Ecology, Evolution, and Systematics*, **40**, 677.

Harris, D.J. (2015) Generating realistic assemblages with a Joint Species Distribution Model. *Methods in Ecology and Evolution*, n/a–n/a.

Harvey, A.C. (1990) *Forecasting, structural time series models and the Kalman filter*, Cambridge university press, Cambridge, UK.

Hooten, M.B. & Wikle, C.K. (2010) Statistical agent-based models for discrete spatio-temporal systems. *Journal of the American Statistical Association*, **105**, 236–248.
Iknayan, K.J., Tingley, M.W., Furnas, B.J. & Beissinger, S.R. (2014) Detecting diversity: emerging methods to estimate species diversity. *Trends in ecology & evolution*, **29**, 97–106.
Ives, A.R., Dennis, B., Cottingham, K.L. & Carpenter, S.R. (2003) Estimating community stability and ecological interactions from time-series data. *Ecological monographs*, **73**, 301–330.
Kerr, J.T., Pindar, A., Galpern, P., Packer, L., Potts, S.G., Roberts, S.M., Rasmont, P., Schweiger, O., Colla, S.R., Richardson, L.L., Wagner, D.L., Gall, L.F., Sikes, D.S. & Pantoja, A. (2015) Climate change impacts on bumblebees converge across continents. *Science*, **349**, 177–180.
Kissling, W.D., Dormann, C.F., Groeneveld, J., Hickler, T., Kühn, I., McInerny, G.J., Montoya, J.M., Römermann, C., Schiffers, K., Schurr, F.M. & others (2012) Towards novel approaches to modelling biotic interactions in multispecies assemblages at large spatial extents. *Journal of Biogeography*, **39**, 2163–2178.
Kristensen, K. (2014) *TMB: General random effect model builder tool inspired by ADMB.*,.
Kristensen, K., Nielsen, A., Berg, C.W. & Skaug, H. (In press) Template Model Builder TMB. *Journal of Statistical Software*.
Latimer, A.M., Banerjee, S., Sang Jr, H., Mosher, E.S. & Silander Jr, J.A. (2009) Hierarchical models facilitate spatial analysis of large data sets: a case study on invasive plant species in the northeastern United States. *Ecology Letters*, **12**, 144–154.
Lindgren, F., Rue, H. & Lindström, J. (2011) An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **73**, 423–498.
Martin, T.G., Wintle, B.A., Rhodes, J.R., Kuhnert, P.M., Field, S.A., Low-Choy, S.J., Tyre, A.J. & Possingham, H.P. (2005) Zero tolerance ecology: improving ecological inference by modelling the source of zero observations. *Ecology Letters*, **8**, 1235–1246.
Merow, C., Lafleur, N., Silander, J.A., Wilson, A.M. & Rubega, M. (2011) Developing dynamic mechanistic species distribution models: predicting bird-mediated spread of invasive plants across northeastern North America. *The American Naturalist*, **178**, 30–43.
Orensanz, J., Ernst, B., Armstrong, D.A., Stabeno, P. & Livingston, P. (2004) Contraction of the geographic range of distribution of snow crab (Chionoecetes opilio) in the eastern Bering Sea: An environmental ratchet? *California Cooperative Oceanic Fisheries Investigations Report*, **45**, 65.
Ovaskainen, O., Abrego, N., Halme, P. & Dunson, D. (2015a) Using latent variable models to identify large networks of species-to-species associations at different spatial scales. *Methods in Ecology and Evolution*.
Ovaskainen, O., Hottola, J. & Siitonen, J. (2010) Modeling species co-occurrence by multivariate logistic regression generates new hypotheses on fungal interactions. *Ecology*, **91**, 2514–2521.
Ovaskainen, O., Roy, D.B., Fox, R. & Anderson, B.J. (2015b) Uncovering hidden spatial structure in species communities with spatially explicit joint species distribution models. *Methods in Ecology and Evolution*, n/a–n/a.
Ovaskainen, O. & Soininen, J. (2011) Making more out of sparse data: hierarchical modeling of species communities. *Ecology*, **92**, 289–295.

Parmesan, C., Ryrholm, N., Stefanescu, C., Hill, J.K., Thomas, C.D., Descimon, H., Huntley, B., Kaila, L., Kullberg, J., Tammaru, T. & others (1999) Poleward shifts in geographical ranges of butterfly species associated with regional warming. *Nature*, **399**, 579–583.

Pearman, P.B., Guisan, A., Broennimann, O. & Randin, C.F. (2008) Niche dynamics in space and time. *Trends in Ecology & Evolution*, **23**, 149–158.

Piacenza, S.E., Thurman, L.L., Barner, A.K., Benkwitt, C.E., Boersma, K.S., Cerny-Chipman, E.B., Ingeman, K.E., Kindinger, T.L., Lindsley, A.J., Nelson, J., Reimer, J.N., Rowe, J.C., Shen, C., Thompson, K.A. & Heppell, S.S. (2015) Evaluating Temporal Consistency in Marine Biodiversity Hotspots. *PLoS ONE*, **10**, e0133301.

Pinsky, M.L., Worm, B., Fogarty, M.J., Sarmiento, J.L. & Levin, S.A. (2013) Marine taxa track local climate velocities. *Science*, **341**, 1239–1242.

Pollard, E. (1977) A method for assessing changes in the abundance of butterflies. *Biological conservation*, **12**, 115–134.

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

Rassweiler, A., Costello, C., Hilborn, R. & Siegel, D.A. (2014) Integrating scientific guidance into marine spatial planning. *Proceedings of the Royal Society B: Biological Sciences*, **281**, 20132252.

R Core Team (2014) *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.

Root, T.L., Price, J.T., Hall, K.R., Schneider, S.H., Rosenzweig, C. & Pounds, J.A. (2003) Fingerprints of global warming on wild animals and plants. *Nature*, **421**, 57–60.

Rota, C.T., Wikle, C.K., Kays, R.W., Forrester, T.D., McShea, W.J., Parsons, A.W. & Millspaugh, J.J. (2015) A two-species occupancy model accommodating simultaneous spatial and interspecific dependence. *Ecology*.

Roy, D.B. & Sparks, T.H. (2000) Phenology of British butterflies and climate change. *Global change biology*, **6**, 407–416.

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

Schumacher, J.D. & Stabeno, P.J. (1998) The continental shelf of the Bering Sea. *The sea*, **11**, 789–822.

Shelton, A.O., Thorson, J.T., Ward, E.J. & Feist, B.E. (2014) Spatial semiparametric models improve estimates of species abundance and distribution. *Canadian Journal of Fisheries and Aquatic Sciences*, **71**, 1655–1666.

Skaug, H. & Fournier, D. (2006) Automatic approximation of the marginal likelihood in non-Gaussian hierarchical models. *Computational Statistics & Data Analysis*, **51**, 699–709.

Soberón, J. (2007) Grinnellian and Eltonian niches and geographic distributions of species. *Ecology letters*, **10**, 1115–1123.

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

Thorson, J.T., Skaug, H.J., Kristensen, K., Shelton, A.O., Ward, E.J., Harms, J.H. & Benante, J.A. (2015b) The importance of spatial models for estimating the strength of density dependence. *Ecology*, **96**, 1202–1212.

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

Wyllie-Echeverria, T. & Wooster, W.S. (1998) Year-to-year variations in Bering Sea ice cover and some consequences for fish distributions. *Fisheries Oceanography*, **7**, 159–170.

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

Zipkin, E.F., Royle, J.A., Dawson, D.K. & Bates, S. (2010) Multi-species occurrence models to evaluate the effects of conservation and management actions. *Biological Conservation*, **143**, 479–484.

Zuur, A.F., Tuck, I.D. & Bailey, N. (2003) Dynamic factor analysis to estimate common trends in fisheries time series. *Canadian Journal of Fisheries and Aquatic Sciences*, **60**, 542–552.

Table 1 – List of symbols used in describing the spatial dynamic factor analysis model

Symbol	Description	Dimension
Variables		
$\boldsymbol{\Psi}$	Dynamic factor	$n_s \times n_j \times n_t$
$\boldsymbol{\Theta}$	Expected log-density from dynamic factors	$n_s \times n_p \times n_t$
$\mathbf{E}$	Spatiotemporal variation	$n_s \times n_j \times n_t$
$\boldsymbol{\Omega}$	Spatial variation	$n_s \times n_j$
$\mathbf{L}$	Loadings matrix	$n_p \times n_j$
$\mathbf{H}$	Varimax rotation matrix	$n_j \times n_j$
$\mathbf{X}$	Measured environmental variables	$n_i \times n_k$
C	Count data for all species, sites, and time periods	n_i
ρ	Autocorrelation in spatiotemporal variation	*1*
χ	Ratio of spatiotemporal and spatial variance	*1*
δ	Lognormal overdispersion	n_i
γ	Effect of covariates	n_k
σ_δ	Standard deviation of overdispersion	n_p
ϕ	Ratio of initial and median value for dynamic factors	n_j
τ_{space}	Magnitude of spatial variation	*1*
τ_{time}	Magnitude of spatiotemporal variation	*1*
κ	Distance of spatial correlation	*1*
η	Overdispersion in observations (used for count data)	n_i
σ	Magnitude of overdispersion in observations (used for count data)	n_p
v	Measurement error parameters (used for zero-inflated continuous-valued data)	$n_p \times 2$
Dimensions		
n_t	Number of time intervals	*1*
n_p	Number of species	*1*
n_s	Number of sites with available data	*1*
n_j	Number of dynamic factors	*1*
n_k	Number of covariates	
n_i	Number of observations	
Indices		
t	Time interval	*1*
p	Species	*1*
s	Site with available data	*1*
j	Factor	*1*
k	Covariate	*1*
i	Observation	*1*

Table 2 – List of species (with common and scientific name) used in each case study, along with "species numbers" used to reference species in later figures.

Species Number	Case study #1: Interannual dynamics of Bering Sea demersal community			Case study #2: Within-season dynamics of Ohio butterfly assemblage		
	Scientific Name	*Common Name*	*Number of encounters (12210 total samples)*	*Scientific Name*	*Common Name*	*Number of encounters (1132 total samples)*
1	*Chionoecetes bairdi*	Tanner crab	7687	*Anatrytone logan*	Delaware Skipper	19
2	*Chionoecetes opilio*	snow crab	8995	*Ancyloxypha numitor*	Least Skipper	221
3	*Gadus chalcogrammus*	walleye pollock	11786	*Asterocampa celtis*	Hackberry Emperor	78
4	*Gadus macrocephalus*	Pacific cod	11682	*Asterocampa clyton*	Tawny Emperor	20
5	*Hippoglossoides elassodon*	flathead sole	9228	*Battus philenor*	Pipevine Swallowtail	39
6	*Hippoglossus stenolepis*	Pacific halibut	8253	*Boloria bellona*	Meadow Fritillary	67
7	*Limanda aspera*	yellowfin sole	8297	*Celastrina ladon*	Spring/Summer Azure	524
8	*Platichthys stellatus*	starry flounder	1591	*Cercyonis pegala*	Common Wood-nymph	190
9	*Pleuronectes quadrituberculatus*	Alaska plaice	8284	*Chlosyne nycteis*	Silvery Checkerspot	27
10	*Podothecus accipenserinus*	sturgeon poacher	6772	*Colias eurytheme*	Orange Sulphur	583
11				*Colias philodice*	Clouded Sulphur	629
12				*Cyllopsis gemma*	Gemmed Satyr	16
13				*Danaus plexippus*	Monarch	483
14				*Enodia anthedon*	Northern Pearly-eye	64
15				*Epargyreus clarus*	Silver-spotted Skipper	476
16				*Erynnis baptisiae*	Wild Indigo Duskywing	126
17				*Erynnis brizo*	Sleepy Duskywing	13

8	*Erynnis horatius*	Horace's Duskywing	42
9	*Erynnis juvenalis*	Juvenal's Duskywing	34
0	*Euphydryas phaeton*	Baltimore Checkerspot	10
1	*Euphyes vestris*	Dun Skipper	43
2	*Eurema lisa*	Little Yellow	40
3	*Eurema nicippe*	Sleepy Orange	11
4	*Eurytides marcellus*	Zebra Swallowtail	58
5	*Everes comyntas*	Eastern Tailed-blue	455
6	*Hermeuptychia sosybius*	Carolina Satyr	11
7	*Hylephila phyleus*	Fiery Skipper	44
8	*Junonia coenia*	Common Buckeye	358
9	*Libytheana carinenta*	American Snout	53
0	*Limenitis archippus*	Viceroy	228
1	*Limenitis arthemis astyanax*	Red-spotted Purple	241
2	*Lycaena hyllus*	Bronze Copper	13
3	*Lycaena phlaeas*	American Copper	44
4	*Megisto cymela*	Little Wood Satyr	224
5	*Nymphalis antiopa*	Mourning Cloak	111
6	*Papilio glaucus*	Eastern Tiger Swallowtail	462
7	*Papilio polyxenes*	Black Swallowtail	289
8	*Papilio troilus*	Spicebush Swallowtail	279
9	*Phoebis sennae*	Cloudless Sulphur	16
0	*Pholisora catullus*	Common Sootywing	32
1	*Phyciodes tharos*	Pearl Crescent	786
2	*Pieris rapae*	Cabbage White	921
3	*Poanes hobomok*	Hobomok Skipper	73
4	*Poanes viator*	Broad-winged	15

		Skipper	
5	*Poanes zabulon*	Zabulon Skipper	102
6	*Polites mystic*	Long Dash	24
7	*Polites origenes*	Crossline Skipper	25
8	*Polites peckius*	Peck's Skipper	263
9	*Polites themistocles*	Tawny-edged Skipper	57
0	*Polygonia comma*	Eastern Comma	201
1	*Polygonia interrogationis*	Question Mark	255
2	*Pompeius verna*	Little Glassywing	73
3	*Pyrgus communis*	Common Checkered-skipper	51
4	*Satyrium calanus*	Banded Hairstreak	16
5	*Satyrodes appalachia*	Appalachian Brown	28
6	*Satyrodes eurydice*	Eyed Brown	11
7	*Speyeria cybele*	Great Spangled Fritillary	357
8	*Strymon melinus*	Gray Hairstreak	101
9	*Thymelicus lineola*	European Skipper	64
0	*Vanessa atalanta*	Red Admiral	358
1	*Vanessa cardui*	Painted Lady	56
2	*Vanessa virginiensis*	American Lady	51
3	*Wallengrenia egeremet*	Northern Broken Dash	55

Fig. 1 – Estimated correlation in spatiotemporal densities among species, $Corr(\psi_i, \psi_j)$ for species i and j, in the first case study ("among-year dynamics of Bering Sea demersal community") where species are labelled by number (see Table 1 for names), red is a positive correlation (with a correlation of 1.0 on the diagonal) and blue is a negative correlation.

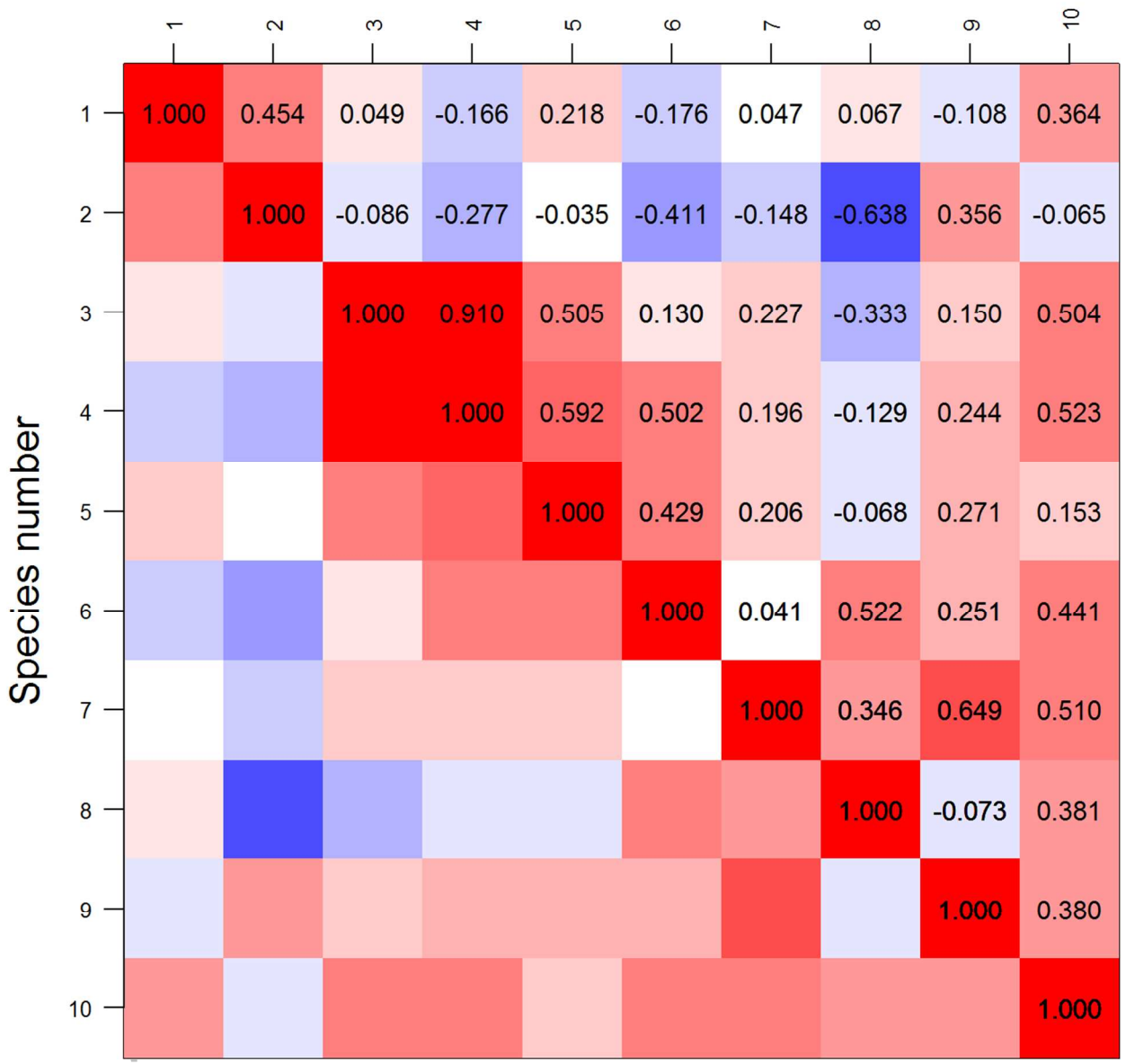


Fig. 2 – Summary of results for the dominant three dynamic factors in the first case study ("among-year dynamics of Bering Sea demersal community"), including a depiction of average spatial effect (the median across years of the value of the factor at a given site, left-column, where blue is a low value, 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 ≈80% of the variance explained by all 6 estimated factors. When plotting spatial estimates, we extrapolate from knots to all locations within the sampling domain of the survey.

e

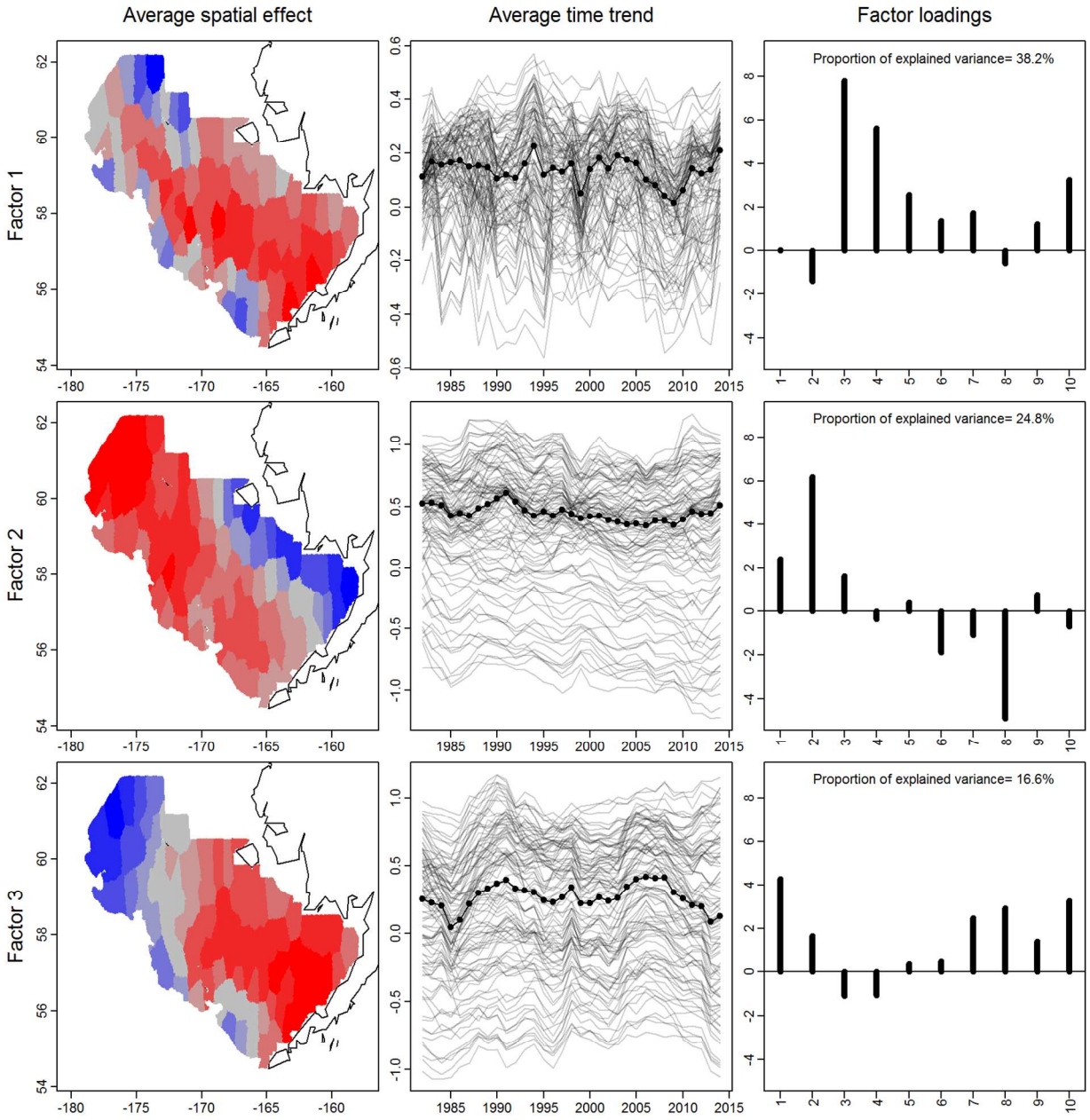


Fig. 3 – Summary of results for the dominant three dynamic factors in the first case study ("within-year dynamics of Ohio butterflies in 2010"), see Fig. 2 caption for details. We display the first two factors, which collectively explain ≈65% of the variance explained by all 6 estimated factors. When plotting spatial estimates, we only plot estimates at knots because data do not arise from a designed survey.

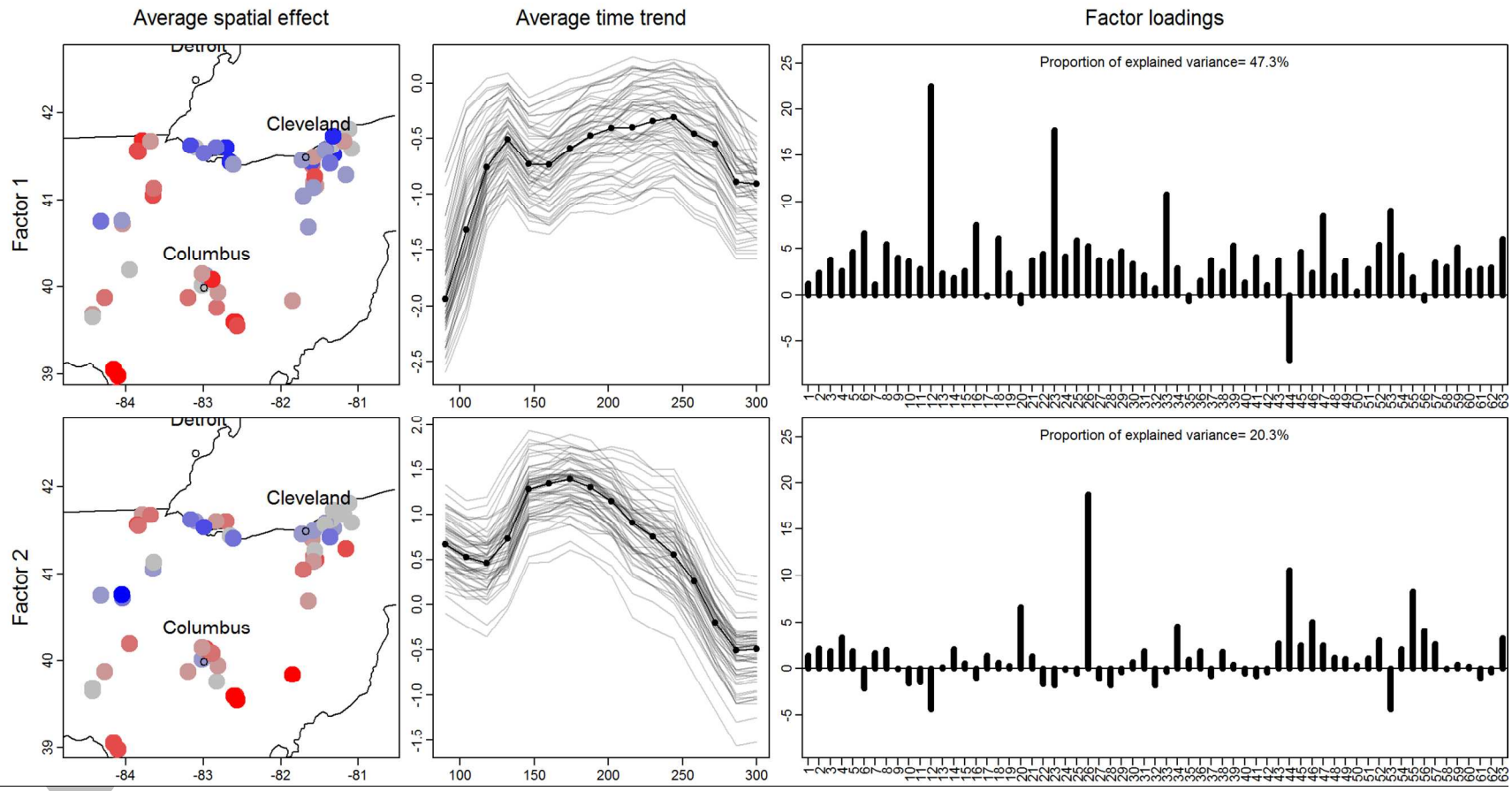


Fig. 4 – Summary of simulation experiment testing the small-sample properties of the dynamic spatial factor analysis model, where each panel shows the true simulated value (dashed vertical line) and distribution of estimates for each simulation replicate (histogram). Panels show the ratio of temporal and spatial variance (λ), the temporal autocorrelation (ρ), the average spatial range (defined as the distance at which correlations drop to 10% as calculated from κ_ω), and the spatial range of spatiotemporal variance (calculated from κ_ξ). In the top-right portion each panel, we list an estimates that exceed the x-axis boundary of that panel.

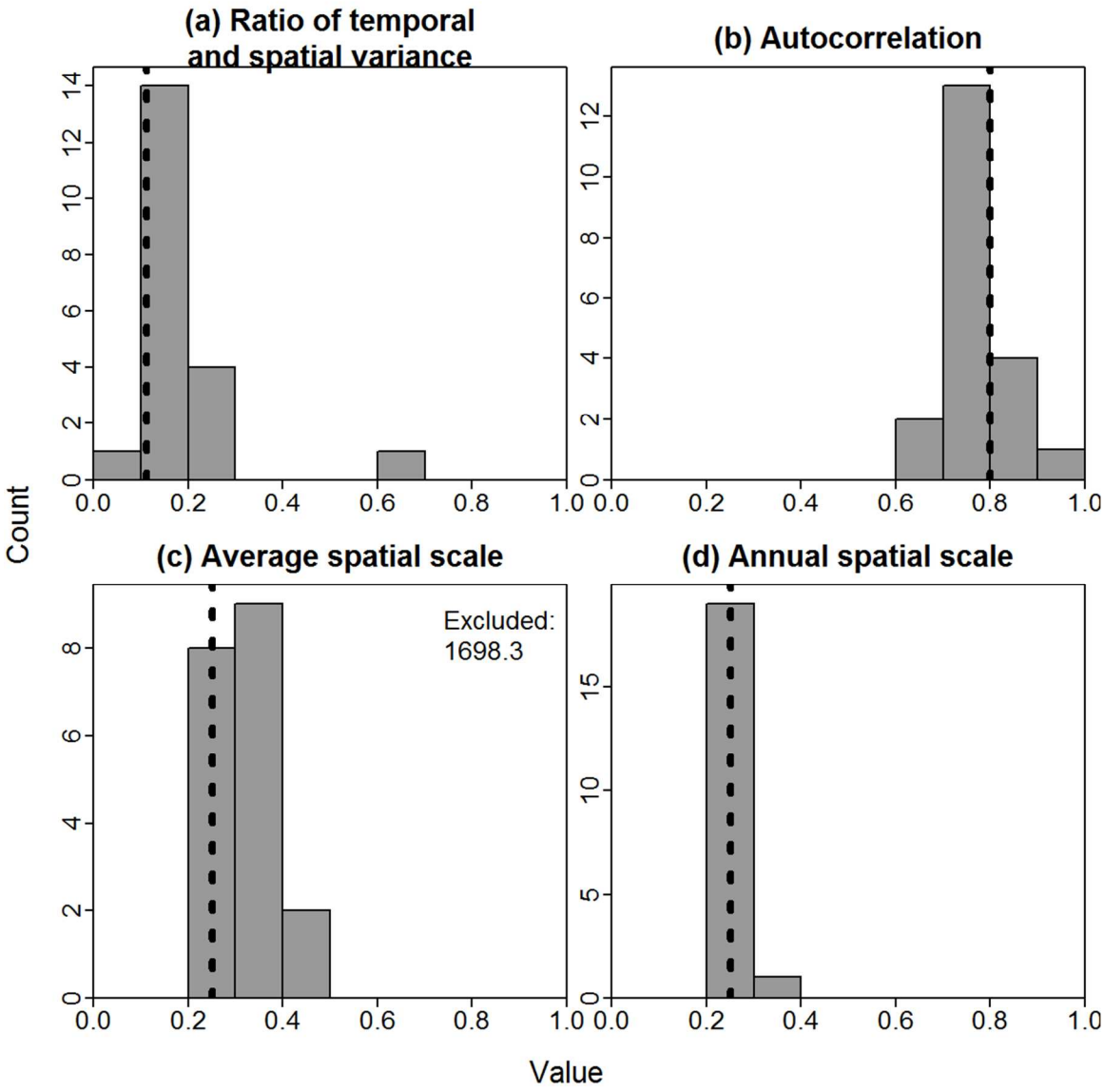
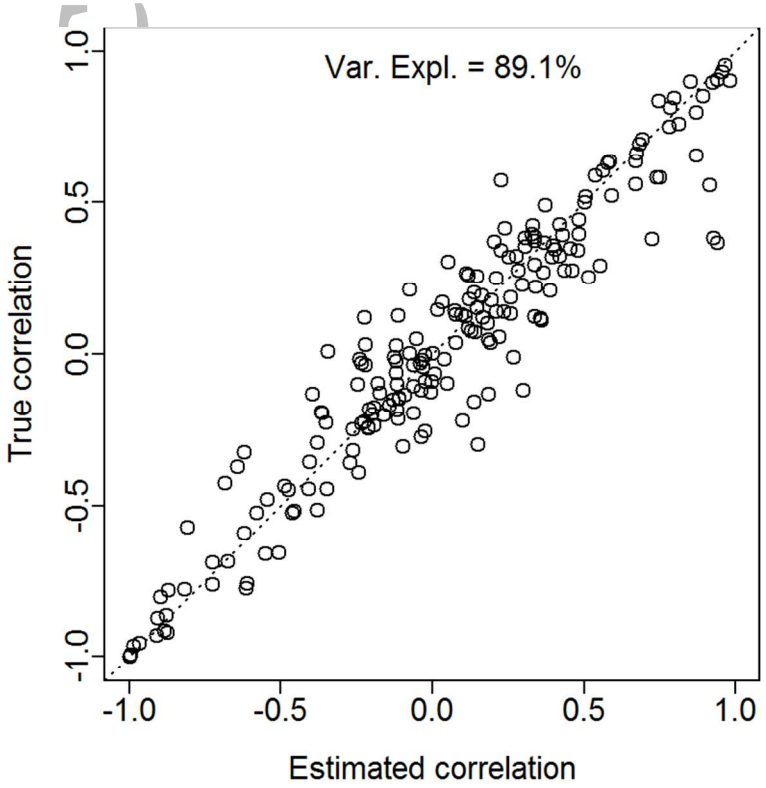


Fig. 5 – A scatterplot showing the estimated correlation in density among species, $Corr(\psi_i, \psi_j)$ for species *i* and *j*, and the true correlation, where each is calculated from the estimated and true loadings matrices $L_{p,j}$, respectively. The proportion of variance explained when regressing the true correlation on the estimated correlation is shown at the top.



Appendix S1 – Linear parameterization for each dynamic factor

Each factor is assumed to exhibit simple density-dependent (Gompertz) dynamics over time (Thorson *et al.*, 2015b). This amounts to a first-order autoregressive process at each site over time. Because dynamics are linear, they can be solved explicitly for any site and time:

$$\psi_j(n,t) = \phi_j \rho_j^t + \varepsilon_j(n,t) + \frac{\omega_j(n)}{1-\rho_j}$$

where $\varepsilon_j(n,t)$ now represents variation in the dynamics of factor ψ_j at site n between times t and $t+1$:

$$\boldsymbol{\varepsilon}_j \sim \mathrm{MVN}(\mathbf{0}, \boldsymbol{\Sigma}_\xi \otimes \Sigma_{AR(1)})$$

where

$$\mathrm{Var}\left(\varepsilon_j(s,t), \varepsilon_j(s+h, t+\Delta t)\right) = \rho_j^{\Delta t} f(|h|, \kappa_\xi, \tau_\xi^2)$$

such that $\Sigma_{AR(1)}$ is defined as the covariance for first-order autoregressive process, where the magnitude of autoregression ρ_j is identical to the strength of density dependence.

Appendix S2 – Identifiability conditions

Parameters for the generic spatial dynamic factor analysis model are not uniquely identifiable for two reasons. First, predicted log-densities $\theta_p(n,t)$ remain unchanged whenever the columns of the matrix of dynamic factors $\boldsymbol{\Psi} = (\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_j)$ and rows of the loadings matrix **L** are relabeled. Elsewhere, this phenomenon is termed "label switching", and we note that the ordering of factors is arbitrary and can be changed. Second, predicted densities are unchanged when the loadings matrix and **L** and dynamic factors are rotated by a rotation matrix (i.e., non-singular linear transformation **H** with determinant det(**H**)=1):

$$\psi_j^*(n,t) = \sum_{j2} H_{j,j2}\psi_{j2}(n,t)$$

$$L_{p,j}^* = \sum_{j2} L_{p,j2}(\mathbf{H}^{-1})_{j2,j}$$

We therefore ensure that estimated parameters are uniquely identifiable (except for label switching) by specifying that the upper-right corner of the loadings matrix **L** is fixed at zero. This constraint is analogous to a similar constraint in dynamic factor analysis (Harvey, 1990; Zuur *et al.*, 2003). Third, predicted densities remain unchanged whenever factors and loadings are transformed by any "scaling" matrix **H** (i.e., a diagonal and intertible matrix H, diag(**H**)=**h**, where all $h_j \neq 0$). In this case, the model will generally seek a solution in which $\mathrm{Var}[\boldsymbol{\psi}_j] \to 0$ and $\mathrm{Var}[\mathrm{L}_{p,j}] \to \infty$. We therefore specify that $\mathrm{Var}[\boldsymbol{\psi}_j] = 1$ for each dynamic factor *j* (Eq. 1), following an analogous constraint for spatial factor analysis (Thorson *et al.*, 2015a-a) and dynamic factor analysis (Harvey, 1990; Zuur *et al.*, 2003). The variance for each dynamic factor can be calculated as follows:

$$\mathrm{Var}\left(\boldsymbol{\psi}_j\right) = \mathrm{Var}\left(\phi_k\rho_j^t + \boldsymbol{\varepsilon}_j + \frac{\boldsymbol{\omega}_j}{1-\rho_j}\right) = \mathrm{Var}\left(\phi_k\rho_j^t\right) + \mathrm{Var}\left(\boldsymbol{\varepsilon}_j\right) + \mathrm{Var}\left(\frac{\boldsymbol{\omega}_j}{1-\rho_j}\right)$$

where this reduces to:

$$\mathrm{Var}[\boldsymbol{\psi}_j] = \sigma^2_{time,j} + \sigma^2_{space,j} = \frac{1}{4\pi\tau^2_{\varepsilon,j}\kappa_j{}^2(1-\rho_j{}^2)} + \frac{1}{4\pi\tau^2_{\omega,j}\kappa_j{}^2(1-\rho_j)^2}$$

We additionally define a parameter $\chi_j \equiv \frac{\sigma^2_{time,j}}{\sigma^2_{space,j}}$, such that:

$$\tau_{\omega,j} = \left(\frac{1+\chi_j}{4\pi\kappa_j{}^2(1-\rho)^2}\right)^{1/2}$$

$$\tau_{\varepsilon,j} = \tau_{\varepsilon,j}(1-\rho_j)\left(\chi_j(1-\rho_j{}^2)\right)^{-1/2}$$

In summary, we calculate a parameter χ_j for each dynamic factor j, which represents the relative magnitude of spatiotemporal and spatial variance, and other parameters are uniquely defined given the constraint that the marginal variance of each dynamic factor is one (up to the effect of label switching).

Appendix S3 -- Transformation of loadings matrix and factors to aid interpretation

Given the identifiability conditions listed in Appendix S2, estimates of the dynamic factors $\boldsymbol{\Psi}$ and the loadings matrix $\mathbf{L}$ are uniquely identifiable. However, we use a post-hoc transformation of the factors and loadings matrix to aid interpretation. We calculate a linear transformation matrix $\mathbf{H}$, and then interpret $\mathbf{L}'=\mathbf{LH}$ and $\boldsymbol{\Psi}'=\boldsymbol{\Psi}\mathbf{H}^{-1}$. Any transformation matrix $\mathbf{H}$ is permissible. However, we additionally specify that $\mathbf{H}$ is an orthogonal transformation, such that $\mathbf{H}\mathbf{H}^{T}=\mathbf{I}$. This property implies that:

$$\mathrm{Cov}(\boldsymbol{\Theta}') = \mathbf{LH}(\mathbf{LH})^{T} = \mathbf{LHH}^{T}\mathbf{L}^{T} = \mathbf{LL}^{T} = \mathrm{Cov}(\boldsymbol{\Theta})$$

such that transformation leaves the explicit computation of covariance between species densities unchanged. We seek a transformation that results in factor 1 explaining the greatest possible proportion of Cov($\boldsymbol{\Theta}$), factor 2 explaining the greatest possible proportion of Cov($\boldsymbol{\Theta}$) after controlling for factor 1, etc. These two criteria are achieved by defining:

$$\mathbf{L}' = \mathbf{LH} = (e_1^{0.5}\mathbf{v}_1, ..., e_n^{0.5}\mathbf{v}_n)$$

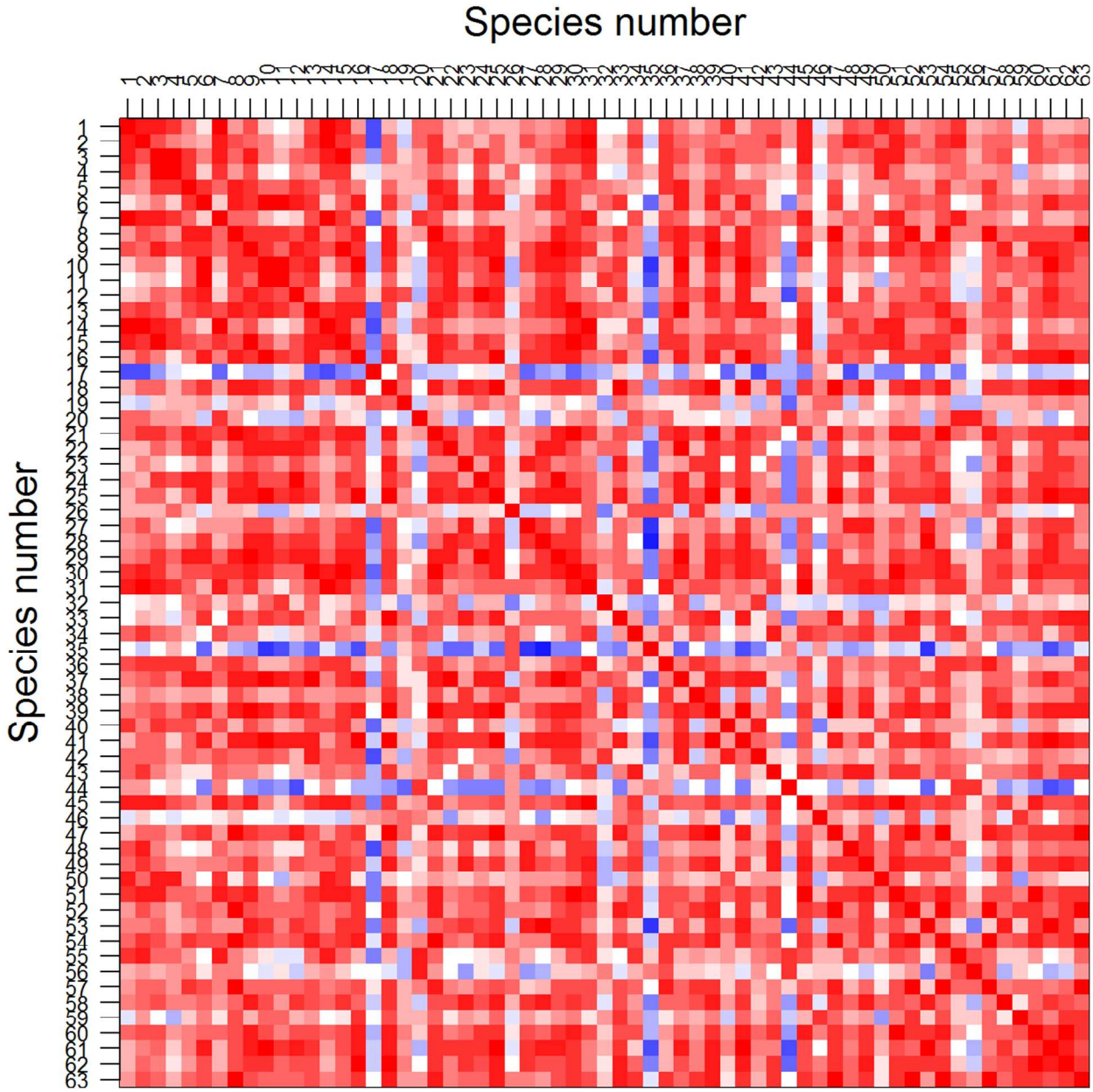
where $\mathbf{L}'$ is the loadings matrix after transformation, $\mathbf{H}$ is the transformation matrix, e_1 and $\mathbf{v}_1$ are the first eigenvalue and eigenvector of Cov($\boldsymbol{\Theta}$), and n is the number of estimated factors. Linear predictions of log-density $\boldsymbol{\Theta}$ are unchanged after transformation of the factors:

$$\boldsymbol{\psi}'(n,t) = (e_1^{0.5}\mathbf{v}_1, ..., e_n^{0.5}\mathbf{v}_n)^{-1}\mathbf{L}\boldsymbol{\psi}(n,t)$$

where $(e_1^{0.5}\mathbf{v}_1, \ldots, e_n^{0.5}\mathbf{v}_n)^{-1}$ is the Moore-Penrose pseudoinverse of $(e_1^{0.5}\mathbf{v}_1, \ldots, e_n^{0.5}\mathbf{v}_n)$. A quick check confirms that this transformation is orthogonal and results in identical predictions of log-density $\boldsymbol{\Theta}$ as well as covariance Cov($\boldsymbol{\Theta}$). Therefore, we refer to the loadings matrix and dynamic factors after transformation unless otherwise noted.

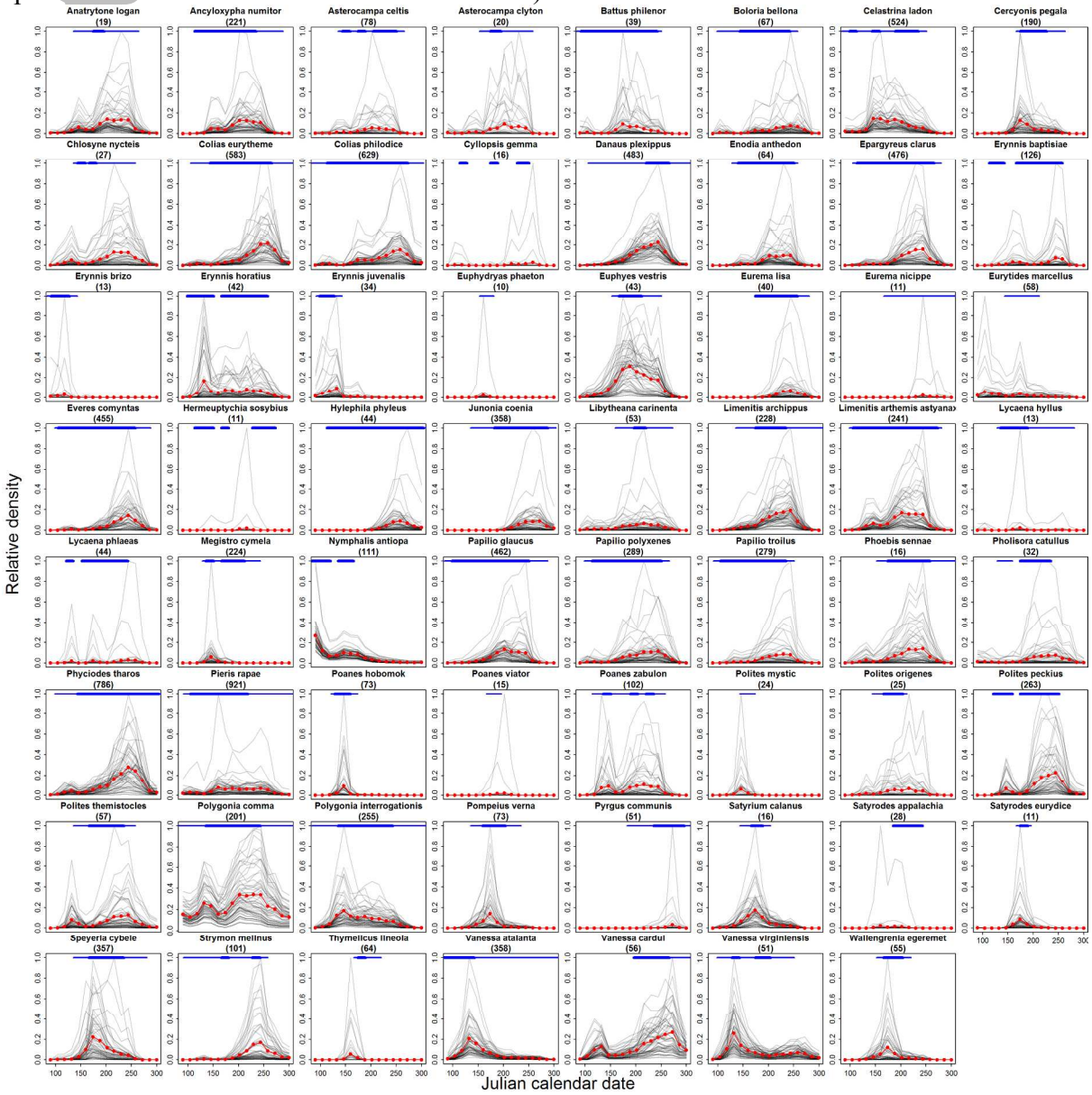
Appendix S4 – Estimated correlation among species in Case Study #2

Estimated correlation in spatiotemporal densities among species, $Corr(\psi_i, \psi_j)$ for species i and j, in the second case study ("Ohio butterflies") where species are labelled by number (see Table 1 for names), red is a positive correlation (with a correlation of 1.0 on the diagonal) and blue is a negative correlation.



Appendix S5 – Validation of estimated flight curves by comparison with independent reports

Supplementary Figure S1 – A comparison of estimated and previously reported flight curves for each butterfly species (see Table 1 for list of species). For each species, we show estimated density for each site (thin black lines) and overall (average among sites: red line). We also display independent reports of emergence times (thin blue lines) along with times of peak emergence (thick blue lines) for each species. These emergence times are based on reported phenograms, obtained from an independent data set of surveys throughout Indiana as published in Belth (2012). Indiana is adjacent to Ohio and in an ecologically similar zone, spans a similar latitudinal gradient, and has a similar butterfly community (where all 63 of the most common species seen in Ohio are also found in Indiana).



Bibliography

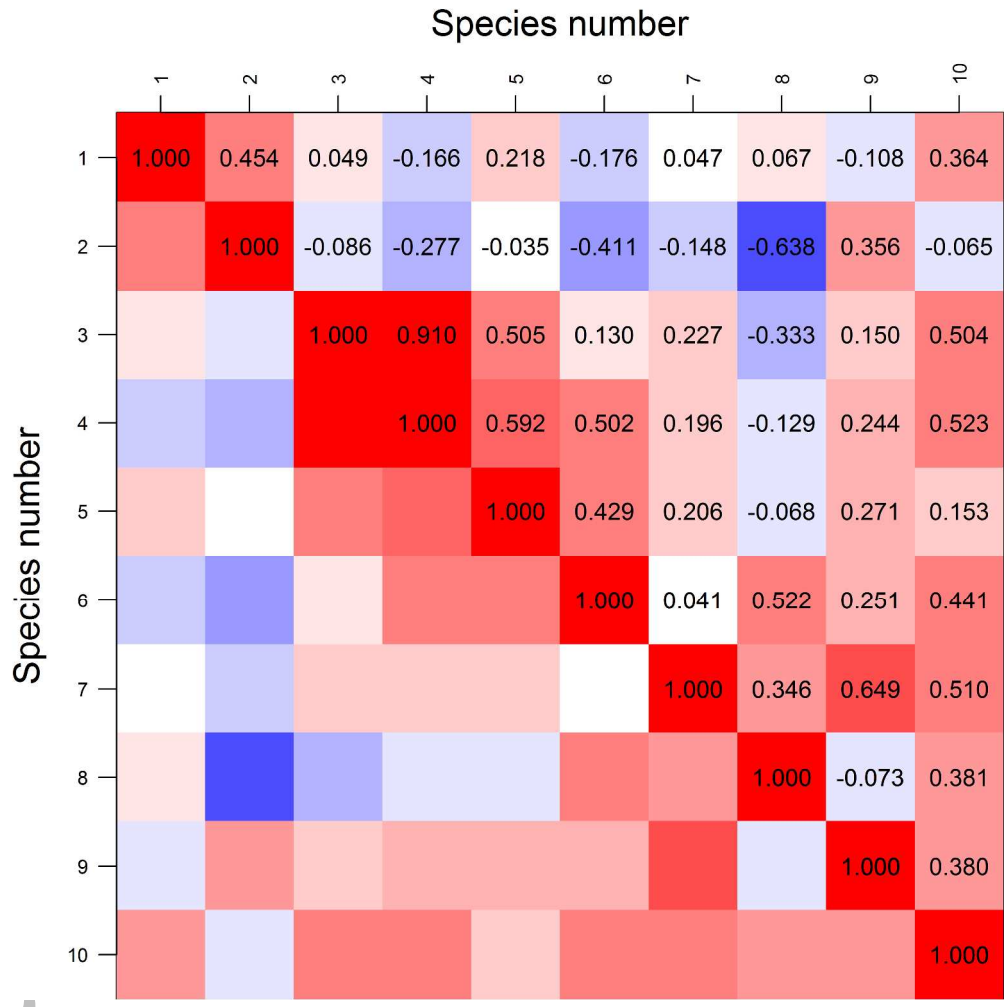
Belth, J.E. (2012) *Butterflies of Indiana: a field guide*, Indiana University Press, Bloomington, Indiana.

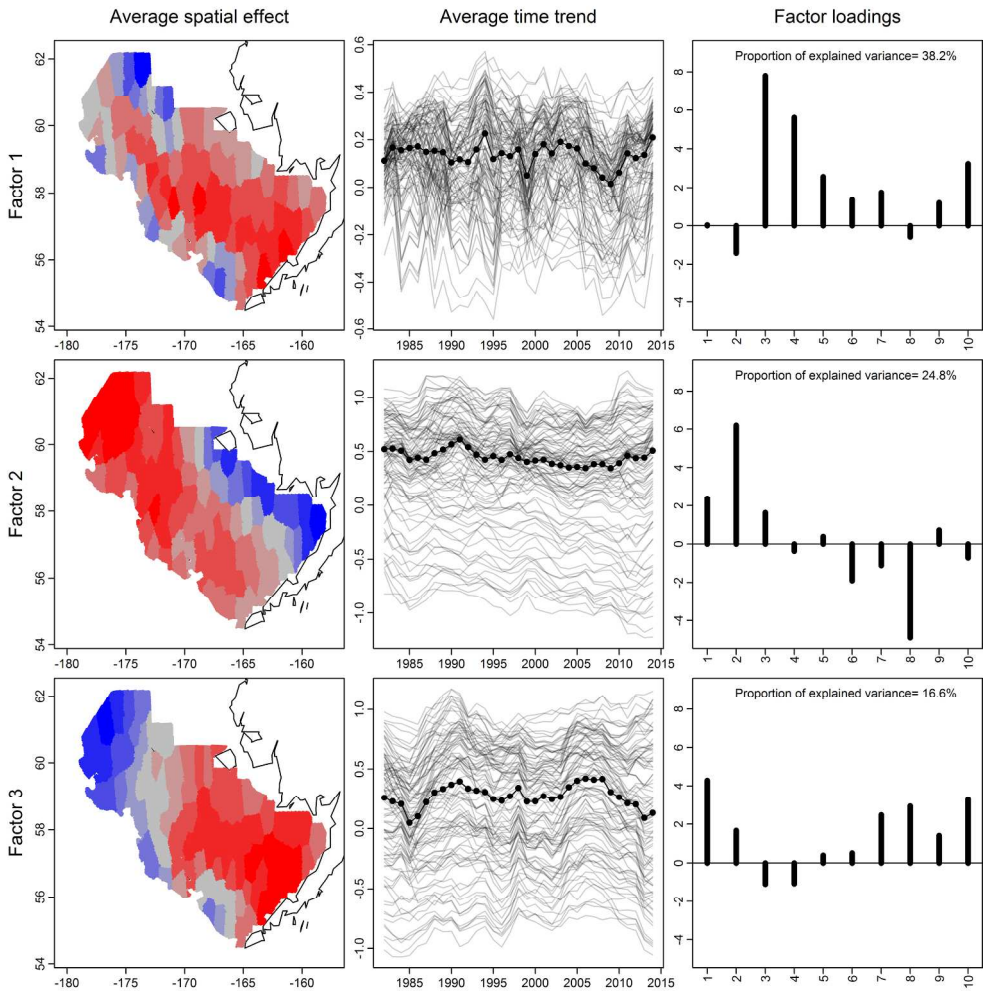
Harvey, A.C. (1990) *Forecasting, structural time series models and the Kalman filter*, Cambridge university press, Cambridge, UK.

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

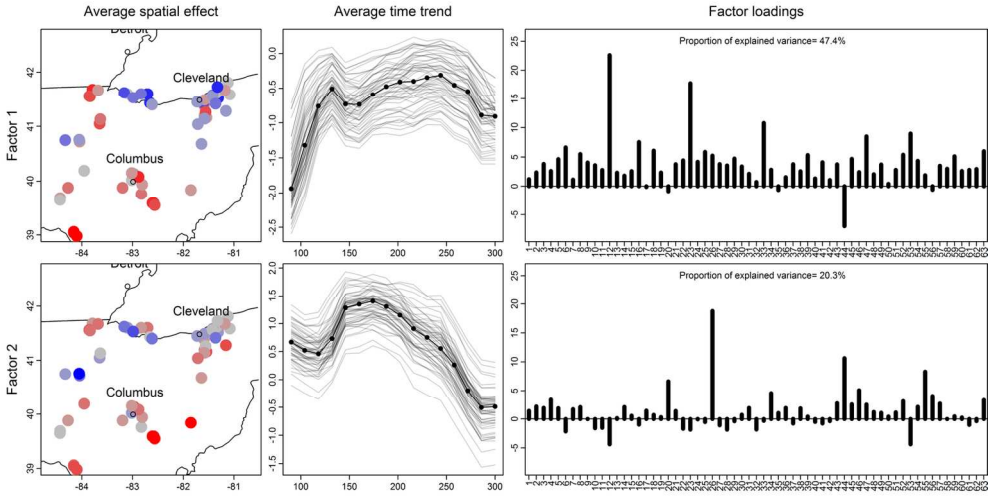
Thorson, J.T., Skaug, H.J., Kristensen, K., Shelton, A.O., Ward, E.J., Harms, J.H. & Benante, J.A. (2015b) The importance of spatial models for estimating the strength of density dependence. *Ecology*, **96**, 1202–1212.

Zuur, A.F., Tuck, I.D. & Bailey, N. (2003) Dynamic factor analysis to estimate common trends in fisheries time series. *Canadian Journal of Fisheries and Aquatic Sciences*, **60**, 542–552.

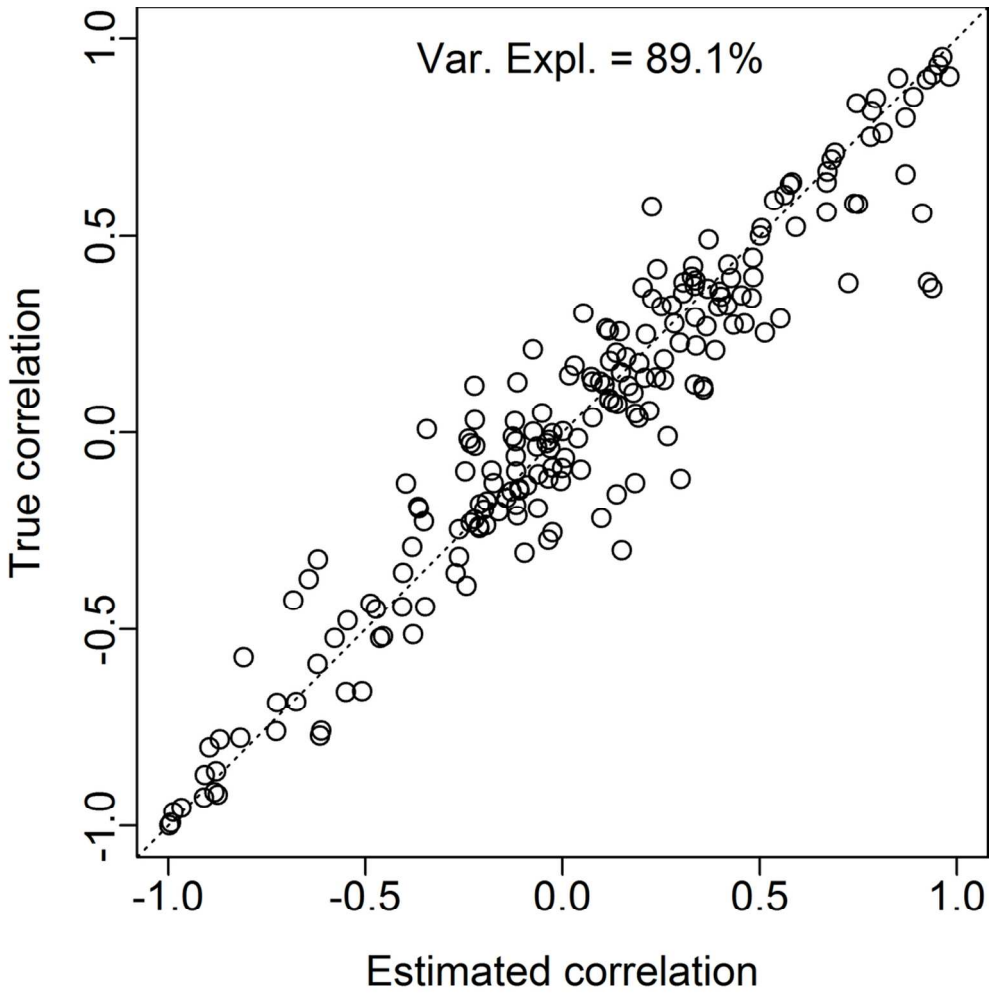




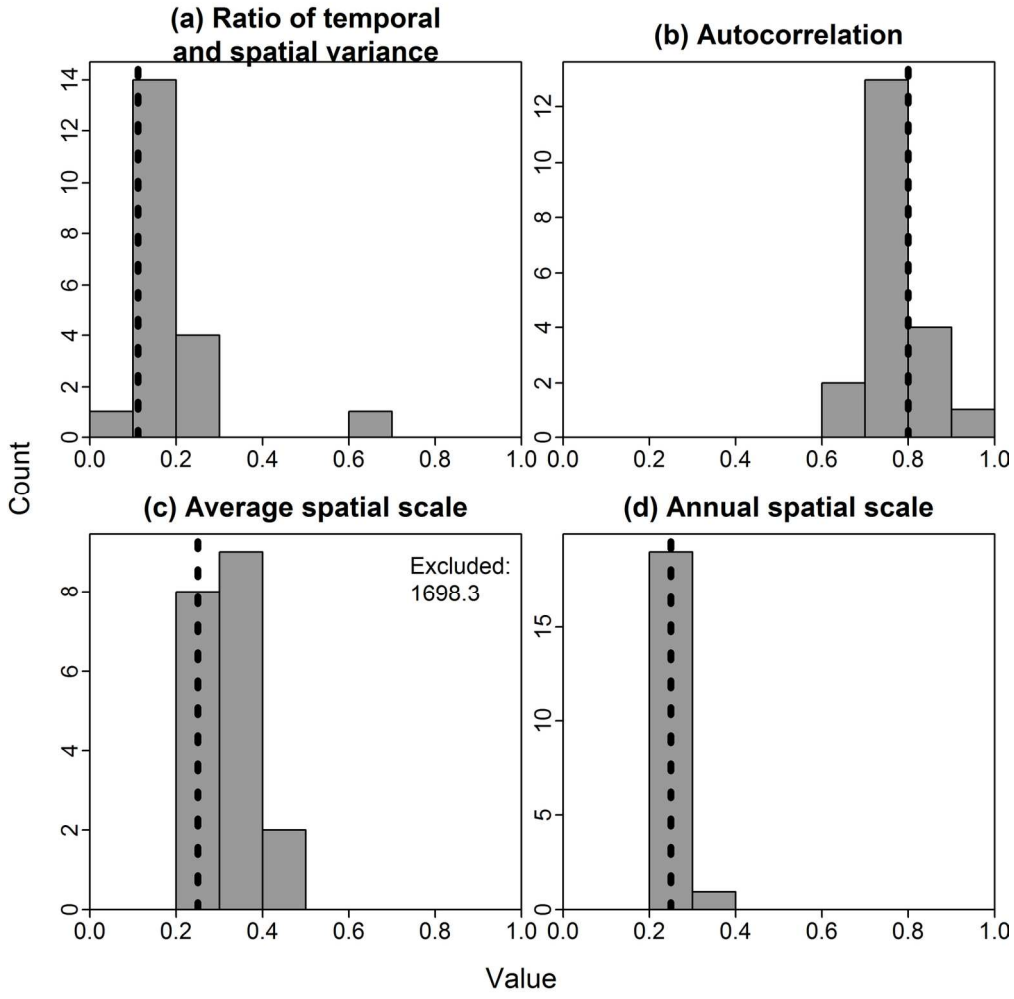
228x228mm (300 x 300 DPI)



152x76mm (300 x 300 DPI)



101x101mm (300 x 300 DPI)



152x152mm (300 x 300 DPI)