Hierarchical Spatiotemporal Matrix Models for Characterizing Invasions

Mevin B. Hooten,[1,*] **Christopher K. Wikle,**[2] **Robert M. Dorazio,**[3] **and J. Andrew Royle**[4]

[1]Department of Mathematics and Statistics, Utah State University, Logan, Utah 84322-3900, U.S.A.
[2]Department of Statistics, University of Missouri, Columbia, Missouri 65211, U.S.A.
[3]USGS Florida Integrated Science Center and Department of Statistics, University of Florida,
Gainesville, Florida 32611-0339, U.S.A.
[4]USGS Patuxent Wildlife Research Center, 12100 Beech Forest Road, Laurel, Maryland 20708, U.S.A.
[*]*email:* mevin.hooten@usu.edu

SUMMARY. The growth and dispersal of biotic organisms is an important subject in ecology. Ecologists are able to accurately describe survival and fecundity in plant and animal populations and have developed quantitative approaches to study the dynamics of dispersal and population size. Of particular interest are the dynamics of invasive species. Such nonindigenous animals and plants can levy significant impacts on native biotic communities. Effective models for relative abundance have been developed; however, a better understanding of the dynamics of actual population size (as opposed to relative abundance) in an invasion would be beneficial to all branches of ecology. In this article, we adopt a hierarchical Bayesian framework for modeling the invasion of such species while addressing the discrete nature of the data and uncertainty associated with the probability of detection. The nonlinear dynamics between discrete time points are intuitively modeled through an embedded deterministic population model with density-dependent growth and dispersal components. Additionally, we illustrate the importance of accommodating spatially varying dispersal rates. The method is applied to the specific case of the Eurasian Collared-Dove, an invasive species at mid-invasion in the United States at the time of this writing.

KEY WORDS: Animal dispersal; Detection probability; Hierarchical Bayesian models; Invasive species; Population models; Spatiotemporal dynamics.

1. Introduction

1.1 *Modeling Invasive Species*

The growth in population and dispersal of biotic organisms as a function of time has been recognized as an important subject throughout the relatively short history of ecology as a science (Bullock, Kenward, and Hails, 2002). Ecologists have long been able to accurately describe survival and fecundity in plant and animal populations (Renshaw, 1991, Chapter 1; Bullock et al., 2002) and have also developed quantitative model-based approaches to study the dynamics of dispersal (Clark et al., 2003; Nathan et al., 2003). Of particular interest are the dynamics of invasive species (Kolar and Lodge, 2001). Such nonindigenous animals (and plants) can levy significant impacts on native biotic communities (Elton, 1958; Shigesada and Kawasaki, 1997, Chapter 1). Multiple approaches have been proposed to model invasions, many of which stem from early work by Fisher (1937) and Skellam (1951) in diffusion models and biological waves (Shigesada and Kawasaki, 1997, Chapter 3). These led ultimately to more complex metapopulation models, matrix models, and state-space models (e.g., Hanski, 1999, Chapter 12; Caswell, 2001, Chapter 4; Borchers, Buckland, and Zucchini, 2002, Chapter 13). Often such models have been considered in the deterministic

sense or considered with known stochastic components for purposes of simulation (Renshaw, 1991, Chapter 2; Kot, 2001, Chapter A; Turchin, 2003, Chapter 5; Hastings et al., 2005). Statistical estimation of parameters in such models has conventionally consisted of likelihood-based methods, linear regressions, time-series methods, and more recently, hierarchical approaches (Shigesada and Kawasaki, 1997; Caswell, 2001; Calder et al., 2003; Buckland et al., 2004; Thomas et al., 2005).

Recent advances in computational efficiency have allowed for the implementation of sophisticated hierarchical Bayesian models (e.g., Calder et al., 2003; Clark, 2003; Wikle, 2003; Buckland et al., 2004; Su, Peterman, and Haeseker, 2004; Thomas et al., 2005; Wikle and Hooten, 2006; Hooten and Wikle, 2007). It has been shown that such models have much to offer, including more precise and less biased parameter estimation (Calder et al., 2003), accounting for multiple sources of uncertainty (Wikle, 2003), and the ability to describe nonlinear spatiotemporal dispersal and growth in relative abundance (Hooten and Wikle, 2007), as well as dynamics of population demographics (Buckland et al., 2004). Many previous studies have been concerned with the rate of spread only (e.g., Arim et al., 2006). However, models specifically

formulated for describing the spread of invasive species explicitly on spatio-temporal domains have relied on a state-space framework for implementing discretized partial differential equations to model a log-transformed Poisson intensity that represents "relative" abundance (Wikle, 2003; Wikle and Hooten, 2006; Hooten and Wikle, 2007). Although proven to be informative and useful, these formulations lack an intuitive appeal, as parameter estimates arise from a latent, log-transformed process that may also induce a growth model misspecification (Hooten and Wikle, 2007). Conjugate full-conditional distributions highlight the primary mechanical advantage of these models in that large multivariate parameter updates can easily be obtained from a Markov chain Monte Carlo (MCMC) algorithm. The combination of non-conjugacy, non-Gaussianity, and high-dimensional parameter spaces presents significant implementation challenges. At the same time, such model specifications allow for a more intuitive setting where identification and estimation of posterior parameter distributions comes more naturally.

So-called matrix models (i.e., first-order Markov models) offer a very flexible framework for describing population dynamics in time, space, and age classes (Caswell, 2001). Much can be learned about population dynamics through matrix model simulations. Neubert and Caswell (2000) provide an extensive discussion of demographic matrix models and the effects of demographics on dispersal; however, they do not estimate parameters on large spatiotemporal domains explicitly from a model-based statistical perspective. Conventional parameter estimation in such models requires extensive datasets from well-designed experiments (Caswell, 2001, Chapter 6). Although large-scale spatiotemporal data are becoming more readily available, they are often subject to multiple sources of uncertainty (e.g., observer bias and irregularly sampled locations and times) and not specifically collected to accommodate complicated space–time demographic models (Hastings et al., 2005).

Although much more general as a class of models, matrix models can be specifically employed to model the dynamics of invasive species in space and time and, therefore, such models are the focus of this paper. We approach the problem hierarchically, with an intuitive data model, flexible matrix process model comprised of growth and dispersal components, and parameter model consisting mostly of non-conjugate priors. This approach has the advantage of appearing less complex than those proposed previously (e.g., Wikle and Hooten, 2006; Hooten and Wikle, 2007). The general form of the spatiotemporal matrix model, however, is suggested by discretized partial differential (Hooten and Wikle, 2007) and integro-difference equations (Wikle, 2002), yet the direct matrix model approach proposed here allows for more general dispersal.

In most ecological studies, the state variable of interest cannot be observed directly because of bias related to imperfect detection. In fact, studies ignoring such sampling assumptions generally focus on modeling "relative abundance" only (e.g., Wikle, 2003; Hooten and Wikle, 2007). While models for relative abundance are useful, it is often necessary to make direct inference about the size of a population (i.e., numbers of organisms) at a given set of locations. Monitoring efforts for invasive species often result in count data, that is, the number of individuals observed at each location (or area). In cases where less than the total number of individuals is observed (i.e., imperfect detection), a binomial data model is suggested. In general, nonidentifiability issues arise when considering such models (Olkin, Petkau, and Zidek, 1981; Carroll and Lombard, 1985; Raftery, 1988). For cases where prior information about the probability of detection is available, we propose a data model with identifiable population size parameters that accounts for inherent uncertainty associated with nondetection bias.

1.2 *The Eurasian Collared-Dove in the United States*

Large scale spatiotemporal ecological datasets consisting of quantitative population information are not common, especially for invasive species (Caswell, 2001; Hastings et al., 2005). Fortunately, long-term monitoring efforts such as the North American Breeding Bird Survey (BBS; Robbins, Bystrak, and Geissler, 1986) exist and can provide high-dimensional space–time invasive species data (North American BBS data can be obtained at the following web site: `http://www.pwrc.usgs.gov/bbs/`).

Specifically, the Eurasian Collared-Dove (ECD, *Streptopelia decaocto*) is a nonindigenous species in North America for which data have been collected since its introduction in the early 1980s (Romagosa and Labisky, 2000). Although similar in appearance to the Ringed Turtle-Dove, ECD presence is not so benign as it rapidly spread through Europe in the 1900s and poses a possible threat to native ecosystems in North America (Hengeveld, 1993).

During the peak of the avian breeding season each year for most of the United States and Canada, skilled observers collect bird population data along roadside survey routes. Each survey route is 24.5 miles long with stops at 0.5-mile intervals. At each stop, a 3-minute point count is conducted, during which every bird seen or heard within a 0.25-mile radius is recorded. The spatial unit of observation for the routes along which data are collected presents a significant challenge to spatial modeling, thus, certain simplifying assumptions about the data must be considered in the construction of a spatiotemporal modeling framework. Herein we assume the data occur as counts located at the route centers; at a continental scale, we assume the effect of aggregating over routes will be minimal. Note that because of this aggregation assumption, any parameters corresponding to route locations should be interpreted with the original spatial unit of observation in mind. Additionally, the BBS data are subject to multiple sources of uncertainty, especially within-site variability (Link et al., 1994; Sauer, Peterjohn, and Link, 1994). Knowledge of such intricacies in the data provides a significant motivation for the development of models that explicitly account for known sources of uncertainty, particularly at the data level.

ECD currently occupies the Eastern United States although the invasion is ongoing (see Figure 1). Models for relative abundance have suggested that the ECD displays typical invasive behavior (Hooten and Wikle, 2007). That is, the underlying dynamics of the invasion can be characterized by common density-dependent population growth and dispersal processes. Romagosa and Labisky (2000) speculated that the ECD would likely colonize the entire United States in a few
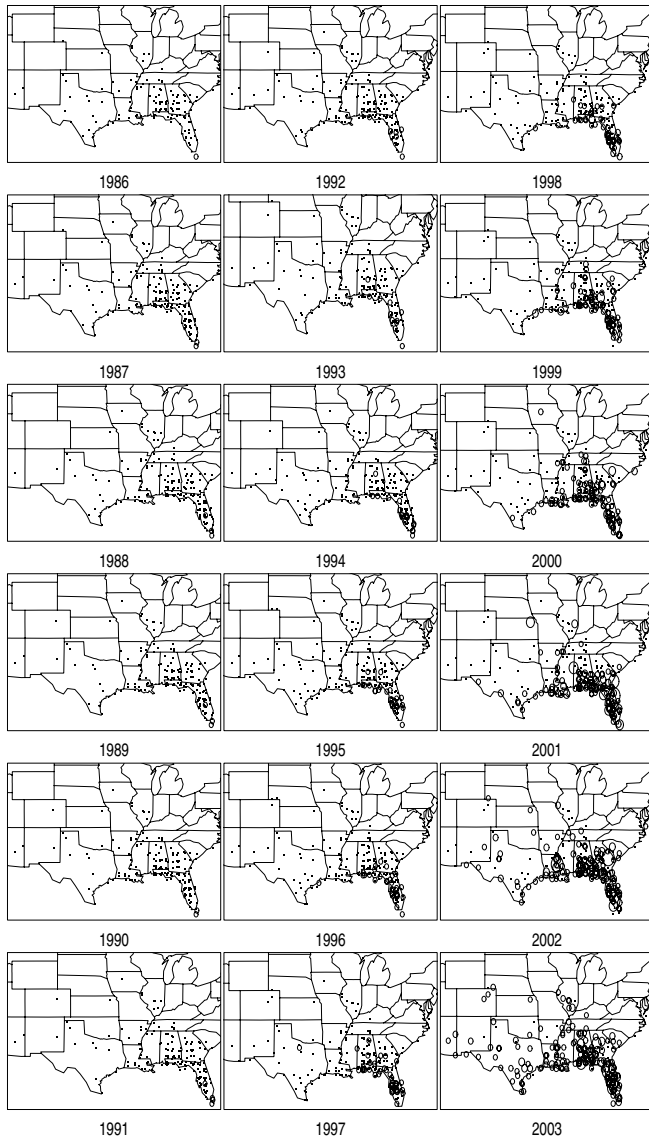
Figure 1. BBS counts for ECD at BBS route centers for 1986–2003. Points denote sample locations at which ECD was observed and circles illustrate magnitude of count data at each route location.

decades. The approach that follows allows for the rigorous probabilistic evaluation of such claims.

2. Methods

We adopt the general hierarchical model framework for spatiotemporal processes as in Wikle, Berliner, Cressie (1998), Calder et al. (2003), Wikle (2003), and Hooten and Wikle (2007), as proposed for time-series by Berliner (1996), whereby we construct the model in three stages: data, process, and parameters. The hierarchical approach allows us to construct and estimate a complex joint distribution via a sequence of simpler, and more intuitive, conditional distributions. This allows us to incorporate different forms of uncertainty within each level of the hierarchy. The square

bracket and conditional notation (i.e., $[a \mid b]$) is used hereafter for simplicity and denotes a conditional probability distribution. That is, $[a \mid b]$ implies the conditional probability distribution of a given b.

2.1 *Data Model*

In general, recording counts of individuals (or groups) at various locations over time is a common data collection scheme for monitoring plant and animal abundance. Because animals are usually detected imperfectly in most animal surveys, such data often represents a subset of the true, unknown, number of organisms at each time and location. If the assumption is made that each organism at a given time and location is observed independently *given* that the organism is present, then a binomial data model would be reasonable. Let $n_{i,t}$ represent the observed number of organisms at location i and time t for a total of m spatial locations and T times, while the true, unknown number of organisms at that location and time is denoted $n_{i,t}$. In the case of the ECD, the ith location refers to the location of the BBS route center, thus $n_{i,t}$ and $N_{i,t}$ refer to the observed and true number of ECDs aggregated over the ith BBS route, respectively. If the probability of observing (i.e., detecting) each organism is θ, then we would have (temporarily) the following data model specification:

$$n_{i,t} \mid N_{i,t}, \theta \sim \text{Binom}(N_{i,t}, \theta) \quad i = 1, \ldots, m; \quad t = 1, \ldots, T.$$

It would be preferable to use heterogeneous detection probabilities, yielding the data model: $n_{i,t} \sim \text{Binom}(N_{i,t}, \theta_{i,t})$. However, substantial data would be required to estimate such highly parameterized models in the absence of additional information. One feasible approach is to absorb much of the uncertainty regarding the probability of detection with a distribution on θ.

2.1.1 Estimation of detection probability. Although simple point counts are perhaps the most common sampling protocol employed in avian surveys, they present a problem in the estimation of $N_{i,t}$ when the detection probability (θ) is unknown and must be estimated as well (Royle, 2004). The two parameters are not identifiable in the absence of additional information. A minor extension of the sampling protocol can provide information on both θ and $N_{i,t}$, though its implementation specifically involving the ECD is not feasible using BBS observations due to the lack of replicate data. Thus, in order to estimate the true population size ($N_{i,t}$) we need to gain some a priori understanding of the probability of detection. Consider separately a sampling protocol wherein each local population (i.e., each sample unit) is visited on k separate occasions, yielding a sequence of independent binomial counts $n_{i,t}$ (for $t = 1, \ldots, k$). Assuming the true population size, N_i, remains fixed during the period required to collect k observations, the likelihood is the product binomial,

$$f(n_{i,1}, \ldots, n_{i,k} \mid N_i, \theta)$$
$$= \prod_t \frac{N_i!}{(N_i - n_{i,t})! n_{i,t}!} \theta^{n_{i,t}} (1 - \theta)^{N_i - n_{i,t}}. \tag{1}$$

The general instability of the maximum likelihood estimate for N_i based on (1) is well known (Olkin et al., 1981; Carroll and Lombard, 1985; Raftery, 1988), and this deficiency may explain why the sampling protocol is not widely used in

practice. However, when this sampling protocol is applied at replicate spatial locations and additional model structure is imposed on the location-specific abundance parameters, the estimation problem becomes more tractable (Royle, 2004). For example, if $N_i \sim$ Poisson (λ), then λ and θ are well identified.

This approach can be used in the presence of replicate data to obtain a posterior distribution for θ under a Bayesian implementation of equation (1) (Royle and Dorazio, 2006). As previously mentioned, in the case of the ECD, the operational BBS survey does not yield replicate counts at each BBS sample location. However, a study conducted by Link et al. (1994) obtained replicate counts on a number of BBS routes within the ECD range. Specifically, we used replicate data from a different species (Mourning Dove, *Zenaida macroura*) that shares similar detectability characteristics with ECD to inform θ. The posterior for the detection probability of the Mourning Dove is the standard form of a beta distribution with parameters α_θ and β_θ. We do not expect the estimate of θ (in terms of α_θ and β_θ) to represent the detectability of the ECD exactly; however, based on similarities in dove behavior we can expect it to be a reasonable estimate, especially considering that we are allowing for the uncertainty in θ through the beta distribution.

2.1.2 Incorporation of detection probability. The information about detection probability based on the Mourning Dove data provides the best estimate of θ we could hope to attain given available data and methods. In fact, given the identifiability issues between $N_{i,t}$ and θ we cannot learn anything further about the probability of detection in a hierarchical spatiotemporal model without additional replicate data. However, by allowing the Mourning Dove estimate to be our prior for θ and then treating it as a nuisance parameter, we can integrate it out of our data model while accounting for its uncertainty (Berger, Liseo, and Wolpert, 1999). This suggests the following modified data model,

$$[n_{i,t} \mid N_{i,t}]$$
$$= \int_0^1 [n_{i,t} \mid N_{i,t}, \theta][\theta]d\theta \quad i = 1, \ldots, m; \quad t = 1, \ldots, T, \quad (2)$$

which analytically yields a beta-binomial data model when the prior distribution for θ is indeed specified as a beta distribution (i.e., $\theta \sim$ Beta $(\alpha_\theta, \beta_\theta)$):

$$n_{i,t} \mid N_{i,t}, \alpha_\theta, \beta_\theta \sim \text{Beta-Binom}(N_{i,t}, \alpha_\theta, \beta_\theta) \quad i = 1, \ldots, m;$$
$$t = 1, \ldots, T. \quad (3)$$

Hence, this beta-binomial model is used hereafter as the data model component in the hierarchical matrix model.

2.2 Process Model

Although the estimation, spatial prediction, and temporal forecasting of population size are of primary concern in this setting, the underlying ecological process is also of interest. That is, knowledge about the true process from which the data arise is useful for characterizing and studying the invasion. Many previously proposed methods split the process model into two stages where the first stage is a nonlinear (and hence nonconjugate) transformation of the second stage (a Gaussian state equation). This provides a computational advantage yet detracts from the intuitive nature of the process

itself. We propose a process model with intuitive construction while retaining scientifically meaningful dynamic behavior.

We adopt a natural and common model for the true number of individuals at a given location and time. That is, the true population $(N_{i,t})$ is assumed to be conditionally Poisson with intensity $(\lambda_{i,t})$ depending hierarchically on population growth and dispersal parameters.

$$N_{i,t} \mid \lambda_{i,t} \sim Pois(\lambda_{i,t}) \quad i = 1, \ldots, m; \quad t = 1, \ldots, T.$$

Assuming the Poisson intensity $(\boldsymbol{\lambda}_t \equiv [\lambda_{1,t}, \ldots, \lambda_{m,t}]')$ evolves according to first-order Markovian dynamics, we can utilize a matrix model framework (Caswell, 2001, Chapter 4). In the specific setting where we lack information about the age structure of a species, we have $\boldsymbol{\lambda}_t$, the state vector, being propagated by a transition matrix $\mathbf{H}$ in the following evolution equation.

$$\boldsymbol{\lambda}_t = \mathbf{H}\boldsymbol{\lambda}_{t-1}$$
$$= \mathbf{MG}\boldsymbol{\lambda}_{t-1}$$
$$= \mathbf{M}(\boldsymbol{\tau})\mathbf{G}(K, r, \boldsymbol{\lambda}_{t-1})\boldsymbol{\lambda}_{t-1} \quad t = 1, \ldots, T. \quad (4)$$

Notice in equation (4) that we have decomposed the transition matrix $\mathbf{H}$ into a product of two distinct $m \times m$ matrices corresponding to growth $(\mathbf{G})$ and movement $(\mathbf{M})$ of the population. This formulation allows the process, which is random through its parameters (i.e., $K, r, \boldsymbol{\tau}$, and $\boldsymbol{\lambda}_1$ are all probabilistically specified in the next section) and quite flexible, to evolve in a conventional dynamical system. Additionally, it provides a convenient and ecologically meaningful specification of the growth and dispersal parameters. Similar decompositions of the transition matrix $\mathbf{H}$ in settings where population demographics is also being studied have been discussed (e.g., Neubert and Caswell, 2000; Buckland et al., 2004), though the specification presented here differs from those used previously in order to accommodate the specific ECD dataset as well as prediction of a high-dimensional spatially explicit process.

The matrix controlling population growth $(\mathbf{G})$ can be parameterized many ways. We propose a diagonal form where population growth at each location is described by a common density-dependent formulation given by the Ricker growth equation (Turchin, 2003, Chapter 3). Although other growth equations exist and are commonly used (e.g., Beverton-Holt, Malthusian, Gompertz), the Ricker form of growth is flexible and allows for potentially chaotic behavior and oscillatory characteristics in population stability (Kot, 2001, Chapter A). Additionally, investigators wishing to perform inference on the form of growth model specifically could implement various process models separately and compare their effect using any number of model selection criteria (e.g., Deviance Information Criterion [DIC] or Bayes factors). Another approach would be to consider multiple growth models simultaneously using a mixture-based specification for the process model and reversible jump MCMC. In simulation, we have found the Ricker growth model, with random parameters, to be very flexible and capable of exhibiting behavior similar to that of other forms of theoretical population growth. Thus, here we focus on Ricker growth and let the ith diagonal element of $\mathbf{G}$ be defined as,

$$G(r, K, \lambda_{i,t}) = \exp\{r(1 - \lambda_{i,t}/K)\}, \quad t = 1, \ldots, T, \quad (5)$$

where K and r correspond to carrying capacity and growth rate, respectively. In cases where there are significant differences in the carrying capacity (K) and growth rate (r) across spatial locations, one could consider parameterizations where K and r are allowed to vary in space. Effective estimation of spatially varying growth rates may be difficult in models already heavily parameterized without informative prior knowledge. In the presence of data describing a nontransient process (i.e., a process where population size has reached carrying capacity at all locations) it may be possible to estimate spatially varying carrying capacities (K); in cases where a population is still growing, however, these parameters will be difficult to estimate.

The dispersal of organisms in space is described in equation (4) by the matrix **M**. Dispersal in this case is assumed to be characterized by a spatially varying Gaussian kernel. That is, the population in a given location at the current time is a spatially weighted average of the population at surrounding locations depending on the distances ($d_{i,j}$) between locations and dispersal parameters, $\boldsymbol{\tau}$. We let **M** be defined as:

$$\mathbf{M}(\boldsymbol{\tau}) = [(M_{i,j})]_{m \times m},$$
$$M_{i,j} \propto \exp\left\{-\frac{d_{i,j}^2}{\tau_j}\right\}, \tag{6}$$

where each row of **M** is constrained to sum to one with each value between 0 and 1, as motivated by discretized diffusion equations (e.g., Wikle, 2002; Hooten and Wikle, 2007).

Other parameterizations of this dispersal matrix could be considered. For example, an alternative specification of the dispersal matrix, say $\mathbf{M}^{(\mathrm{alt})}$, is one where each column (rather than row) of $\mathbf{M}^{(\mathrm{alt})}$ is constrained to sum to one with $0 \leq M_{i,j}^{(\mathrm{alt})} \leq 1$ for all i and j. This dispersal matrix is especially suited to the "multisite" situation where the $M_{i,j}^{(\mathrm{alt})}$ refer to transition probabilities. In such a situation, where individual animals are moving from one distinct "population" to another with a certain probability, it is sensible to think of the transition probabilities as: $M_{i,j}^{(\mathrm{alt})} = \mathrm{P}(\text{animal moves to location } i \,|\, \text{animal is currently in location } j)$. In a spatial setting, with data pertaining to irregularly observed locations on a landscape where there are open boundaries (as is typically the case in invasion problems), a dispersal model needs to have the flexibility that the quantity being dispersed can move out of the domain through the boundary. To help visualize the difference between **M** and $\mathbf{M}^{(\mathrm{alt})}$, consider the simplest case, where $\mathbf{G} = \mathbf{I}$ (i.e., no growth in population); here $\mathbf{M}^{(\mathrm{alt})}$ preserves the number of total organisms over time and only allows them to occur in distinct areas, whereas **M** does not account for every organism at each time and thus the total number of organisms may fluctuate over time even though there is no growth component in the model. This effect of parameterization implies that when $\mathbf{G} \neq \mathbf{I}$, the parameters in **M** and **G** may not be completely identifiable because they could both be contributing to the overall growth in population. However, in the situation considered here, with the ECD and BBS count data, the preservation of population size resulting from the "probability parameterization" of $\mathbf{M}^{(\mathrm{alt})}$ is not as important as the dispersal behavior, thus we retain the parameterization of **M** as in equation (6). Though the chosen specification for

M implies that the growth (i.e., K and r) and dispersal (i.e., $\boldsymbol{\tau}$) parameters may not be separable, simulations (detailed in Hooten, 2006) verify that for the spatial and temporal domains considered in this application, the contribution from **M** to overall growth is minimal. Still, rigorous interpretation of growth and dispersal parameters, separately, could be misleading, and thus it is best to interpret the effect of K, r, and $\boldsymbol{\tau}$ simultaneously. Here, the emphasis is on the prediction of true population size ($N_{i,t}$) in time and space and thus there is no need to interpret the parameters separately; however, we provide the posterior results of these parameters for the sake of completeness.

As with the growth parameters, one could specify this Gaussian kernel to be spatially homogeneous in cases where limited data exist. For the application considered here, a multivariate parameterization has been shown to accommodate regional differences in dispersal (Hooten and Wikle, 2007) when adequately estimated. The dispersal mechanism employed here is more flexible than those previously proposed; thus we would expect patterns of dispersal to be different. In fact, because our emphasis is on prediction and estimation of the true population size ($N_{i,t}$), by specifying a simpler model with the constraint that all τ_i are equal, we can illustrate the risk of assuming homogeneous dispersal. Additionally, we can compare model complexity and effectiveness using DIC (see Gelman et al., 2004, Chapter 6). Though no methods of model comparison and selection are without criticism, DIC focused on the most important level in complex hierarchical models (such as the data model in this case) can provide a means of comparing the overall effect of different parameterizations while taking into account model complexity.

2.3 *Parameter Model*

The underlying process (equation 4) in this hierarchical model is a dynamical system with random components. Conventionally, latent processes are formulated with additive error to account for misspecification of the model. Such specifications are very general and often result in nonidentifiable variance components (e.g., Hooten and Wikle, 2007) and prohibit long-range prediction due to additive error propagating through the predictive forward model. Accounting for the process model error through probabilistic growth and dispersal parameter specifications alleviates such complications while offering sensible flexibility in the latent process.

With the nuisance parameter, θ, integrated out, we propose a very simple specification for prior distributions of the remaining model parameters. Assuming that we have independence between priors (i.e., $[K, r, \boldsymbol{\tau}, \boldsymbol{\lambda}_1] \propto [K][r][\boldsymbol{\tau}][\boldsymbol{\lambda}_1]$), we let:

$$K \sim \mathrm{Gamma}(\alpha_K = 4, \beta_K = 50),$$
$$r \sim \mathrm{N}(\mu_r = 0.5, \sigma_r^2 = 0.25),$$
$$\log\begin{pmatrix}\boldsymbol{\tau} \\ \tilde{\boldsymbol{\tau}}\end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix}\boldsymbol{\mu}_\tau \\ \tilde{\boldsymbol{\mu}}_\tau\end{pmatrix}, \quad \boldsymbol{\Sigma}_\tau = \begin{pmatrix}\boldsymbol{\Sigma}_\tau^{1,1} \boldsymbol{\Sigma}_\tau^{1,2} \\ \boldsymbol{\Sigma}_\tau^{2,1} \boldsymbol{\Sigma}_\tau^{2,2}\end{pmatrix}\right), \tag{7}$$
$$\log\begin{pmatrix}\boldsymbol{\lambda}_1 \\ \tilde{\boldsymbol{\lambda}}_1\end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix}\boldsymbol{\mu}_\lambda \\ \tilde{\boldsymbol{\mu}}_\lambda\end{pmatrix}, \quad \boldsymbol{\Sigma}_\lambda = \begin{pmatrix}\boldsymbol{\Sigma}_\lambda^{1,1} \boldsymbol{\Sigma}_\lambda^{1,2} \\ \boldsymbol{\Sigma}_\lambda^{2,1} \boldsymbol{\Sigma}_\lambda^{2,2}\end{pmatrix}\right), \tag{8}$$

where $\tilde{\boldsymbol{\tau}}$ and $\tilde{\boldsymbol{\lambda}}_1$ are the parameter values at a set of locations where we are interested in making predictions. Also, the Gamma distribution on K has a mean 200 and variance of 10,000 and is parameterized such that β_K is the scale parameter. $\boldsymbol{\Sigma}_\tau$ and $\boldsymbol{\Sigma}_\lambda$ are characterized by exponential spatial hyperpriors on the covariance structure of the dispersal and intensity process at the initial time (i.e., $t = 1$), respectively. Specifically, $[\mu_\tau]_i = [\tilde{\mu}_\tau]_j = \log(2), \forall i, j$ and $[\boldsymbol{\Sigma}_\tau]_{i,j} = \log(1.1)[\exp(-2\,||d_{i,j}||)]$ represent the hyperpriors for $\boldsymbol{\tau}$ while $\mu_\lambda = \tilde{\mu}_\lambda = \log(15)$ for locations in extreme southern Florida (i.e., the location where ECD was introduced into the United States) with zero elsewhere and $[\boldsymbol{\Sigma}_\lambda]_{i,j} = \log(1.1)[\exp(-2\,||d_{i,j}||)]$ represent the hyperpriors for $\boldsymbol{\lambda}_1$. The Euclidean distance between locations i and j is represented by $||d_{i,j}||$. The spatial specification of $\boldsymbol{\tau}$ and $\boldsymbol{\lambda}_1$ allows for posterior predictive estimation at locations and times without data as well as likely autocorrelation in the parameters (see Web Appendix A). Hyperpriors are specified based on their ability to exhibit reasonable growth and dispersal behavior in simulations. For example, we know that the population size is generally growing over time and thus μ_r was specified to be positive, although r could still be negative depending on σ_r^2 and the influence of the data. Recall that prior specification for θ (i.e., $\theta \sim \text{Beta}\,(1.7, 7.7)$, implying a mean of 0.18 and standard deviation of 0.12) was based on posterior estimates of detection probability from a separate model and dataset with repeated measurements and then integrated out, as discussed in the data model section of this article.

3. Results

Methods for estimating the posterior distribution for the models considered here include MCMC and various forms of importance sampling (Gelman et al., 2004, Chapter 11). Importance sampling is especially desirable in settings with nonconjugate full-conditional distributions and nonlinear relationships between parameters; however, it suffers from degeneracy problems, especially in situations with high-dimensional parameter spaces, as is the case here. Metropolis–Hastings updating within a Gibbs Sampler (one form of MCMC) is an alternative approach for sampling, but requires careful selection of proposal distributions that can sufficiently explore the parameter space. We concentrate on the MCMC approach for estimation although other sampling algorithms may also provide adequate results.

Several multivariate Metropolis–Hastings parameter updates slow the mixing of MCMC samples, thus the Gibbs sampler was allowed to run for 200,000 iterations to ensure that the posterior parameter space was explored sufficiently. Resampling was used to remove correlation between realizations. Convergence, which was assessed by visually inspecting parameter chains, occurred before the burn-in period of 20,000 iterations, which were used for calculation of posterior summary statistics. Metropolis–Hastings acceptance rates for the parameter updates varied between 20% and 40%. For additional details about the sampling algorithm and predictions, see Web Appendix A.

Estimation of the posterior distributions of model parameters provides useful information about the underlying dynamics of the ECD invasion. Although the focus here is on the estimation and prediction of the population size at various locations and times, knowledge about the process driving the population size can also be informative. In this setting we can evaluate population dispersal as well as growth parametrically, though, inference on dispersal and growth parameters, separately, should only serve as a general guideline because the degree of separation is unknown. However, the combined effect of growth and dispersal is valid and can be assessed by interpreting the process itself (i.e., $\boldsymbol{\lambda}_t$). Simulated processes on similar spatial and temporal domains suggested that the shared contribution from K, r, and $\boldsymbol{\tau}$ to overall population size is negligible (as evidenced by only a slight bias in the posterior estimates of parameters based on simulated data); however, we provide the following posterior summaries with a cautionary warning.

Posterior distributions for the univariate parameters K (carrying capacity) and r (growth rate) yielded 95% credible intervals of [103, 201] and [0.26, 0.32], respectively. Posterior standard deviations for K and r (i.e., $(V(K\,|\,n))^{1/2} = 24.3, (V(r\,|\,n))^{1/2} = 0.02$) were very precise compared to their prior standard deviations (i.e., $(V(K))^{1/2} = 100, (V(r))^{1/2} = 0.5$), thus implying that there was sufficient data to inform the growth of the process.

Statistics from the posterior distribution of multivariate parameters can be expressed as spatial maps. Prediction and estimation at a set of regularly spaced locations provides a convenient way to view the posterior parameter spaces. Note that all figures containing maps herein are displayed as images with grid cell intensity representing parameter magnitude at
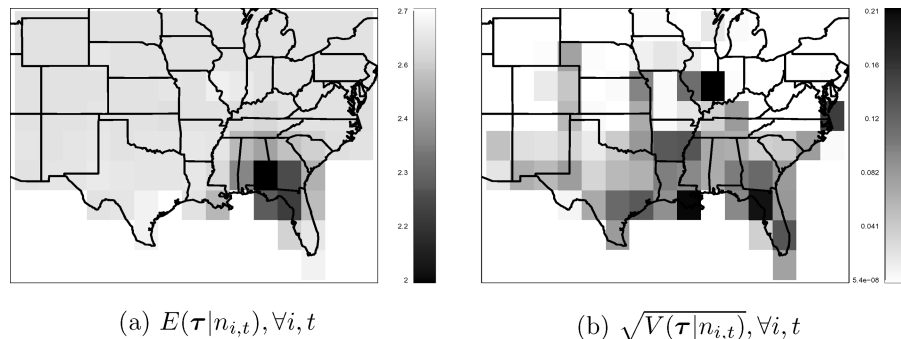

(a) $E(\boldsymbol{\tau}|n_{i,t}), \forall i, t$　　　　　　(b) $\sqrt{V(\boldsymbol{\tau}|n_{i,t})}, \forall i, t$

Figure 2. Maps representing the posterior mean ($E(\boldsymbol{\tau}|n_{i,t}), \forall i, t$) and standard deviation ($(V(\boldsymbol{\tau}|n_{i,t}))^{1/2}, \forall i, t$) of dispersal.

the location of the grid cell center. The image format is for visualization only (as an alternative to variable sized points) and should not be misinterpreted as "areal" support. That is, each pixel in the images represents a point at the pixel center, rather than the area of the pixel itself.

Consider the posterior mean and standard deviation maps of dispersal τ (Figure 2). As with the growth parameters, the posterior standard deviations for dispersal were much smaller than the prior standard deviation (i.e., $\max(V(\tau_i\,|\,n)) \ll V(\tau_i)$). A primary concern of ecologists and resource managers related to invasive species is the extent and magnitude of the invasion at a given time (including future times). In the case of the ECD, the mean posterior population size presented in Figure 3a illustrates the estimated magnitude and extent of the invasion at a set of times (including predictions) given the sample data. By the year 2016, the ECD population is predicted to be at or near carrying capacity for the portion of the United States shown in the figures. Posterior credible interval growth curves for individual locations can reveal at what time the population is projected to reach its carrying capacity (i.e., the total number of ECDs a sampled location can support).

By implementing a simpler form of the model with homogeneous dispersal rates (i.e., $\tau_i = \tau,\ \forall i$) we can compare and assess what, if any, advantage spatially varying dispersal contributes to estimation and prediction. Using DIC (e.g., Gelman et al., 2004, Chapter 6) focused on the data model, we can compare the two models in terms of their predictive capacity while penalizing for model complexity. Denote model 1 as the proposed model allowing for heterogeneous dispersal and model 2 as the simplified model with homogeneous dispersal. Then, the resulting number of effective parameters for each model is: $pD_1 = 133$ and $pD_2 = 130$. The DIC for each model is: $DIC_1 = 13{,}052$ and $DIC_2 = 13{,}107$.

4. Discussion

Consider, first, the posterior 95% credible interval for the Ricker growth rate (r). In situations where there is no change in population size due to dispersal, a growth rate of $r < 0$ implies a decrease in population while $r > 0$ implies an increase. Here r does not overlap zero, implying that there is indeed growth in the population, in addition to dispersal. In this case, where dispersal contributes to population size, and because an invasion generally constitutes a growing population, we can expect r to be positive but smaller than it would be in a population without dispersal. Also, recall from equation (5) that when $r \to 1$ it implies that $G(K,\ r,\ \lambda_{i,t}) \to \exp\{1 - \lambda_{i,t}/K\}$; a much simpler growth model. Here, the posterior credible interval of r suggests that it is significantly different from 1, thus the assumption of a two parameter growth equation is reasonable.

Recall that the carrying capacity is modeled with an overall term (K); thus, the estimate for $E(K\,|\,n)$ could be thought of as representing the overall mean while $V(K\,|\,n)$ represents the uncertainty associated with carrying capacity. The possible spatial heterogeneity in carrying capacity is not accounted for explicitly here for the sake of parsimony; situations may exist, however, where such a multivariate carrying capacity parameter is indeed identifiable. The posterior credible interval for
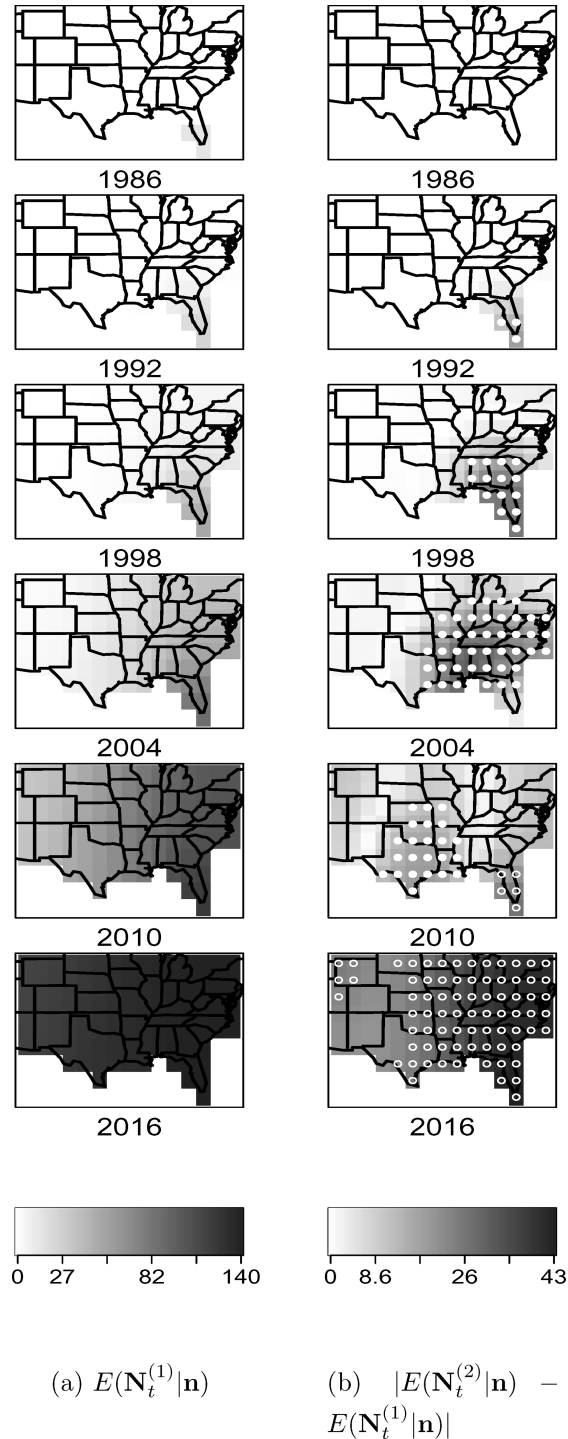

$$\text{(a) } E(\mathbf{N}_t^{(1)}|\mathbf{n}) \qquad \text{(b) } |E(\mathbf{N}_t^{(2)}|\mathbf{n}) - E(\mathbf{N}_t^{(1)}|\mathbf{n})|$$

Figure 3. (a) Posterior (and posterior predicted) mean of ECD population size throughout the United States from 1986 to 2016. (b) Also, the absolute difference between posterior means for $\mathbf{N}_t^{(2)}$ and $\mathbf{N}_t^{(1)}$ (i.e., the difference in estimates of population size for the model with homogeneous dispersal versus the model with heterogeneous dispersal). Overall shade corresponds to absolute difference and solid white circles represent large positive differences while empty white circles represent large negative differences.
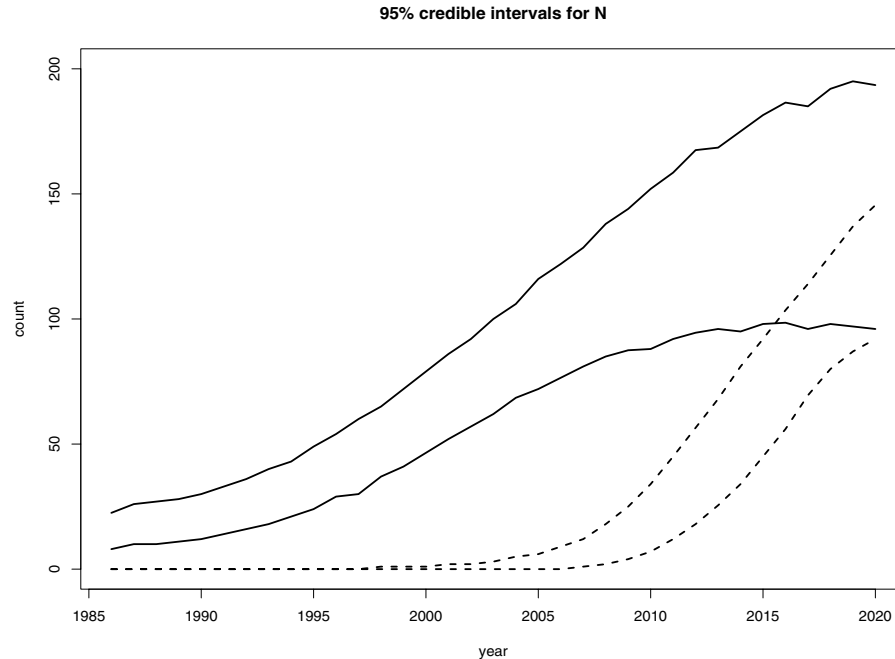
95% credible intervals for N

Figure 4. Credible intervals for posterior population growth of ECD ($N_{i,t}$) from 1986 to 2020 for locations in south Florida (solid lines) and northwest Wyoming (dashed lines) simultaneously.

the carrying capacity used here (K) suggests that the number of individuals a given location can support is quite variable. This is not surprising given the lack of prior information about this parameter and the fact that the invasion is still ongoing and has not reached saturation at all locations. For situations where the population size at more locations had reached carrying capacity, the variability in K could likely be reduced by conditioning on spatial covariates.

From the parameter estimates in Figure 2 we can see that there is a clear and significant area of low dispersal in the region of northern Florida (in the southeast portion of the map in Figure 2a). The effect of this will slow the spread of ECD westward out of Florida while northward, along the east coast, ECD dispersal remains uninhibited. Note that the values for the mean parameter estimates of τ are directly related to the scale of the map distance units, and therefore the general pattern rather than the magnitude should be interpreted here. That is, the use of a different coordinate system would result in different parameter values, but retain the same pattern.

Consider 95% credible intervals for a location in south Florida and a location in northwest Wyoming (i.e., Northwest United States) simultaneously (Figure 4). At the time of this writing, very few, if any, ECDs have likely dispersed as far as Wyoming, whereas the population sizes in most of Florida are nearly at carrying capacity. Although, by the year 2020, Wyoming will be near carrying capacity with high probability as well, as speculated by Romagosa and Labisky (2000). Such extrapolations (in time and space) are generally not advisable; however, the predictions in this case are a result of the estimated dynamics of the invasion and serve as the only insight into the probable population sizes at future times and locations based on the available data.

The DIC analysis for comparing the two models suggests that after correcting for the number of unconstrained parameters, model 1 still provides better predictions due to the lower DIC value. Additionally, Figure 3b illustrates the relative differences between models and shows how assuming homogenous dispersal can lead to inflated and deflated population size estimates in both time and space. In Figure 3b the areas with solid white circles imply that model 2 (homogeneous dispersal) is overestimating the population size relative to the predictions resulting from model 1, while the areas with empty white circles indicate that model 2 is underestimating population size relative to the predictions resulting from model 1. These areas of misestimation clearly evolve over time and can be quite misleading. Overall, the DIC analysis and difference maps suggest that a spatially varying dispersal parameter is indeed important in the model.

Overall, we have found that by using a priori knowledge gained from models proposed by Royle (2004) and Royle and Dorazio (2006), it is possible to estimate parameters that characterize a nonlinear and non-Gaussian dynamical system underlying the dispersal of invasive species. This use of subjective probability illustrates a strength of Bayesian models in complicated hierarchical settings where conventional likelihood-based parameter estimation would not be feasible. From a scientific perspective, it is important to note that previous dynamical spatiotemporal population models have largely ignored the detectability of the species, whereas here we are directly accounting for detection as well as uncertainty associated with detection. In the specific case of the ECD, this represents the only way to gain an understanding of the spatiotemporal dynamics in true population size (i.e., $N_{i,t}$).

The model proposed in this article adopts the intuitive and simple yet powerful form of a matrix model (e.g., Caswell, 2001). Although information about the age distribution of ECD over time and space is not available, this model framework could easily accommodate such data. In fact, several approaches have been taken to implement the demographic version of matrix models in a hierarchical framework (e.g., Buckland et al., 2004; Thomas et al., 2005).

It is important to note that this form of the proposed model relies heavily on the estimated dynamics of the invasion in order to predict population size at future times and unsampled locations. For cases where there exists more complete data for less of a generalist species, the model could easily be extended to accommodate covariate effects on model parameters, such as growth rate (r), carrying capacity ($\mathbf{K}$), and dispersal (τ). In cases where such information is available, predictions could also serve as a way to evaluate scenarios on how to control the spread of invasions. Many other extensions to the model are possible concerning detectability. For example, if detection was known (or estimated a priori) to vary by location or time (possibly through association with covariates), then it could be included in the exact same framework only with $\theta_{i,t}$ in the data model rather than θ. Furthermore, in cases where repeated counts are observed for each location and time, a modeling framework similar to that of Royle and Dorazio (2006) could be implemented to estimate $\theta_{i,t}$, in addition to the process and dynamics.

The main advantage of this model over other similar invasive species models is that it is intuitive and simple yet flexible enough to accommodate various specifications and sources of uncertainty (e.g., such as the uncertainty in detection). It provides tangible graphical output (in addition to numerical output) and is based on rigorous data-driven statistical methodology, the basic principles of which ecologists and resource managers can understand and utilize for scientific endeavor and decision making.

Supplementary Materials

Web Appendix A referenced in Sections 2.3 and 3 is available under the Paper Information link at the Biometrics web site `http://www.tibs.org/biometrics`.

References

Arim, M., Abades, S. R., Neill, P. E., Lima, M., and Marquet, P. A. (2006). Spread dynamics of invasive species. *Proceedings of the National Academy of Sciences, USA* **103,** 374–378.

Berger, J. O., Liseo, B., and Wolpert, R. (1999). Integrated likelihood methods for eliminating nuisance parameters. *Statistical Science* **14,** 1–28.

Berliner, L. M. (1996). Hierarchical Bayesian time-series models, In *Maximum Entropy and Bayesian Methods*, 15–22. Dordrecht, The Netherlands: Kluwer Academic Publishers.

Borchers, D. L., Buckland, S. T., and Zucchini, W. (2002). *Estimating Animal Abundance: Closed Populations*. London: Springer-Verlag.

Buckland, S. T., Newman, K. B., Thomas, L., and Koesters, N. B. (2004). State-space models for the dynamics of wild animal populations. *Ecological Modelling* **171,** 157–175.

Bullock, J. M., Kenward, R. E., and Hails, R. S. (2002). *Dispersal Ecology*. Malden, Massachusettes: Blackwell Science Ltd.

Calder, C., Lavine, M., Muller, P., and Clark, J. S. (2003). Incorporating multiple sources of stochasticity into dynamic population models. *Ecology* **84,** 1395–1402.

Carroll, R. J. and Lombard, F. (1985). A note on N estimators for the binomial distribution. *Journal of the American Statistical Association* **80,** 423–426.

Caswell, H. (2001). *Matrix Population Models*. Sunderland, Massachusettes: Sinauer Associates, Inc.

Clark, J. S. (2003). Uncertainty and variability in demography and population growth: A hierarchical approach. *Ecology* **84,** 1370–1381.

Clark, J. S., Lewis, M., McLachlan, J. S., and HilleRisLambers, J. (2003). Estimating population spread: What we can forecast and how well? *Ecology* **84,** 1979–1988.

Elton, C. A. (1958). *The Ecology of Invasions by Animals and Plants*. London: Methuen.

Fisher, R. A. (1937). The wave of advance of advantageous genes. *Annals of Eugenics* **7,** 355–369.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004). *Bayesian Data Analysis*, 2nd edition. New York: Chapman & Hall.

Hanski, I. (1999). *Metapopulation Ecology*. New York: Oxford University Press.

Hastings, A., Cuddington, K., Davies, K., et al. (2005). The spatial spread of invasions: New developments in theory and practice. *Ecology Letters* **8,** 91–101.

Hengeveld, R. (1993). What to do about the North American invasion by the Collard Dove? *Journal of Field Ornithology* **64,** 477–489.

Hooten, M. B. (2006). *Hierarchical SpatioTemporal Models for Ecological Processes*. Ph.D. Dissertation, University of Missouri. Columbia, Missouri.

Hooten, M. B. and Wikle, C. K. (2007). A hierarchical Bayesian nonlinear spatiotemporal model for the spread of invasive species with application to the Eurasian Collared-Dove. *Environmental and Ecological Statistics*, in press.

Kolar, C. S. and Lodge, D. M. (2001). Progress in invasion biology: Predicting invaders. *Trends in Ecology and Evolution* **16,** 199–204.

Kot, M. (2001). *Elements of Mathematical Biology*. Cambridge: Cambridge University Press.

Link, W. A., Barker, R. J., Sauer, J. R., and Droege, S. (1994). Within-site variability in surveys of wildlife populations. *Ecology* **74,** 1097–1108.

Nathan, R., Perry, G., Cronin, J. T., Strand, A. E., and Cain, M. L. (2003). Methods for estimating long-distance dispersal. *Oikos* **103,** 261–273.

Neubert, M. G. and Caswell, H. (2000). Demography and dispersal: Calculation and sensitivity analysis of invasion speed for structured populations. *Ecology* **81,** 1613–1628.

Olkin, I., Petkau, A. J., and Zidek, J. V. (1981). A comparison of N estimators for the binomial distribution. *Journal of the American Statistical Association* **76,** 637–642.

Raftery, A. E. (1988). Inference for the binomial N parameter: A hierarchical Bayes approach. *Biometrika* **75,** 223–228.

Renshaw, E. (1991). *Modelling Biological Populations in Space and Time.* Cambridge: Cambridge University Press.

Robbins, C. S., Bysrak, D. A., and Geissler, P. H. (1986). The Breeding Bird Survey: Its first fifteen years, 1965–1979. Fish and Wildlife Service Resource Publication 157, US-DOI, Washington D.C.

Romagosa, C. M., and Labisky, R. F. (2000). Establishment and dispersal of the Eurasian Collared-Dove in Florida. *Journal of Field Ornithology* **71,** 159–166.

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

Royle, J. A. and Dorazio, R. M. (2006). Hierarchical models of animal abundance and occurrence. *Journal of Agricultural, Biological, and Environmental Statistics* **11**(3), 249–263.

Sauer, J. R., Peterjohn, B. G., and Link, W. A. (1994). Observer differences in the North American Breeding Bird Survey. *Auk* **111,** 50–62.

Shigesada, N. and Kawasaki, K. (1997). *Biological Invasions: Theory and Practice.* New York: Oxford University Press.

Skellam, J. G. (1951). Random dispersal in theoretical populations. *Biometrika* **38,** 196–218.

Su, Z., Peterman, R. M., and Haeseker, S. L. (2004). Spatial hierarchical Bayesian models for stock-recruitment analysis of pink salmon (*Oncorhynchus gorbuscha*). *Canadian Journal of Fisheries and Aquatic Sciences* **61,** 2471–2486.

Thomas, L., Buckland, S. T., Newman, K. B., and Harwood, J. (2005). A unified framework for modelling wildlife population dynamics. *Australian and New Zealand Journal of Statistics* **47,** 19–34.

Turchin, P. (2003). *Complex Population Dynamics.* Princeton, New Jersey: Princeton University Press.

Wikle, C. K. (2002). A kernel-based spectral model for non Gaussian spatiotemporal processes. *Statistical Modelling: An International Journal* **2,** 299–314.

Wikle, C. K. (2003). Hierarchical Bayesian methods for predicting the spread of ecological processes. *Ecology* **84,** 1382–1394.

Wikle, C. K. and Hooten, M. B. (2006). Hierarchical Bayesian spatiotemporal models for population spread. In *Applications of Computational Statistics in the Environmental Sciences: Hierarchical Bayes and MCMC methods.* New York: Oxford University Press.

Wikle, C. K., Berliner, L. M., and Cressie, N. (1998). Hierarchical Bayesian space-time models. *Environmental and Ecological Statistics* **5,** 117–154.