Supplementary materials for this article are available at 10.1007/s13253-013-0147-9.

Computationally Efficient Statistical Differential Equation Modeling Using Homogenization

Mevin B. HOOTEN, Martha J. GARLICK, and James A. POWELL

Statistical models using partial differential equations (PDEs) to describe dynamically evolving natural systems are appearing in the scientific literature with some regularity in recent years. Often such studies seek to characterize the dynamics of temporal or spatio-temporal phenomena such as invasive species, consumer-resource interactions, community evolution, and resource selection. Specifically, in the spatial setting, data are often available at varying spatial and temporal scales. Additionally, the necessary numerical integration of a PDE may be computationally infeasible over the spatial support of interest. We present an approach to impose computationally advantageous changes of support in statistical implementations of PDE models and demonstrate its utility through simulation using a form of PDE known as "ecological diffusion." We also apply a statistical ecological diffusion model to a data set involving the spread of mountain pine beetle (*Dendroctonus ponderosae*) in Idaho, USA.

This article has supplementary material online.

Key Words: Change of support; Harmonic mean; Multi-scale analysis; Partial differential equation; Spatio-temporal model; Upscaling.

1. INTRODUCTION

With the exception of a few ground-breaking efforts (e.g., Fisher 1937; Hotelling 1927), the use of sophisticated ecological models involving derivatives in a rigorous statistical setting has only gained ground recently (e.g., Cangelosi and Hooten 2009; Hooten and Wikle 2008; Wikle 2003; Wikle and Hooten 2010; Zheng and Aukema 2010). As these recent studies have shown, scientifically motivated statistical models can be specified and

Mevin B. Hooten (✉) is Associate Professor, U.S. Geological Survey, Colorado Cooperative Fish and Wildlife Research Unit, Department of Fish, Wildlife, and Conservation Biology, Department of Statistics, Colorado State University, Fort Collins, CO, USA (E-mail: *Mevin.Hooten@colostate.edu*). Martha J. Garlick is Assistant Professor, Department of Mathematics and Computer Science, South Dakota School of Mines and Technology, Rapid City, SD, USA, James A. Powell is Professor, Department of Mathematics and Statistics, Utah State University, Logan, UT, USA.

Journal of Agricultural, Biological, and Environmental Statistics, ***Accepted for publication***
DOI: 10.1007/s13253-013-0147-9

Published online: 29 May 2013

fit readily in a hierarchical modeling context and often with a computational Bayesian implementation (i.e., Markov Chain Monte Carlo, MCMC).

Such modeling efforts have shown great promise for making science-based inference on dynamic natural processes, though they are not without their challenges (Wikle and Hooten 2010). Issues that commonly arise include: the chosen form of stochasticity, numerical stability of the deterministic equations, and discrepancy between the measurements and process in terms of spatial and temporal support.

With regard to the change of support issue in particular, the effects of spatial support changes on statistical inference are well described in the literature (e.g., Gotway and Young 2002; Wikle and Berliner 2005). In the case of upscaling spatial support specifically, common methods generally involve some form of spatial averaging or integration of the variable of interest over space (Ferreira and Lee 2007). Despite numerous suggestions in the literature to the contrary, many forms of aggregation in spatio-temporal statistical modeling projects still rely on over-simplified and non-scientific spatial and/or temporal scaling methods (e.g., arithmetic averaging) without regard for the inherent properties or dynamic features of the process under study. When more sophisticated approaches to handling the scaling are adopted, they are often based on statistical or computational properties and less on the mathematical underpinnings of the process being modeled (e.g., Ferreira et al. 2006). In fact, a popular method in the broader context of dimension reduction is to model the process on some lower-dimensional manifold through the use of a transformation involving a set of phenomenologically justified basis functions (e.g., Hooten and Wikle 2007; Royle and Wikle 2005; Wikle 2010) rather than mechanistically justified (as we present herein).

In what follows, we present a general approach to change of support that can be advantageous when utilizing partial differential equation (PDE) models in a statistical framework. The methodology we present reconciles the statistical and mathematical forms of multi-scale analysis and can be beneficial, both in terms of dimension reduction and consistency with the dynamics implied by the chosen mathematical model. A detailed presentation of the mathematical approach to multi-scale analysis can be found in the mathematical literature (e.g., Holmes 1995; Murray 2002; Okubo and Levin 2001), thus, here we focus on the use of multi-scale methods for approximating PDE solutions for statistical inference. We illustrate both the mathematical and statistical analyses through a specific example involving ecological diffusion and provide a simulation study and discussion of the advantages and potential applications of this approach. We show that the specific nature of ecological diffusion suggests an appropriate form of upscaling based on harmonic averaging that yields a more computationally efficient and stable statistical algorithm and sacrifices very little in the way of statistical inference. We then apply this upscaling approach in the analysis of a real data set involving the spread of an insect species over a large spatial domain.

1.1. PARTIAL DIFFERENTIAL EQUATION EXAMPLE: ECOLOGICAL DIFFUSION

To begin, we will focus our analysis on the specific example of ecological diffusion. To gain an understanding of how this form of mathematical model arises and how it may be useful, we derive it from first principles. Following Turchin (1998), we first consider

an individual-based or Lagrangian understanding of animal movement. To illustrate the approach, we will use a one-dimensional spatial domain throughout, which easily generalizes to higher dimensions.

Suppose that, in a certain time interval $(t-\Delta t, t]$, an animal at location x can move left, right, or remain where it is with probabilities: $\phi_L(x,t)$, $\phi_R(x,t)$, and $\phi_N(x,t)$, respectively (where, $\phi_L(x,t)+\phi_R(x,t)+\phi_N(x,t)=1$). Then the probability of the animal occupying location x at time t is:

$$\begin{aligned} p(x,t) = {} & \phi_L(x+\Delta x, t-\Delta t)\,p(x+\Delta x, t-\Delta t) \\ & + \phi_R(x-\Delta x, t-\Delta t)\,p(x-\Delta x, t-\Delta t) \\ & + \phi_N(x, t-\Delta t)\,p(x, t-\Delta t), \end{aligned} \tag{1.1}$$

where the Δ notation refers to changes in time and space (i.e., $+\Delta x$ represents the change in spatial location in the positive direction, for a 1-D spatial domain). If we ultimately seek an Eulerian model on the probability of occupancy, $p(x,t)$, then we need to replace the Δ notation with differential notation. Turchin (1998) proceeds by expanding each of the probabilities in a Taylor series, truncating to remove higher order terms, and then substituting the truncated expansions back into (1.1). This yields a recurrence equation involving partial derivatives:

$$\begin{aligned} p = {} & (\phi_L+\phi_N+\phi_R)p - \Delta t(\phi_L+\phi_N+\phi_R)\frac{\partial p}{\partial t} - \Delta t p\frac{\partial}{\partial t}(\phi_L+\phi_N+\phi_R) \\ & - \Delta x(\phi_R-\phi_L)\frac{\partial p}{\partial x} - \Delta x p\frac{\partial}{\partial x}(\phi_R-\phi_L) + \frac{\Delta x^2}{2}(\phi_L+\phi_R)\frac{\partial^2 p}{\partial x^2} \\ & + \Delta x^2\frac{\partial p}{\partial x}\frac{\partial}{\partial x}(\phi_L+\phi_R) + p\frac{\Delta x^2}{2}\frac{\partial^2}{\partial x^2}(\phi_L+\phi_R) + \cdots, \end{aligned}$$

where we have defined $p \equiv p(x,t)$, $\phi_L \equiv \phi_L(x,t)$, $\phi_N \equiv \phi_N(x,t)$, and $\phi_R \equiv \phi_R(x,t)$ to simplify the expressions. Combining like terms results in a PDE of the form:

$$\frac{\partial p}{\partial t} = -\frac{\partial}{\partial x}(\beta p) + \frac{\partial^2}{\partial x^2}(\delta p),$$

where $\beta = \Delta x(\phi_R-\phi_L)/\Delta t$ and $\delta = \Delta x^2(\phi_R+\phi_L)/2\Delta t$. This resulting model is now Eulerian and known as the Fokker–Planck or Kolmogorov equation (Risken 1989), and when thought of in an ecological setting, the spatial density $u(x,t)$, of some number of total animals (M), can be considered by letting $u(x,t) \equiv Mp(x,t)$. In this context, assuming for the moment that there is no advection component (i.e., $\beta = 0$), we have the ecological diffusion equation:

$$\frac{\partial u}{\partial t} = \frac{\partial^2}{\partial x^2}(\delta u), \tag{1.2}$$

where the process of interest is $u \equiv u(x,t)$, and $\delta \equiv \delta(x,t)$ represents the diffusion coefficients that could vary over space and time. In the specific example that follows, δ represents animal motility and is only assumed to vary in space (albeit at two scales). We note here that one could arrive at an alternative reduction of the Fokker–Planck equation by assuming

that $\delta = 0$, thus implying that animal movement is driven by advection only. Though, perhaps less intuitive, we might expect such behavior in wind- or water-advected populations (e.g., egg dispersal in a river system). The methods we present in what follows focus on the diffusion-only case, as these may be applicable to a greater range of real data scenarios, but they certainly apply to other forms of PDEs.

Returning to the ecological diffusion model, we should note other ways to arrive at Equation (1.2) (Turchin 1998); however, we feel that this perspective may be directly beneficial to those modeling spatio-temporal population dynamics, as the recent literature suggests (e.g., Wikle and Hooten 2010; Lindgren, Rue, and Lindstrom 2011). The properties of (1.2) are quite a bit different than those of plain or Fickian diffusions. The fundamental difference is that the diffusion coefficient is on the inside of the two spatial derivatives rather than between them (Fickian, $\frac{\partial u}{\partial t} = \frac{\partial}{\partial x}\delta\frac{\partial}{\partial x}(u)$) or on the outside (plain, $\frac{\partial u}{\partial t} = \delta\frac{\partial^2}{\partial x^2}(u)$). Ecological diffusion describes a much less smooth process $u(x,t)$ than Fickian or plain diffusion, and allows for motility-driven congregation to sharply differ between neighboring habitat types. In some areas, animals may move slow, perhaps to forage, whereas in other areas they move fast, as in exposed terrain. The resulting behavior shows a congregative effect in areas of low motility (i.e., $\delta \downarrow$) and a dispersive effect in areas of high motility (i.e., $\delta \uparrow$). In fact, depending on the boundary conditions, the steady-state solution has u proportional to the inverse of δ.

1.2. STATISTICAL IMPLEMENTATION

Recent efforts in spatio-temporal statistical modeling (e.g., Cressie and Wikle 2010; Hooten and Wikle 2008; Wikle 2003; Wikle et al. 2001; Wikle and Hooten 2010; Zheng and Aukema 2010) suggest that deterministic equations, such as (1.2), can be placed into a statistical context by assuming three things: first, that the dynamic process can be measured, and those observations are subject to sampling error; second, that the Eulerian model in (1.2) does not define the exact dynamics of the system, but, rather, serves as a well-founded scientific motivation for parameterizing the dynamics; third, we seek to estimate the model parameters (e.g., δ) given the observed data. Thus, we have three explicit sources of uncertainty in the model: observation, process, and parameter uncertainty. As discussed throughout the recent statistical literature on the subject, this situation is ideal for the use of a hierarchical model (e.g., Berliner 1996; Cressie et al. 2009). In this setting, we can think of the measurements as observed counts of animals, $N(x,t)$, at a set of locations and times. Then, by modeling the latent process $u(x,t)$ we can learn about the dynamics. Thus, using vector notation to denote finite sets of state variables and parameters (e.g., $\mathbf{u}(t) \equiv (u(1,t), \ldots, u(x,t), \ldots, u(n,t))'$ for n spatial locations of interest), let the statistical model be specified as

$$N(x,t) \sim \left[N(x,t) \,\middle|\, u(x,t)\right], \quad \forall x,t,$$
$$u(x,t) \sim \left[u(x,t) \,\middle|\, f\left(\mathbf{u}(t-\Delta t), \boldsymbol{\delta}\right)\right], \quad \forall x,t,$$

where the '$[\cdot]$' notation refers to a probability distribution, $\boldsymbol{\delta}$ represents a set of diffusion coefficients corresponding to each of the spatial locations, and the ecological diffusion

model (1.2) is represented by a discretized approximation $f(\mathbf{u}(t-\Delta t),\boldsymbol{\delta})$. Typically the discretized approximation amounts to a difference equation, though more sophisticated solvers could be employed. Note that, in some situations, the specific form of f could facilitate estimation (e.g., if linear with Gaussian error, then conjugacy is a byproduct). Then, assuming that a finite set of unique coefficients can be modeled by $\boldsymbol{\delta} \sim [\boldsymbol{\delta}]$, in a Bayesian framework we seek to find the conditional distribution (i.e., posterior) of the parameters $\boldsymbol{\delta}$ and latent state variables $\mathbf{u}(t)$ given the data $\mathbf{N}(t)$:

$$\begin{aligned}\left[\left\{\mathbf{u}(t)\right\},\boldsymbol{\delta} \;\middle|\; \left\{\mathbf{N}(t)\right\}\right] &\propto [\text{Data} \mid \text{Process}][\text{Process} \mid \text{Parameters}][\text{Parameters}] \\ &\propto \prod_t \left[\mathbf{N}(t) \;\middle|\; \mathbf{u}(t)\right] \prod_t \left[\mathbf{u}(t) \;\middle|\; f\left(\mathbf{u}(t-\Delta t),\boldsymbol{\delta}\right)\right][\boldsymbol{\delta}], \end{aligned} \tag{1.3}$$

where, for the time being, a set of initial and boundary conditions are assumed. In the generalized case, additional conditions could be (and have been) further modeled (Wikle, Berliner, and Milliff 2003).

This framework has proven to be a powerful means for directly incorporating scientific information into the modeling process and provides an intuitive way to account for observational and process uncertainty while estimating the parameters that control the dynamics. In principle, this statistical framework could be employed for any differential equation model if sufficient data were available to provide the learning. In practice, several complications can arise when fitting such models. First, if the model is fit using MCMC, the algorithm will require iterative evaluation of the discretized forward model f; this can be very computationally demanding if the discretization is spatially and/or temporally fine. Additionally, there is often a discrepancy between the support of the measurements and that of the process, but inference is desired at the finer of the two scales. In the specific situation with a spatially heterogeneous set of diffusion coefficients (i.e., $\boldsymbol{\delta}(x)$), the dimension of the parameter space is potentially huge. Finally, there is the issue of stability of the forward model (f) itself and the fact that ecological diffusion is inherently less numerically stable than simpler forms (e.g., Fickian and plain diffusion; Mitchell and Griffiths 1980).

In the following section, we take an analytical approach to change of support in PDE modeling by introducing additional temporal and spatial scales in what is referred to as "homogenization" in the PDE literature (Holmes 1995; Pavliotis and Stuart 2008). We will see that such analyses can lead to a choice of statistical quantities that are consistent with the mathematics while inducing an advantageous change of support in the inverse implementation of the model. The result is a faster, more robust algorithm for fitting PDE models to data in a rigorous statistical framework.

2. HOMOGENIZATION FOR ECOLOGICAL DIFFUSION

2.1. MODEL-SPECIFIC UPSCALING METHOD

Homogenization is a mathematical technique that uses the "method of multiple scales" and can be utilized to reframe the ecological diffusion equation in terms of a larger spatial

scale, while preserving the small scale variability in the diffusion (i.e., motility) coefficient. As described in Section 1.1, the spread of a population or disease is connected to the movement of individuals in local habitats as well as over large landscapes. Therefore, we assume the diffusion coefficient depends on two spatial scales, one which varies quickly over short distances (e.g., due to vegetation and soil characteristics) and one which varies over larger landscape scales (e.g., due to climate and geography). We introduce a parameter, ϵ, which is the ratio of the small and large scales. That is, let x be the large spatial scale with an associated slow time scale, t. Then the spatial grain (or resolution) can be expressed as $\Delta x = \frac{\Delta y}{\epsilon}$, while the associated time grains are related by $\Delta t = \frac{\Delta \tau}{\epsilon^2}$. For example, if 1 unit of x is the same as 10 units of y, then $\epsilon = \frac{1}{10}$, and changes on the order of $\mathcal{O}(\epsilon)$ in x become order $\mathcal{O}(1)$ changes in y.

We assume that the dependent variable u can be written as a power series in ϵ, $u = u_0 + \epsilon u_1 + \epsilon^2 u_2 + \cdots$. Since ϵ is small (i.e., closer to zero than one), the first term, u_0 contributes most to the approximation of u, with the other terms correcting the approximation. The solution u_0 is referred to as the leading order approximation of u. The idea behind the homogenization of ecological diffusion is not to find an analytical solution, but rather to rewrite the diffusion equation in terms of u_0 as a function of the large landscape scale and slow time scale with an "averaged" motility coefficient.

Transforming the derivatives in terms of the new variables, x, y, t, and τ, and replacing u with the above power series, the ecological diffusion equation becomes

$$\left(\frac{\partial}{\partial \tau} + \epsilon^2 \frac{\partial}{\partial t}\right)\left(u_0 + \epsilon u_1 + \epsilon^2 u_2 + \cdots\right) = \left(\frac{\partial^2}{\partial y^2} + 2\epsilon \frac{\partial^2}{\partial x \partial y} + \epsilon^2 \frac{\partial^2}{\partial x^2}\right)\delta(x, y)\left(u_0 + \epsilon u_1 + \epsilon^2 u_2 + \cdots\right). \tag{2.1}$$

Even though it appears that this greatly complicates things, it actually simplifies them, because this equation can be divided into a series of simpler equations by equating like powers of ϵ. We begin by gathering the terms of order ϵ^0 to form the equation $\frac{\partial u_0}{\partial \tau} = \frac{\partial^2}{\partial y^2}(\delta(x, y)u_0)$. Note that this is a diffusion equation with derivatives in terms of the small spatial scale y and the fast time scale, τ. Therefore the solution of this equation decays quickly to the solution of the steady state equation $\frac{\partial^2}{\partial y^2}(\delta(x, y)u_0) = 0$, which is $u_0 = c(x, t)/\delta(x, y)$. We proceed now with the equation formed by terms with ϵ^1; it is solved in the same manner and the result is similar, yielding $u_1 = b(x, t)/\delta(x, y)$. This is precisely the same equation as for u_0, and since any relevant c is already satisfied by u_0, the b for u_1 is homogeneous, and consequently $b = 0$. The terms b and c are effectively integration "constants" (i.e., constant in y) and we return to them in what follows.

Finally, gathering the terms with ϵ^2 forms the equation

$$\frac{\partial u_2}{\partial \tau} + \frac{\partial u_0}{\partial t} = \frac{\partial^2}{\partial y^2}\left(\delta(x, y)u_2\right) + \frac{\partial^2}{\partial x^2}\left(\delta(x, y)u_0\right). \tag{2.2}$$

As in the equations for ϵ^0 and ϵ^1, there is a derivative in the fast time scale τ, so the solution to this equation decays rapidly to the solution of the steady state equation with respect to τ (i.e., the equation without the term $\frac{\partial u_2}{\partial \tau}$). After substituting our expressions for

u_0 and u_1 into (2.1), and retaining the portion of the equation that influences the steady state with respect to the fast time scale τ, the equation becomes

$$\frac{\partial^2}{\partial y^2}\big(\delta(x,y)u_2\big) = \frac{\partial}{\partial t}\left(\frac{c(x,t)}{\delta(x,y)}\right) - \frac{\partial^2}{\partial x^2}c(x,t). \tag{2.3}$$

This is due to the simplification of the mixed derivative term in (2.1). The partial derivative with respect to y, of a function without y dependence, is zero. Since it does not contribute to the end result, we can set $u_1 = 0$.

Focusing our attention on (2.3), we do not attempt to solve for u_2, but rather we extract the result by integrating this equation once and analyzing the behavior of the terms. Integrating (2.3) once yields:

$$\frac{\partial}{\partial y}\big(\delta(x,y)u_2\big) = a(x,t) + \frac{\partial}{\partial t}c(x,t)\int_{\mathcal{Y}}\frac{1}{\delta(x,s)}\,ds - y\frac{\partial^2}{\partial x^2}c(x,t), \tag{2.4}$$

where, $a(x,t)$, like b and c, is constant in y, and $\mathcal{Y}$ represents the subset of spatial support over which averaging is desired. The last two terms on the right-hand side of (2.4) grow unbounded as y tends toward infinity. That behavior would not occur in a valid solution, so a necessary condition is that the sum of those two terms equals zero. That is,

$$\lim_{y\to\infty}\left[\frac{\partial}{\partial t}c(x,t)\int_{\mathcal{Y}}\frac{1}{\delta(x,s)}\,ds - y\frac{\partial^2}{\partial x^2}c(x,t)\right] = 0. \tag{2.5}$$

This imposed condition, with a few algebraic manipulations, yields our homogenized diffusion equation:

$$\frac{\partial}{\partial t}c(x,t) = \bar{\delta}(x)\frac{\partial^2}{\partial x^2}c(x,t), \tag{2.6}$$

where

$$\bar{\delta}(x) \equiv \left(\lim_{y\to\infty}\frac{1}{y}\int_{\mathcal{Y}}\frac{1}{\delta(x,s)}\,ds\right)^{-1}. \tag{2.7}$$

The diffusion coefficient $\bar{\delta}$ is now outside of the derivatives (2.6) and is the harmonic mean of δ over the small scale. The variable $c(x,t)$ in (2.6) is related to the population density u_0 by $u_0 = \frac{c(x,t)}{\delta(x,y)}$. Therefore, after (2.6) is solved numerically for $c(x,t)$ over the large scale, the small scale variability is returned to the solution via division by δ.

2.2. ADVANTAGES OF UPSCALING USING HOMOGENIZATION

The harmonic mean (i.e., the reciprocal of the arithmetic mean of the reciprocals) is not only a natural result of the homogenization procedure, but it is also very appropriate for characterizing rates of change (e.g., diffusion coefficients), as it is less than both the arithmetic and geometric means (by Jensen's inequality) and thus less sensitive to extreme values. A brief example can illuminate the importance of this simple, yet elegant statistic, for models involving speed and distance. That is, consider the situation where an animal moves at a certain speed s_1 for a specific distance (Δx), then switches to a different speed, say s_2, for another trip of the same distance. In this situation, the total time should be $\frac{2\Delta x}{\bar{s}}$,

where $\bar{s}$ is average speed over both trips. If $s_1 = 6$, $s_2 = 4$, and $\Delta x = 1$, then the total travel time is $\frac{1}{6} + \frac{1}{4} = \frac{10}{24}$, but if the animal traveled at the arithmetic mean speed for the entire distance ($2\Delta x$), the total travel time would be $\frac{10}{25}$, whereas the total travel time would be $\frac{10}{24}$ if it had traveled at the harmonic mean speed for the entire distance. Clearly, in the above example, the total travel time would have been too small if the arithmetic mean were used to average the speeds, whereas the harmonic mean provides the correct travel time. This result holds in general and demonstrates how the smaller harmonic form of mean can be better for averaging rates.

From a deterministic perspective, the homogenization result derived in the previous section allows one to solve the PDE with a spatially varying diffusion coefficient over large domains using an efficient computational algorithm. The averaging alone can dramatically reduce calculations, however the homogenized equation (2.6) also describes the dynamics of a smooth potential field and this increases computational stability. One can then use a much coarser discretization in the numerical solver for the forward problem (i.e., by a factor of $\frac{1}{\epsilon^2}$ in time and $\frac{1}{\epsilon}$ in space). However, from a statistical perspective, we need to solve the inverse problem. That is, having observed $u(x,t)$ (or some discrete representation of it), we seek to learn about the parameters controlling the dynamics of the system (i.e., $\delta(x)$). If a Monte Carlo based estimation method (e.g., MCMC or importance sampling) were used, the above result would be absolutely necessary if inference were desired over large spatial domains due to the required iterative evaluation of a PDE solver. By mathematically incorporating an explicit change of support we have effectively simplified the entire problem and allowed for efficient computation while limiting the information loss associated with the averaging. Finally, unlike in Fickian homogenization, this result (i.e., the harmonic averaging) holds for ecological diffusion in higher dimensions (Garlick, Powell, and Hooten 2011). In fact, we present the mathematical details describing the homogenization procedure for a generalized version of the 2-D ecological diffusion model in Supplementary Appendix A.

2.3. IMPLEMENTATION OF PHYSICAL-STATISTICAL MODEL WITH HOMOGENIZATION

In terms of implementation, one can proceed as described in Section 1.2 where, depending on the type of observations that are made, a suitable statistical data model can be chosen along with an error distribution for the process model and prior for the parameters. Assuming that the data consist of observations of animal abundance (i.e., $N(x,t)$, counts of animals; discussed in more detail in the following section) and we make the appropriate homogenization transformations, we can treat the harmonic mean as an operator using vector notation (i.e., $\bar{\boldsymbol{\delta}} \equiv \bar{\boldsymbol{\delta}}(\boldsymbol{\delta})$) in the following general hierarchical statistical model:

$$N(x,t) \sim \left[N(x,t) \,\middle|\, u_0(x,t)\right], \quad \forall x,t \tag{2.8}$$

$$u_0(x,t) \sim \left[u_0(x,t) \,\middle|\, f_h\left(\mathbf{u}_0(t-\Delta t), \bar{\boldsymbol{\delta}}(\boldsymbol{\delta})\right)\right], \quad \forall x,t \tag{2.9}$$

$$\boldsymbol{\delta} \sim [\boldsymbol{\delta}], \tag{2.10}$$

where the function f_h represents the plain diffusion solver as a difference equation. Note that the process stage (2.9) of the hierarchical model could either be stochastic or a degenerate distribution implying no additional process uncertainty beyond that provided through the data model (2.8).

Though perhaps possible, it would not be trivial to fit such a model and obtain uncertainty estimates for model parameters using maximum likelihood. Therefore, we describe a Bayesian approach to fit the model that uses an MCMC algorithm as described in recent literature pertaining to physical-statistical modeling (e.g., Wikle and Hooten 2010). The posterior distribution corresponding to the model specified in (2.8)–(2.10) can be expressed as

$$\big[\{\mathbf{u}_0(t)\}, \boldsymbol{\delta} \mid \{\mathbf{N}(t)\}\big] \propto \prod_t \big[\mathbf{N}(t) \mid \mathbf{u}_0(t)\big] \prod_t \big[\mathbf{u}_0(t) \mid f\big(\mathbf{u}_0(t-\Delta t), \boldsymbol{\delta}\big)\big][\boldsymbol{\delta}]. \tag{2.11}$$

Notice that the posterior in (2.11) is similar to that in (1.3) except that it involves a product over the homogenized process distribution rather than the non-homogenized process. This model is completely non-conjugate, implying that full-conditional distributions cannot be found analytically, thus, a Metropolis–Hastings approach must be used to sample from the full-conditionals sequentially.

The portion of an MCMC algorithm where the homogenization technique helps the most is when sampling the diffusion coefficients $\boldsymbol{\delta}$. The Metropolis-Hastings ratio:

$$\frac{(\prod_t [\mathbf{u}_0(t) \mid f_h(\mathbf{u}_0(t-\Delta t), \boldsymbol{\delta}^{(*)})])[\boldsymbol{\delta}^{(*)}][\boldsymbol{\delta}^{(k-1)} \mid \boldsymbol{\delta}^{(*)}]}{(\prod_t [\mathbf{u}_0(t) \mid f_h(\mathbf{u}_0(t-\Delta t), \boldsymbol{\delta}^{(k-1)})])[\boldsymbol{\delta}^{(k-1)}][\boldsymbol{\delta}^{(*)} \mid \boldsymbol{\delta}^{(k-1)}]}, \tag{2.12}$$

needs to be computed at each MCMC step, where $\boldsymbol{\delta}^{(k-1)}$ is the last MCMC sample for $\boldsymbol{\delta}$ and it is assumed that $\mathbf{u}_0(t)$ represents the current MCMC sample for the homogenized PDE process. From (2.12) we can see the advantage of having the homogenized PDE solver f_h, as on every MCMC iteration we must forward solve the entire PDE using the proposal value for $\boldsymbol{\delta}^{(*)}$. Thus, the faster the solver, the faster the MCMC algorithm. The plain diffusion solver f_h, operating on a coarser spatial support, is much faster and more stable than the original PDE solver f. Homogenization suggests an optimal statistical change of support in the latent dynamic process, and, though we are only burdened with calculating $u_0(x,t)$, the statistical model provides inference about $\boldsymbol{\delta}$ based on $u(x,t)$. Note that when we use the term "optimal" here and after we are referring to the fact that the form of averaging is appropriate and directly relevant for the model we're considering.

3. SIMULATION: 1-D ECOLOGICAL DIFFUSION WITH A FINITE SET OF DISTINCT ENVIRONMENTS

In order to illustrate the methods presented above, we consider a simulation of animal movement in a 1-D spatial setting with three different land types. The idea would be that animal motility is homogeneous within each distinct land type, but the landscape itself could be made up of irregular patterns involving those land types. This is a relatively realistic scenario where covariate information may be available for a certain spatial domain

(e.g., remotely sensed data) and animals are counted at a set of locations over that landscape.

In a computational setting where we must discretize the spatial domain using n finite areal spatial units, suppose we code a set of p distinct land types using dummy variables in an $n \times p$ design matrix $\mathbf{X}$, then we could write the spatial field of diffusion coefficients for our domain of interest as $\boldsymbol{\delta} = \mathbf{X}\mathbf{d}$, where $\mathbf{d}$ is a vector of dimension $p \times 1$, for $p << n$, containing the unique coefficients. As a simulation landscape, Figure 1a illustrates the irregularity of the land types and their associated diffusions. In this case, though we evaluated numerous scenarios, the specific one presented here has $\mathbf{d} = (4, 5, 6)'$. For comparison, we have also shown the harmonic mean $\bar{\delta}$, averaged over one thirtieth of the spatial domain ($\mathcal{Y}$) or two units in the discretized case.

To simulate data, we first need to simulate the underlying dynamic u process. Using no-flux boundary conditions and an initial state where a population is placed in the center of the spatial domain and allowed to spread out over time through the PDE in (1.2), we can visualize the solution as a spatio-temporal process (Figure 1b) where the x-axis represents space, the y-axis represents population density, and the intensity of the shaded lines correspond to the time evolution of the system. That is, in Figure 1b, lighter shaded lines illustrate the process $u(x, t)$ at earlier stages of the temporal evolution and darker shaded lines correspond to the latter time periods. In Figure 1b, we can see the density spreading out in such a manner that there tends to be congregation in areas of low motility and dispersal in areas of higher motility. Also note that we are focusing on the transient dynamics here and thus the process has not reached the steady-state in the period of time being considered. This is for two reasons, first because most real data are only available for the transient period of a natural dynamical system, and second, a different estimation procedure could be used if the steady state were observed directly due to the relationship between δ and the steady state of u.

In order to visualize the differences between $u(x, t)$ and $c(x, t)$, notice the smoothness of $c(x, t)$ in Figure 1c as compared to Figure 1b. Also, notice that we can recover the approximation $u_0(x, t)$, by dividing $c(x, t)/\delta(x, t)$ (Figure 1d).

Now, assuming a Poisson distribution for a data model, we can "observe" the simulated process by sampling $N(x, t) \sim \text{Pois}(u(x, t))$ for some finite set of locations x and times t (i.e., 60 locations and 10 times). These observations will serve as simulated data and can be visualized in the same manner as the underlying dynamical processes (Figure 1e). Note that there is substantially more noise in the observations $N(x, t)$ than in the underlying process $u(x, t)$.

For clarity in this example, and because we wish to compare the homogenized statistical model with the non-homogenized model in terms of parameter estimation, we will consider a Bayesian statistical model where the process component is a degenerate probability distribution and thus collapsed into the likelihood. In both the homogenized and non-homogenized cases, we use a vague Gaussian prior on the diffusion coefficients $\mathbf{d} \sim N(\mathbf{1}, 1000^2\mathbf{I})$. Alternatively, a truncated Gaussian or other prior with positive support could be used for $\mathbf{d}$, however, in this case for our simulation, the chosen $\mathbf{d}$ values were sufficiently far from the physically limited zero lower bound and thus the posterior distributions will not be affected by the unconstrained prior. For real data analysis, a model

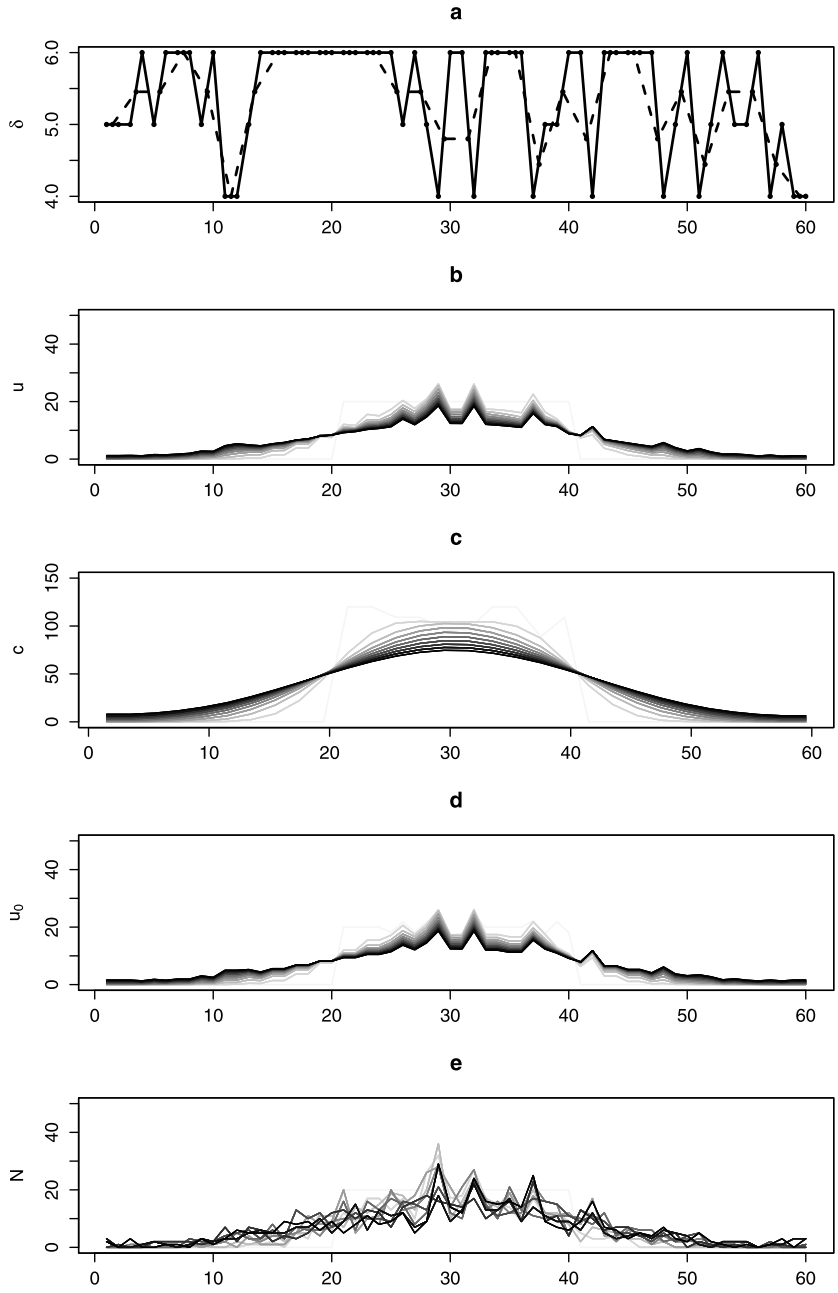

Figure 1. (a) Spatial landscape (x-axis) with ecological diffusion coefficients (y-axis) shown in solid ($\boldsymbol{\delta}$) and dashed ($\bar{\boldsymbol{\delta}}$) where the averaging is one thirtieth of the spatial domain or two units of space in this case. (b) Forward simulation of ecological diffusion with $u(x,t)$ shown on the y-axis, through space (x-axis) and over time (gray shading). Lighter shading represents the state of the system at earlier time points, whereas darker shading represents later times. Note that the initial state $u(x,0)$ was specified with population mass in the middle of the spatial domain, hence the population is spreading out based on the landscape and $\boldsymbol{\delta}$. (c) Forward simulation of $c(x,t)$ via plain diffusion, shown on the same spatial domain. (d) PDE approximation, $u_0(x,t)$, over time as recovered from $c(x,t)$ and $\boldsymbol{\delta}$. (e) Simulated data, $N(x,t) \sim \text{Pois}(u(x,t))$, where the lines represent a set of discrete counts of "organisms" (i.e., observations) simulated using the hierarchical model.

with positive support on $\mathbf{d}$ might be warranted. In this non-hierarchical setting, the ratio of observations to parameters is quite high (i.e., 200 to 1) and thus there will be plenty of information to estimate $\mathbf{d}$ despite their vague proper prior. In the hierarchical setting, a more precise prior may be useful for improving identifiability.

In this example, we wish to compare the regular (or non-homogenized) and homogenized PDEs in terms of the statistical inference they provide and their computational efficiency, thus we consider two separate statistical models. In this case, we assume that an aggregated spatio-temporal point process is observed with measurements corresponding to animal abundance in space over time (the vector $\mathbf{N}_t$). The inhomogeneous point process in our example represents animal locations on a landscape and when these points are aggregated over a contiguous set of grid cells (i.e., discrete spatial support), the likelihoods for the regular model and homogenized model can then be expressed as Poisson (e.g., Warton and Shepherd 2010; Aarts, Fieberg, and Matthiopoulos 2012):

$$\mathbf{N}(t) \sim \text{Pois}(f\big(\mathbf{u}(t - \Delta_t), \boldsymbol{\delta}(\mathbf{d})\big), \tag{3.1}$$

and,

$$\mathbf{N}(t) \sim \text{Pois}\big(f_h\big(\mathbf{u}_0(t - \Delta_t), \bar{\boldsymbol{\delta}}(\mathbf{d})\big)\big), \tag{3.2}$$

$\forall t$, where f and f_h correspond to the original and homogenized PDE solvers, respectively.

Conditioning on the initial state (i.e., $\mathbf{u}(0)$) and no-flux boundary conditions, we then seek the posterior distribution for the diffusion coefficients from each model, both of which are conditional distributions of the form: $[\mathbf{d} \mid \{\mathbf{N}(t), \forall t\}]$. Due to the nonlinearities in the PDE solver as well as the specific forms of the likelihood and prior, the posterior distribution will not be conjugate, and thus, not analytically tractable. Following the basic approach outlined in the previous section, we constructed an MCMC algorithm to sample from the posterior using Metropolis–Hastings updates for the diffusion coefficients $\mathbf{d}$. This algorithm was developed using the R Statistical Computing Environment (R Development Core Team 2012).

For this simulation study, we sampled $L = 500$ realizations of $\mathbf{N}_t^{(l)}$, for $l = 1, \ldots, L$ from the "true" model and then fit both the original and homogenized statistical PDE models to each of the data sets by obtaining 5000 MCMC samples (assessing convergence visually and discarding the first 1000 MCMC samples as the burn-in). The resulting chains for each model converged rapidly, though we noticed improved mixing in the case of the homogenized PDE model. We attributed this to the improved stability of the plain diffusion solver as compared with the ecological diffusion solver. Table 1 summarizes our simulation-based findings pertaining to the bias and credible interval coverage of the diffusion coefficients using both models. In this case, both models captured the truth well and demonstrated very little bias, while the 95 % credible intervals for each of the diffusion coefficients had very close to the correct coverage.

In terms of computational performance, the homogenized PDE model took only 40 % of the time it took the original PDE model to be fit (the entire simulation study took approximately 14 hours for both models combined, on a 6-Core 2.93 GHz Processor Workstation with 32 GB of RAM). It should be noted that the gain in computing efficiency is increased

Table 1. Marginal posterior bias and 95 % credible interval coverage for each of the ecological diffusion coefficients **d** resulting from both the model fit using the original PDE (u) and the homogenized PDE (u_0).

Parameter	Truth	Original bias (95 % CI coverage)	Homogenized bias (95 % CI coverage)
d_1	4	0.062 (0.931)	0.096 (0.929)
d_2	5	0.024 (0.939)	0.018 (0.958)
d_3	6	0.043 (0.946)	−0.028 (0.946)

as the extent of averaging increases. That is, in this case, we only averaged over one thirtieth of the spatial domain. Through simulation we also evaluated the change in computation time as a function of both the change of support for averaging and the magnitude of the diffusion coefficients (i.e., $\max(\boldsymbol{\delta})$). The latter is important for the stability of the numerical PDE solver being employed; in this case, the solver is a centered difference equation. These simulation results indicated that the computational savings generally improve with larger diffusion coefficients and larger changes in support. Specifically, in this 1-D example, the improvement in computational efficiency is asymptotic in both the change of support and increase in diffusion coefficients (i.e., increasing to an upper bound) and thus the best we can do with homogenization is an algorithm that is 5 times faster. In the 2-D spatial setting, we can improve on that by a power of two (Garlick, Powell, and Hooten 2011).

4. APPLICATION: SPREAD OF MOUNTAIN PINE BEETLE

4.1. BACKGROUND

The mountain pine beetle (MPB, *Dendroctonus ponderosae*) infests and kills *Pinus* host trees throughout western North America; rapid population growth and inherent positive feedbacks are hallmarks of this economically important species (Raffa et al. 2008). Unlike many phytophagous insects, successful MPB reproduction almost always results in death of all or part of the host. Host trees, however, have evolved effective resin response mechanisms to defend themselves against bark beetle attacks (Raffa, Phillips, and Salom 1993). At low population densities trees with less defensive capacity, and hence less food resources, are selected. Once populations reach outbreak densities, however, MPB can successfully overcome all suitable hosts (Boone et al. 2011).

MPB has been successful across a broad latitudinal and elevational thermal regime as evidenced by the numerous population outbreaks recorded during the past 100 to 150 years across western North America (Crookston, Stark, and Adams 1977; McGregor 1978; Perkins and Swetnam 1996; Alfaro et al. 2004). The severity and distribution of some recent outbreaks, however, differ from what can be inferred from historical records. Increasing temperature associated with climate change is believed to be a significant factor in recent outbreaks, with positive influences on development timing and cold survival (Logan and Powell 2001; Raffa et al. 2008; Sambaraju et al. 2012). Models describing the effect of temperature on MPB developmental timing and survival have been developed (Bentz, Logan, and Amman 1991; Gilbert et al. 2004; Régnierè and Bentz 2007), and have been used

to analyze MPB population response to historic and future climate regimes (Bentz et al. 2010; Safranyik et al. 2010). Powell and Bentz (2009) combined temperature-driven phenology, a controller of adult emergence timing, with a daily threshold of beetles required to overwhelm defenses of new host trees under attack (i.e., Allee effect) to produce a model for predicting local population growth rate within a watershed. Using observed phloem temperatures to drive their mechanistic model, year-to-year details of density-dependent growth and the transition from incipient to epidemic levels of MPB were predicted. This effort provided an accurate description of MPB population growth through time, yet the spatial behavior of MPB eruption and spread remains undescribed.

One poorly-understood aspect of MPB spatial ecology is dispersal. Proliferation of small spots of infested trees is largely dependent on short-range movement under the stand canopy, conditioned by host tree availability and size, MPB population levels, weather, and behavior-modifying chemicals (Mitchell and Preisler 1991; Safranyik et al. 1992). Through a chemically-mediated synergistic reaction with host defensive compounds, MPB release aggregation pheromones that attract additional beetles (Pitman 1971; Hughes 1973) resulting in a mass attack on a single focus tree. Individual trees are finite resources that can be overexploited, however, and it is therefore advantageous to redirect attacks. A complex suite of derived compounds and behaviors have evolved resulting in a close-range redirection of responding beetles to nearby trees (Borden et al. 1987; McCambridge 1967; Bentz, Powell, and Logan 1996). A spatially explicit model describing host tree switching at <10 m scales was developed by Powell, Logan, and Bentz (1996), parameterized using the spatio-temporal pattern of attacked trees in a stand (Biesinger et al. 2000), and used in an analysis to test hypotheses of MPB outbreak ecology (Logan et al. 1998). These modeling results support the notion that overall dispersal of MPB and subsequent successful attacks depend directly on host density.

Ignoring the effects of behavior-modifying chemicals, Heavilin and Powell (2008) fit deterministic integrodifference equation models to USDA Forest Service Aerial Detection Survey (ADS) data that describes spatial patterns of MPB killed trees. The models incorporated density dependence and possible Allee effects at 30-m resolution using type II and III functional responses, and described MPB dispersal via Gaussian and exponential dispersal kernels (Heavilin, Powell, and Logan 2007). The models predicted that mean dispersal distances were between 15 and 50 meters, a finding similar to other studies (Safranyik et al. 1992; Robertson, Nelson, and Boots 2007). No parameters tested reconciled predictions with observed spatial pattern at any scale, suggesting that dispersal cannot result from a random walk with fixed step sizes. Although more than three quarters of new attacks occur within 100 m (Robertson, Nelson, and Boots 2007), a significant aspect of MPB dispersal is long distance movement (>2 km, Robertson et al. 2009), contributing to the development of new spots. Localized infestations can erupt in areas geographically disjunct from other spots (Aukema et al. 2006). Understanding how MPB disperse is a critical element in understanding outbreak progression, and this information would allow for more robust forecasting across the entire range of mountain pine beetle.

4.2. SAWTOOTH STUDY AREA

The Sawtooth National Recreation Area (SNRA) in central Idaho was chosen as our study area for several reasons. A single host, lodgepole pine, predominates and grows in stands with relatively homogeneous demographics at the lowest elevations; host demographics are consistent across the valley due to historical disturbance patterns. Lodgepole pine stands in the SNRA are within a coherent geographic unit, bounded by 3000+ meter mountains on three sides, minimizing factors such as dispersal and immigration that have been shown to be important in MPB outbreaks (Aukema et al. 2008). The valley opens to the north, cross-wise to prevailing weather patterns, and is small enough that the entire valley experiences the same general climate. Consequently, per-capita growth rates are approximately constant (in space), although they vary significantly from year to year (Powell and Bentz 2009)

The study area is a rectangular region from approximately 44°22′N to 43°44′N (∼60 km) and 115°10′W to 114°28′W (∼30 km), comprising over 1800 km^2. The landscape is characterized by the Sawtooth valley and the surrounding mountains, nearly all of which are administered by the SNRA, Sawtooth National Forest. Elevation ranges from 1651 m to 3605 m; vegetation types range from shrub and grasslands to coniferous forests dominated by Douglas fir (*Pseudotsuga menziesii*), subalpine fir (*Abies lasiocarpa*), and lodgepole pine and whitebark pine (*Pinus albicaulis*). Extensive barren areas exist above tree-line at the highest elevations. The climate is characterized by very cold winters and mild summers. Extensive studies on MPB phenology and life history have been conducted within the study area boundary (Bentz and Mullins 1999; Bentz 2006; Powell and Bentz 2009).

4.3. PINE STEM DENSITY AND AERIAL DETECTION SURVEY DATA

Spatially explicit data sets of pine density at 30-m resolution were derived for both study areas using existing geospatial data sets of vegetation composition and structure (Crabb, Powell, and Bentz 2012). For the Sawtooth study area forest density (trees per acre >2.54 cm diameter at breast height (DBH)) at 250 m resolution developed by the USDA Forest Service FIA (Blackard et al. 2008) were downscaled to 30-m resolution using data from the inter-agency Landscape Fire and Resource Management Planning Tools Project (LANDFIRE).

The USDA Forest Service Forest Health Protection (FHP) branch conducted annual ADS from fixed-wing aircraft between 1989 and 2010. During ADS flights, trained observers collect and manually record data on a geo-referenced map based on visual inspection of forest structure, tree species, and foliage color (Halsey 1998). Observers delineate areas of trees with faded foliage, estimating attack and tree mortality caused by MPB the previous year. ADS data sets include ‘damage’ polygon shapefiles with metadata describing the estimated number of trees per acre affected and a code for the damage causal agent(s) (DCA). Only areas with at least 20 trees per hectare were mapped. These georeferenced data serve as our source of information on the spatial location, timing, and intensity of MPB impact in the Sawtooth study area. Rasters of total MPB impact by year were created by summing MPB impacts across observations for each polygon, then converting the

polygons to rasters. The rasters were produced at a 30 meter resolution and were kept in the same coordinate system as the original ADS shapefiles.

To develop one-dimensional slices we chose a transect 6 kilometers wide (east-west) and 60 kilometers long (north-south), running along the western side of the Stanley Valley and intersecting the majority of MPB impact from 2000 to 2005. The strip is bounded on the east by the valley bottom, comprised of hay fields and range land with few or no hosts; on the west the strip is bounded by the ascending slopes of the Sawtooth range itself, thus giving natural boundaries. Average host abundances (ranging from 131 to 731 pine stems per hectare) were calculated as the mean of each row in the host raster and referred to as $w(x)$ (Figure 2a) in what follows; total impacts were calculated by summing the number of impacted stems from ADS data, again along rows. Over the course of these six years, impacts varied from a minimum of zero to a maximum of 5009 stems in an east-west transect; we use these impacted stem counts as a surrogate for MPB abundances ($N(x,t)$, Figure 2b) in what follows. In total, there are 2048 north-south 30 meter spatial units annually over a six year period (2000–2005). This results in 12,288 total response observations.

4.4. MPB MODEL

Recall the original ecological diffusion equation (1.2) described in the previous sections:

$$\frac{\partial u(x,t)}{\partial t} = \frac{\partial^2}{\partial x^2}\big(\delta(x)u(x,t)\big).$$

This deterministic model describes a mechanism for the spread of a population of animals through space while allowing them to congregate in areas where motility is low (e.g., preferable habitat). In the case of the MPB epidemic, the MPB population is not only spreading out, but also growing. Thus, consider an ecological diffusion model with a growth component

$$\frac{\partial u(x,t)}{\partial t} = \frac{\partial^2}{\partial x^2}\big(\delta(x)u(x,t)\big) + \gamma(t)u(x,t). \tag{4.2}$$

This model is a generalization of (1.2), containing a growth parameter $\gamma(t)$ that is allowed to vary over time, and would normally require a new analytical homogenization procedure (as in Section 2.1). An alternative specification would allow the growth parameter $\gamma(x)$ to vary in space rather than time, and though that model is not necessary for our MPB example, we present the details of the homogenization procedure in Supplementary Appendix A for the interested reader. In the case of the MPB model (4.2), however, a very useful result arises if we introduce a change of variables such that

$$z(x,t) = \exp\left(\int_0^t \gamma(v)\,dv\right)u(x,t). \tag{4.3}$$

Then the following PDE for the new variable z results:

$$\frac{\partial z(x,t)}{\partial t} = \frac{\partial^2}{\partial x^2}\big(\delta(x)z(x,t)\big). \tag{4.4}$$

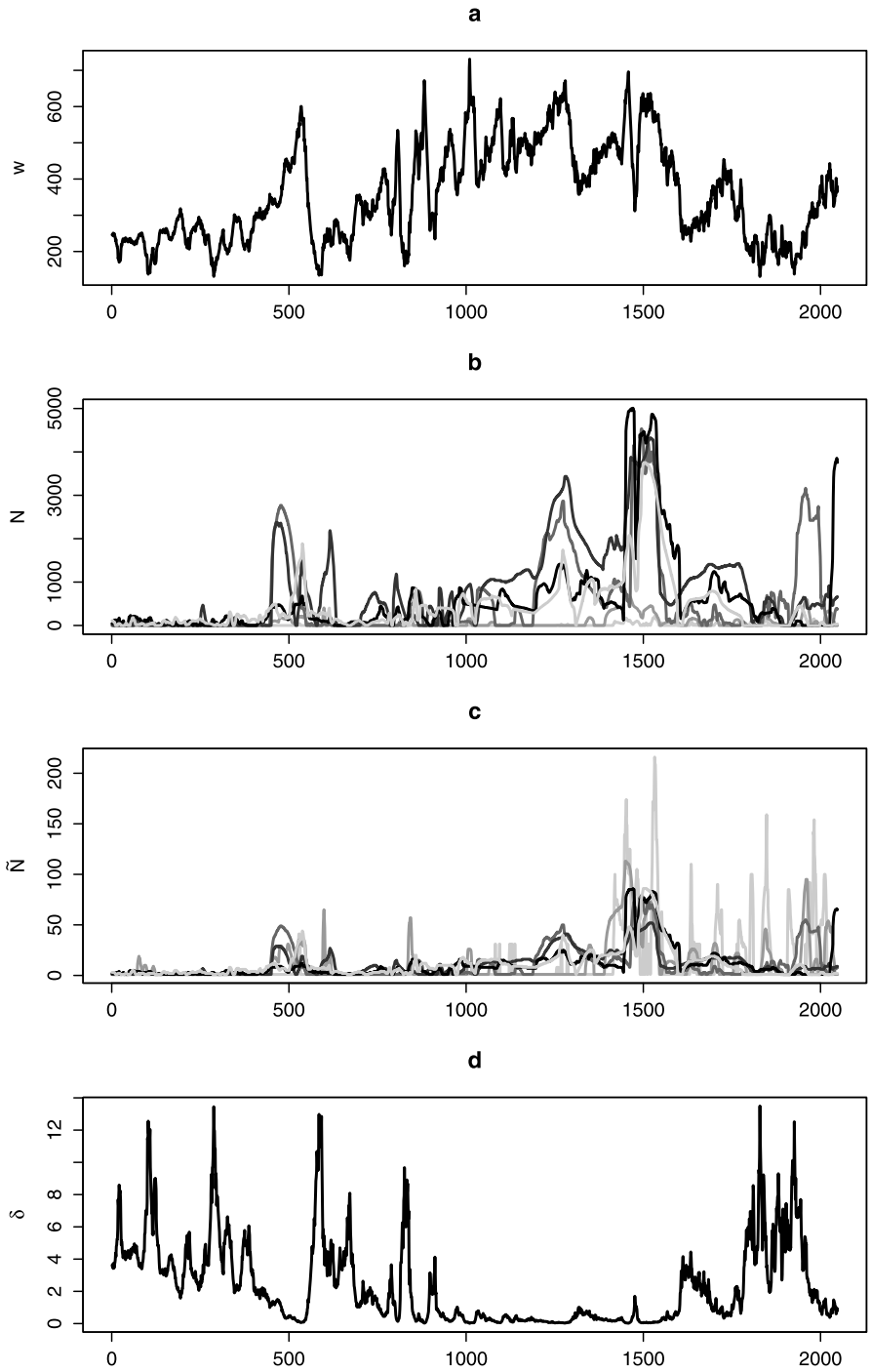

Figure 2. (a) Host abundance (i.e., w in the model) oriented such that North is right and South is left on the x-axis of the figure. (b) MPB abundance (i.e., N); shading represents year (light= 2000, dark= 2005). (c) Scaled MPB abundance ($\tilde{N}$). (d) Posterior mean MPB motility (i.e., spatial diffusion coefficients δ).

Note that this is exactly the ecological diffusion PDE (1.2) for which we have already derived the appropriate homogenization procedure. We merely need to consider the new variable z instead of the population intensity u directly. If we are interested only in inference on motility, then the new PDE (4.2) and the change of variables (4.3) implies that we can standardize our original MPB abundance data $N(x,t)$ such that the growth signal is removed. In effect, this rescales each year of data so that they are on the scale of year 2000 total abundance levels; this is the first year for which we have data. In matrix notation, the resulting statistical model for the standardized MPB abundance data $\tilde{N}(x,t)$ (Figure 2c) using homogenization can be written as:

$$\tilde{\mathbf{N}}(t) \sim \text{Pois}\big(f_h\big(\mathbf{z}_0(t-\Delta_t), \bar{\boldsymbol{\delta}}(\mathbf{w}, \boldsymbol{\beta})\big)\big), \tag{4.5}$$

where, f_h is the homogenized PDE solver for (4.4) and we let $\delta(x) = \exp\{\beta_0 + \beta_1 w(x)\}$, thus allowing the motilities (or diffusion coefficients, $\delta(x)$) to depend on host abundance $w(x)$. We used a multivariate Gaussian prior for $\boldsymbol{\beta} = (\beta_0, \beta_1)'$ with mean vector equal to zero and covariance matrix equal to $100 \times \mathbf{I}$.

4.5. RESULTS

The statistical ecological diffusion model (4.5) described in Section 4.4 was fit to the scaled MPB data (Figure 2c) using the MCMC algorithm from Section 2.3 with 20,000 MCMC iterations (assessing convergence visually and discarding the first 2000 MCMC samples as burn-in). Due to the large size of the MPB data set and necessary computational integration, we used homogenization with a spatial averaging window of $(1/256)$th of the spatial domain (or 8 spatial units in this case). The posterior histograms for $\boldsymbol{\beta}$ shown in Figure 3 indicate the effect of host abundance (w) on MPB motility (δ).

The intercept coefficient β_0 (posterior mean: 4.102) influences the overall motility while β_1 (posterior mean: -0.012) indicates that host abundance (w) is negatively related to MPB motility. That is, when the amount of susceptible forest is large, the MPB motilities are slower, allowing them to aggregate in desirable areas. The effect of host abundance (w) on MPB motility can be visualized by assessing the posterior mean for δ (Figure 2d). In this case, it is evident that the motilities δ are negatively related to host abundance w (by comparing plots (a) and (d) in Figure 2). This supports the findings of previous modeling efforts that link MPB dispersal to host abundance (e.g., Powell, Logan, and Bentz 1996; Logan et al. 1998; Biesinger et al. 2000) but in a spatially explicit context.

5. DISCUSSION

In summary, by taking a mathematical approach to change of support in PDE models, we can obtain similar statistical inference on parameters controlling the dynamics of the process with a fraction of the computational effort. Though traditionally employed in engineering settings, the specific type of singular perturbation theory known as homogenization has both a statistical interpretation and utility. Homogenization allows for an accurate and efficient implementation of PDE models while still accommodating small scale structure in

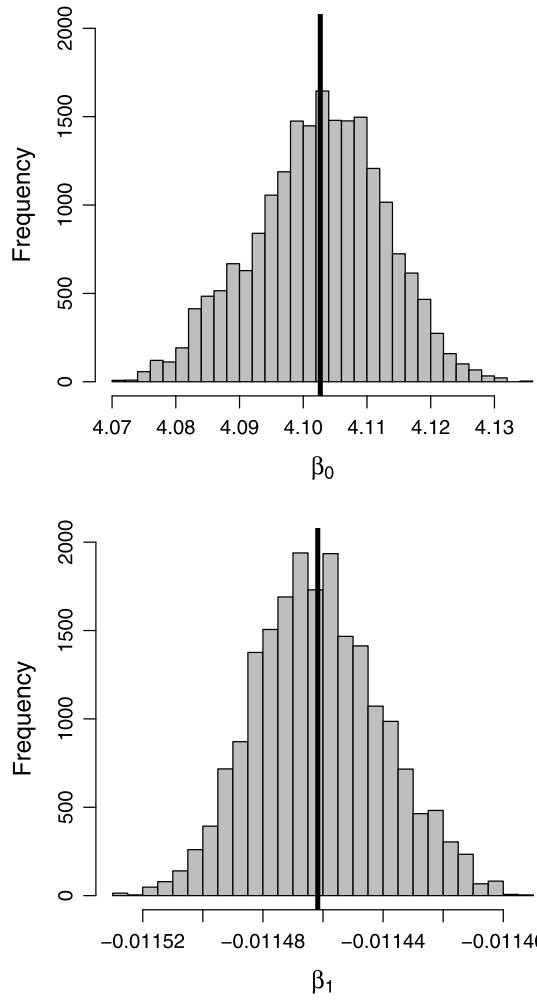

Figure 3. Posterior histograms of the regression coefficients $\boldsymbol{\beta}$ influencing the diffusion coefficients $\boldsymbol{\delta}$. The posterior mean is represented by the bold vertical line on each plot.

the underlying environment and is harmonious with existing multi-scale statistical methods. In the approximation procedure for ecological diffusion, a natural, but uncommon, statistical quantity arises as a means for changing the support (i.e., the harmonic mean), and, as a byproduct, we are left with a smoother dynamical system that can be implemented on a coarser spatial scale. Through simulation, we have demonstrated the use and effectiveness of this method and illustrated the potential savings in computation.

We have demonstrated, with both simulated and real data involving large-scale population spread, how statistical dynamic models can be fit over extensive spatial support using large data sets via the homogenization-based upscaling. Our example pertaining to MPB spread links insect dispersal to host abundance using well-founded mechanistic dynamics in a statistically rigorous manner. In the current age of digital information, a data set like ours which is on the order of tens of thousands of observations could only be con-

sidered moderately large; however, it is important to remember that the data are coupled with a complicated dynamic model that may require hundreds or thousands of computational steps to integrate the process forward in time 6 years. In our case, the underlying process itself is massive and could not be explicitly modeled without some creative statistical dimension reduction. Fortunately, homogenization provides an optimal computational upscaling method that is model-specific but retains the small-scale variability appearing in the data.

In terms of potential extensions to this methodology, choosing the optimal degree of averaging itself (the ϵ in our specification) is a subject of ongoing research. The optimal scale over which to average is a function of both desired computational savings and loss associated with the estimation of model parameters. Most studies would put a much higher weight on estimation accuracy than on computational savings, thus, a potential method that could be employed is to first choose a threshold or tolerance on the approximation error $\|u(x,t) - u_0(x,t)\|$, for all x and t, then find the smallest ϵ that meets the tolerance based on exploratory forward solution of both the original and homogenized PDEs. A promising alternative might be to "model" the ϵ such that the data are used to help find the scale of averaging that provides the maximum information about the parameters while minimizing the computational time.

A distinct but related set of alternative approaches for statistical inference using complicated deterministic process models are the focus of an active area of recent research. One such method is referred to as "Bayesian melding," and employs a clever trick involving implicit and explicit priors on the model inputs and outputs and is implemented using a form of sampling-importance-resampling algorithm (e.g., Poole and Raftery 2000). Another popular approach to fitting computer models is through the use of first-order (e.g., Hooten et al. 2011) and second-order (e.g., Higdon et al. 2008) statistical emulators. These methods allow for inference through an a priori computer experiment and often some modeling of the input–output relationships in a spectral domain. The homogenization approach described here is essentially an emulator. However, rather than phenomenologically mimicking the diffusion model, it actually custom tailors a model-specific mechanistic emulator that retains small-scale structure in the output; this is difficult to achieve using Gaussian process emulators. The model-specific nature of the homogenization approach can also be a disadvantage in some situations. It is not a black box procedure, but rather a custom procedure that requires analytical analysis prior to model fitting. Therefore, it may not be appropriate nor possible in all settings, such as highly complicated agent-based models.

Finally, it is important to note that perturbation theory is applicable to most differential equations. In fact, in the epidemiological/ecological setting, we have found that a similar approach yields advantageous results in the implementation of 2-D ecological diffusion PDEs with growth terms as well as more complicated coupled PDE models (Garlick, Powell, and Hooten 2011). Similarly, the diffusion parameters that govern the rates of movement for individuals could also vary in time (as indicated in (1.2)) and homogenization can be useful for finding optimal upscaling statistics in the temporal domain as well. Other applications of the methodology presented here include modeling of spatio-temporal atmospheric and environmental processes as well as epidemiological processes where the spread of disease is a primary concern.

ACKNOWLEDGEMENTS

This research was funded by USGS 1434-06HQRU1555. Any use of trade names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

REFERENCES

Aarts, G., Fieberg, J., and Matthiopoulos, J. (2012), "Comparative Interpretation of Count, Presence–Absence and Point Methods for Species Distribution Models," *Methods in Ecology and Evolution*, 3, 177–187.

Alfaro, R., Campbell, R., Vera, P., Hawkes, B., and Shore, T. (2004), "Dendroecological Reconstruction of Mountain Pine Beetle Outbreaks in the Chilcotin Plateau of British Columbia," in *Mountain Pine Beetle Symposium: Challenges and Solutions*, eds. T. L. Shore, J. E. Brooks, and J. E. Stone, pp. 245–256. Natural Resources Canada, Information Report BC-X-399.

Aukema, B. H., Carroll, A. L., Zhu, J., Raffa, K. F., Sickley, T. A., and Taylor, S. W. (2006), "Landscape Level Analysis of Mountain Pine Beetle in British Columbia, Canada: Spatiotemporal Development and Spatial Synchrony Within the Present Outbreak," *Ecography*, 29, 427–441.

Aukema, B. H., Carroll, A. L., Zheng, Y., Zhu, J., Raffa, K. F., Moore, R. D., Stahl, K., and Taylor, S. W. (2008), "Movement of Outbreak Populations of Mountain Pine Beetle: Influences of Spatiotemporal Patterns and Climate," *Ecogeography*, 31, 348–358.

Bentz, B. J. (2006), "Mountain Pine Beetle Population Sampling: Inferences from Lindgren Pheromone Traps and Tree Emergence Cages," *Canadian Journal of Forest Research*, 36, 351–360.

Bentz, B. J., Logan, J. A., and Amman, G. D. (1991), "Temperature Dependent Development of the Mountain Pine Beetle (Coleoptera: Scolytidae), and Simulation of Its Phenology," *Canadian Entomologist*, 123, 1083–1094.

Bentz, B. J., and Mullins, D. E. (1999), "Ecology of Mountain Pine Beetle (Coleoptera: Scolytidae) Cold Hardening in the Intermountain West," *Environmental Entomology*, 28, 577–587.

Bentz, B. J., Powell, J. A., and Logan, J. A. (1996), "Localized Spatial and Temporal Attack Dynamics of the Mountain Pine Beetle (*Dendroctonus Ponderosae*) in Lodgepole Pine," USDA/FS Research Paper INT-RP-494, December, 1996.

Bentz, B. J., Régnière, J., Fettig, C. J., Hansen, E. M., Hayes, J. L., Hicke, J. A., and Seybold, S. J. (2010), "Climate Change and Bark Beetles of the Western United States and Canada: Direct and Indirect Effects," *BioScience*, 60, 427–613.

Berliner, L. M. (1996), "Hierarchical Bayesian Time-Series Models," in *Maximum Entropy and Bayesian Methods*, Amsterdam: Kluwer Academic, pp. 15–22.

Biesinger, Z., Powell, J. A., Bentz, B. J., and Logan, J. A. (2000), "Direct and Indirect Parameterization of a Localized Model for the Mountain Pine Beetle Lodgepole Pine System," *Ecological Modelling*, 129, 273–296.

Blackard, J. A., Finco, M. V., Helmer, E. H., Holden, G. R., Hoppus, M. L., Jacobs, D. M., Lister, A. J., Moisen, G. G., Nelson, M. D., Riemann, R., Ruefenacht, B., Salajanu, D., Weyermann, D. L., Winterberger, K. C., Brandeis, T. J., Czaplewski, R. L., McRoberts, R. E., Patterson, P. L., and Tycio, R. P. (2008), "Mapping U.S. Forest Biomass Using Nationwide Forest Inventory Data and Moderate Resolution Information," *Remote Sensing of Environment*, 112, 1658–1677.

Boone, C. K., Aukema, B. H., Bohlmann, J., Carroll, A. L., and Raffa, K. F. (2011), "Efficacy of Tree Defense Physiology Varies with Bark Beetle Population Density: A Basis for Positive Feedback in Eruptive Species," *Canadian Journal of Forest Research*, 41, 1174–1188.

Borden, J. H., Ryker, L. C., Chong, L. J., Pierce, H. D., Johnston, B. D., and Oehlschlage, A. C. (1987), "Response of the Mountain Pine Beetle, *Dendroctonus Ponderosae*, to Five Semiochemicals in British Columbia Lodgepole Pine Forests," *Canadian Journal of Forest Research*, 17, 118–128.

Cangelosi, A. R., and Hooten, M. B. (2009), "Models for Bounded Systems with Continuous Dynamics," *Biometrics*, 65, 850–856.

Crabb, B. A., Powell, J. A., and Bentz, B. J. (2012), "Development and Assessment of 30-m Pine Density Maps for Landscape-Level Modeling of Mountain Pine Beetle Dynamics," Research Paper RMRS-RP-96WWW. Fort Collins, CO: U.S. Department Beetle Dynamics. Res. Pap. RMRS-RP-96WWW. Fort Collins, CO: U.S. Department of Agriculture, Forest Service, Rocky Mountain Research Station, p. 43.

Cressie, N. A. C., and Wikle, C. K. (2010), *Statistics for Spatio-Temporal Data*, Hoboken, NJ: Wiley.

Cressie, N. A. C., Calder, C. A., Clark, J. S., Ver Hoef, J. M., and Wikle, C. K. (2009), "Accounting for Uncertainty in Ecological Analysis: The Strengths and Limitations of Hierarchical Statistical Modeling," *Ecological Applications*, 19, 553–570.

Crookston, N. L., Stark, R. W., and Adams, D. L. (1977), "Outbreaks of Mountain Pine Beetle in Northwestern Lodgepole Pine Forests—1945 to 1975," Forest, Wildlife and Range Experiment Station Bulletin No. 22. University of Idaho, Moscow, p. 7.

Ferreira, M. A. R., and Lee, H. K. H. (2007), *Multiscale Modeling, a Bayesian Perspective*, New York: Springer.

Ferreira, M. A. R., Higdon, D., Lee, H. K. H., and West, M. (2006), "Multiscale and Hidden Resolution Time Series Models," *Bayesian Analysis*, 1, 947–968.

Fisher, R. A. (1937), "The Wave of Advance of Advantageous Genes," *Annals of Eugenics*, 7, 355–369.

Garlick, M. J., Powell, J. A., and Hooten, M. B. (2011), "Homogenization of Large-Scale Movement Models in Ecology," *Bulletin of Mathematical Biology*, 73, 2088–2108.

Gilbert, E., Powell, J. A., Logan, J. A., and Bentz, B. J. (2004), "Comparison of Three Models Predicting Developmental Milestones Given Environmental and Individual Variation," *Bulletin of Mathematical Biology*, 66, 1821–1850.

Gotway, C. A., and Young, L. J. (2002), "Combining Incompatible Spatial Data," *Journal of the American Statistical Association*, 97, 632–648.

Halsey, R. (1998), Aerial Detection Survey Metadata for the Intermountain Region 4. U.S. Department of Agriculture Forest Service, Forest Health Protection.

Heavilin, J., and Powell, J. A. (2008), "A Novel Method for Fitting Spatio-Temporal Models to Data, with Applications to the Dynamics of Mountain Pine Beetle," *Natural Resource Modelling*, 21, 489–524.

Heavilin, J., Powell, J. A., and Logan, J. A. (2007), "Development and Parameterization of a Model for Bark Beetle Disturbance in Lodgepole Forest," in *Plant Disturbance Ecology*, eds. K. Miyanishi and E. Johnson, New York: Academic Press, pp. 527–553.

Higdon, D., Gattiker, J., Williams, B., and Rightley, M. (2008), "Computer Model Calibration Using High-Dimensional Output," *Journal of the American Statistical Association*, 103, 570–583.

Holmes, M. H. (1995), *Introduction to Perturbation Methods*, New York: Springer.

Hooten, M. B., and Wikle, C. K. (2007), "Shifts in the Spatio-Temporal Growth Dynamics of Shortleaf Pine," *Environmental and Ecological Statistics*, 14, 207–227.

——— (2008), "A Hierarchical Bayesian Non-linear Spatio-Temporal Model for the Spread of Invasive Species with Application to the Eurasian Collared-Dove," *Environmental and Ecological Statistics*, 15, 59–70.

Hooten, M. B., Leeds, W. B., Fiechter, J., and Wikle, C. K. (2011), "Assessing First-Order Emulator Inference for Physical Parameters in Nonlinear Mechanistic Models," *Journal of Agricultural, Biological, and Environmental Statistics*, 16, 475–494.

Hotelling, H. (1927), "Differential Equations Subject to Error," *Journal of the American Statistical Association*, 22, 283–314.

Hughes, R. R. (1973), "Dendroctonus: Production of Pheromones and Related Compounds in Response to Host Monoterpenes," *Zeitschrift für Angewandte Entomologie*, 73, 294–312.

Lindgren, F., Rue, H., and Lindstrom, J. (2011), "An Explicit Link Between Gaussian Fields and Gaussian Markov Random Fields: The SPDE Approach (with Discussion)," *Journal of the Royal Statistical Society. Series B*, 73, 423–498.

Logan, J. A., and Powell, J. A. (2001), "Ghost Forests, Global Warming, and the Mountain Pine Beetle (Coleoptera: Scolytidae)," *American Entomologist*, 47 (3), 160–173.

Logan, J. A., White, P., Bentz, B. J., and Powell, J. A. (1998), "Model Analysis of Spatial Patterns in Mountain Pine Beetle Outbreaks," *Theoretical Population Biology*, 53 (3), 236–255.

McCambridge, W. F. (1967), "Nature of Induced Attacks by the Black Hills Beetle, *Dendroctonus Ponderosae* (Coleoptera: Scolytidae)," *Annals of the Entomological Society of America*, 64, 534–535.

McGregor, M. D. (1978), "Status of Mountain Pine Beetle Glacier National Park and Glacier View Ranger District, Flathead National Forest, MT, 1977," Forest Insect and Disease Management Report No. 78-6, Missoula, MT.

Mitchell, A. R., and Griffiths, D. F. (1980), *The Finite Difference Method in Partial Differential Equations*, New York: Wiley.

Mitchell, R. G., and Preisler, H. K. (1991), "Analysis of Spatial Patterns of Lodgepole Pine Attacked by Outbreak Populations of the Mountain Pine Beetle," *Forest Science*, 37 (5), 1390–1408.

Murray, J. D. (2002), *Mathematical Biology* (3rd ed.), New York: Springer.

Okubo, A., and Levin, S. A. (2001), *Diffusion and Ecological Problems: Modern Perspectives* (2nd ed.), New York: Springer.

Pavliotis, G. A., and Stuart, A. M. (2008), *Multiscale Methods: Averaging and Homogenization*, New York: Springer.

Perkins, D. L., and Swetnam, T. W. (1996), "A Dendroecological Assessment of Whitebark Pine in the Sawtooth-Salmon River Region, Idaho," *Canadian Journal of Forest Research*, 26, 2123–2133.

Pitman, G. B. (1971), "Trans-Verbenol and Alpha-Pinene: Their Utility in Manipulation of the Mountain Pine Beetle," *Journal of Economic Entomology*, 64, 426–430.

Poole, D., and Raftery, A. E. (2000), "Inference for Deterministic Simulation Models: The Bayesian Melding Approach," *Journal of the American Statistical Association*, 95, 1244–1255.

Powell, J. A., and Bentz, B. J. (2009), "Connecting Phenological Predictions with Population Growth Rates for Mountain Pine Beetle, an Outbreak Insect," *Landscape Ecology*, 24, 657–672.

Powell, J. A., Logan, J. A., and Bentz, B. J. (1996), "Local Projections for a Global Model of Mountain Pine Beetle Attacks," *Journal of Theoretical Biology*, 179 (3), 243–260.

R Development Core Team (2012), *R: A Language and Environment for Statistical Computing*, Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0, URL *http://www.R-project.org/*

Raffa, K. F., Phillips, T. W., and Salom, S. M. (1993), "Strategies and Mechanisms of Host Colonization by Bark Beetles," in *Beetle-Pathogen Interactions in Conifer Forests*, eds. T. D. Schowalter and G. M. Filip, New York: Academic Press, pp. 103–120.

Raffa, K. F., Aukema, B. H., Bentz, B. J., Carroll, A. L., Hicke, J. A., Turner, M. G., and Romme, W. H. (2008), "Cross-Scale Drivers of Natural Disturbances Prone to Anthropogenic Amplification: Dynamics of Biome-Wide Bark Beetle Eruptions," *BioScience*, 58 (6), 501–518.

Régnierè, J., and Bentz, B. J. (2007), "Modeling Cold Tolerance in the Mountain Pine Beetle, *Dendroctonus Ponderosae*," *Journal of Insect Physiology*, 53 (6), 559–572.

Risken, H. (1989), *The Fokker-Planck Equation: Methods of Solution and Applications*, New York: Springer.

Robertson, C., Nelson, T. A., and Boots, B. (2007), "Mountain Pine Beetle Dispersal: The Spatial-Temporal Interaction of Infestations," *Forestry Science*, 53, 395–405.

Robertson, C., Nelson, T. A., Jelinski, D. E., Wulder, M. A., and Boots, B. (2009), "Spatial-Temporal Analysis of Species Range Expansion: The Case of the Mountain Pine Beetle, *Dendroctonus Ponderosae*," *Journal of Biogeography*, 36 (8), 1446–1458.

Royle, J. A., and Wikle, C. K. (2005), "Efficient Statistical Mapping of Avian Count Data," *Environmental and Ecological Statistics*, 12, 225–243.

Safranyik, L., Linton, D. A., Silversides, R., and McMullen, L. H. (1992), "Dispersal of Released Mountain Pine Beetles Under the Canopy of a Mature Lodgepole Pine Stand," *Journal of Applied Entomology*, 113, 441–450.

Safranyik, L., Carroll, A. L., Régnière, J., Langor, D. W., Riel, W. G., Shore, T. L., Peter, B., Cooke, B. J., Nealis, V. G., and Taylor, S. W. (2010), "Potential for Range Expansion of Mountain Pine Beetle Into the Boreal Forest of North America," *The Canadian Entomologist*, 142, 415–442.

Sambaraju, K. R., Carroll, A. L., Zhu, J., Stahl, K., Moore, R. D., and Aukema, B. H. (2012), "Climate Change Could Alter the Distribution of Mountain Pine Beetle Outbreaks in Western Canada," *Ecography*, 35, 211–223.

Turchin, P. (1998), *Quantitative Analysis of Movement*, Sunderland, MA: Sinauer Associates, Inc. Publishers.

Warton, D. I., and Shepherd, L. C. (2010), "Poisson Point Process Models Solve the "Pseudo-absence Problem" for Presence-Only Data in Ecology," *The Annals of Applied Statistics*, 4, 1383–1402.

Wikle, C. K. (2003), "Hierarchical Bayesian Models for Predicting the Spread of Ecological Processes," *Ecology*, 84, 1382–1394.

——— (2010), "Low-Rank Representations for Spatial Processes," in *Handbook of Spatial Statistics*, eds. A. E. Gelfand, P. J. Diggle, M. Fuentes, and P. Guttorp, Boca Raton: CRC Press, pp. 107–118.

Wikle, C. K., and Berliner, L. M. (2005), "Combining Information Across Spatial Scales," *Technometrics*, 47, 80–91.

Wikle, C. K., Berliner, L. M., and Milliff, R. F. (2003), "Hierarchical Bayesian Approach to Boundary Value Problems with Stochastic Boundary Conditions," *Monthly Weather Review*, 131, 1051–1062.

Wikle, C. K., and Hooten, M. B. (2010), "A General Science-Based Framework for Nonlinear Spatio-Temporal Dynamical Models," *Test*, 19, 417–451.

Wikle, C. K., Milliff, R. F., Nychka, D., and Berliner, L. M. (2001), "SpatioTemporal Hierarchical Bayesian Modeling: Tropical Ocean Surface Winds," *Journal of the American Statistical Association*, 96, 382–397.

Zheng, Y., and Aukema, B. (2010), "Hierarchical Dynamic Modeling of Outbreaks of Mountain Pine Beetle Using Partial Differential Equations," *EnvironMetrics*, 21, 801–816.