Ecological Applications, 21(4), 2011, pp. 1173–1188

Reconciling multiple data sources to improve accuracy of large-scale prediction of forest disease incidence

Ephraim M. Hanks,[1,4] Mevin B. Hooten,[2] and Fred A. Baker[3]

[1]*Department of Statistics, Colorado State University, Fort Collins, Colorado 80523 USA*
[2]*U.S. Geological Survey, Colorado Cooperative Fish and Wildlife Research Unit, Colorado State University, Fort Collins, Colorado 80521 USA*
[3]*Department of Wildland Resources, The Ecology Center, Utah State University, Logan, Utah 84322-5230 USA*

Abstract. Ecological spatial data often come from multiple sources, varying in extent and accuracy. We describe a general approach to reconciling such data sets through the use of the Bayesian hierarchical framework. This approach provides a way for the data sets to borrow strength from one another while allowing for inference on the underlying ecological process. We apply this approach to study the incidence of eastern spruce dwarf mistletoe (*Arceuthobium pusillum*) in Minnesota black spruce (*Picea mariana*). A Minnesota Department of Natural Resources operational inventory of black spruce stands in northern Minnesota found mistletoe in 11% of surveyed stands, while a small, specific-pest survey found mistletoe in 56% of the surveyed stands. We reconcile these two surveys within a Bayesian hierarchical framework and predict that 35–59% of black spruce stands in northern Minnesota are infested with dwarf mistletoe.

Key words: Arceuthobium pusillum; *Bayesian hierarchical models; black spruce; disease monitoring; dwarf mistletoe; multiple data sources; northern Minnesota, USA;* Picea mariana; *spatial autocorrelation.*

Introduction

Reliable ecological data can be difficult and costly to obtain, especially over large landscapes, yet sound management and science require accurate information. No data are perfect in capturing the true ecological state of the system being studied. Recognizing this, much research has focused on accounting for inaccuracy in the data-collection process. For example, models of detection accuracy for wildlife occupancy have received significant attention in recent years (e.g., MacKenzie et al. 2003, Tyre et al. 2003, Gu and Swihart 2004, Royle 2004, MacKenzie et al. 2005), and the Bayesian hierarchical framework has been touted for its ability to model uncertainty in the data collection process separately from uncertainty in the ecological process (e.g., Hooten et al. 2003, Ogle 2009).

Increasing the accuracy of predictions made from an existing set of observations is typically accomplished by gathering new data to replace or supplement the existing data. Including new observations can decrease the variance of predictions made from the data or can allow for a better characterization of measurement bias, thus increasing the reliability of predictions. For example, replicate observations of the same ecological process can be used to more rigorously account for errors in the data collection process (e.g., Royle and Nichols 2003) and can result in greater overall prediction accuracy even though there has been no increase in the reliability of the individual surveys.

Some data sets, however, cannot easily be replicated. The expense and time involved in repeating large-scale monitoring efforts to increase the accuracy of predictions can be prohibitive, and alternate methods must be used. Focusing resources on a small survey can generate more reliable data, as time-consuming techniques and more expensive data-gathering processes can be more easily employed. In this study, we show how multiple data sets, with varying strengths and weaknesses, can be combined to yield improved inference concerning the process of interest. This approach allows us to compare and reconcile sets of observations within a Bayesian hierarchical modeling framework. In our application, involving forest pathogen incidence, the accuracy of predictions based on an extensive forest inventory are improved through reconciliation with a small, specific-pest survey.

Bayesian data reconciliation

In what follows, we present an approach for reconciling two sets of observations within a Bayesian hierarchical modeling (BHM) framework. We have chosen to call this particular application of BHMs "Bayesian data reconciliation." We assume that the response variable of interest has been observed in two separate surveys, one of which, denoted here as D_A, is assumed to be more accurate than the other, denoted as D_L. We assume that the observations in D_A are collected

Manuscript received 27 August 2009; revised 6 July 2010; accepted 15 July 2010; final version received 24 August 2010. Corresponding Editor: D. McKenzie.

[4] E-mail: hanks@stat.colostate.edu

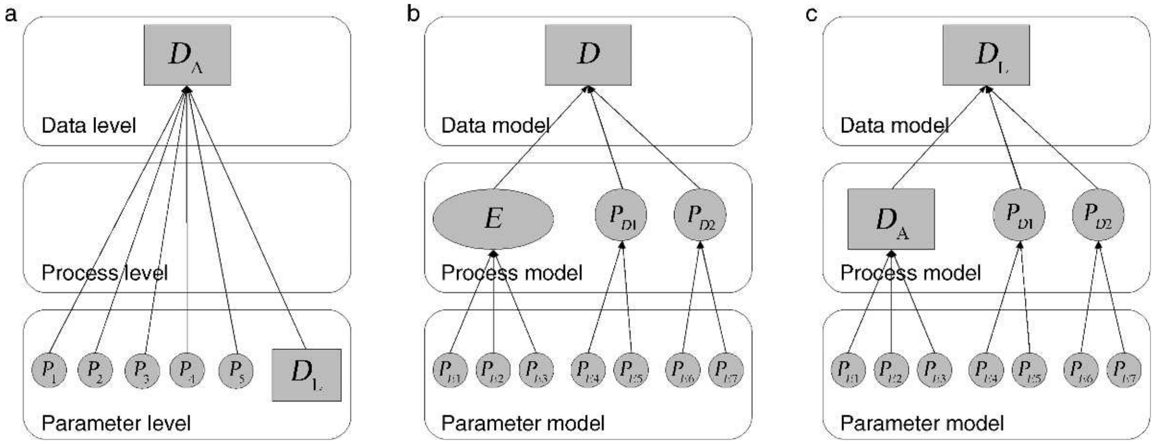

FIG. 1. Comparison of traditional prediction methods such as linear regression and regression trees with the Bayesian hierarchical modeling (BHM) framework and the Bayesian data reconciliation (BDR) approach. (a) In attempting to reconcile two data sets, one (D_A) more accurate than the other (D_L), traditional prediction methods such as linear regression and regression trees make no distinction between the effect of the less accurate data (D_L) and the effects of environmental covariates (P). (b) Traditionally, BHMs use one set of data (D) to make inference about an unobserved ecological process (E). (c) BDR utilizes a BHM by replacing the latent ecological process (E) with the more accurate set of data (D_A). In this way, a small, accurate set of data (D_A) can be used to update and improve the accuracy of a large, inaccurate set of data (D_L).

on a representative subset of the observation locations for which we have the less accurate set of data, D_L, and on the same spatial scale. When the situation allows for a selection of the survey locations for D_A, optimal sampling methods (e.g., Hooten and Wikle 2009) could be employed. We seek to combine the accuracy of D_A with the scale and extent of the large, less accurate D_L. In an analogous generic ecological study we might collect data from a sample of locations across a landscape, fit a statistical model of an ecological process using the collected data, and then use the results of the statistical analysis to make predictions at unsurveyed locations across the landscape. Similarly, in the data reconciliation situation described above, we can construct a statistical model that specifies the relationship between the accurate data D_A, the inaccurate data D_L, and the relevant ecological process. We can then use the results to predict what would be found if the more accurate survey were conducted at all locations across the extent of D_L.

Statistical methods commonly used for prediction include linear regression, logistic regression, classification and regression trees (e.g., Breiman et al. 1984), and random forests (Breiman 2001). For a given set of data, these methods may produce differing levels of accuracy, and it is common to apply multiple methods to the problem and choose the one that delivers the best accuracy, often measured by cross-validation (e.g., Rejwan et al. 1999). In the context of reconciling data, D_L could be conditioned on as an independent variable in one of these traditional methods with D_A as the dependent variable (Fig. 1a). One weakness shared by these traditional methods is the lack of distinction in the prediction model between the effect of environmental covariates and the effect of the less accurate data. While this may not adversely affect the accuracy of predictions made from the method, it makes little ecological sense and can lead to difficulties in interpreting the results of the inference made on the parameters in the model.

In contrast, hierarchical models (e.g., Cressie et al. 2009) are well-suited to the task of synthesizing multiple data sets, such as those described here, because of the flexibility and interpretability they provide in modeling interrelated processes. Specifying the relationship between the two data sets separately from the ecological process allows us to model each process in a scientifically meaningful way, and the hierarchical modeling framework allows us to link these separate processes and rigorously make inferences about both processes simultaneously. Cressie et al. (2009) provide an excellent summary of the strengths and limitations of hierarchical statistical modeling in ecology, especially Bayesian hierarchical models (BHMs), and interested readers are referred to that paper for a full treatment of the subject.

In brief, the traditional BHM is a series of three linked statistical models (Berliner 1996), each dependent on the next (Fig. 1b). The data model links the observed data, D, to the true, but usually unobserved, ecological state of interest, E, through data model parameters P_D. The process model describes the ecological process that gives rise to this latent ecological state, often through the use of parameters, P_E, related to environmental covariates. The parameter model specifies the prior knowledge about the parameters in the data and process models. The observed data, D, is used as the response variable in the data model and linked to corresponding environmental variables in the process model through the latent, true ecological state, E. Thus the available data and the environmental covariates are used to make inference about the true, latent ecological state. Prediction can be made within this framework through the posterior predictive distribution at locations for which we lack observations of the response variable, but have all information needed for the process model (e.g., environmental covariates). This posterior predictive distribution can be found using composition sampling (e.g., Banerjee et al. 2003) simultaneously with the joint posterior distribution of the model parameters using Markov chain Monte Carlo (MCMC) techniques.

Bayesian data reconciliation (BDR) fits within the traditional BHM framework, with one key distinction: our goal is no longer to use one set of observed data to make inference about a latent ecological state, but instead to reconcile two data sets, one more reliable than the other. The data reconciliation process is as follows: the data model links the less accurate data, D_{L}, to the more accurate data, D_{A}, using model parameters, P_D. The more accurate data, and possibly the data reconciliation parameters, P_D, are then in turn linked to environmental parameters P_E within the process model (Fig. 1c). Thus, BDR is, in effect, a BHM in which the latent process is actually observed ($D_{\mathrm{A}} \equiv E$).

The data model represents the statistical relationship of D_{L} to the more accurate survey, D_{A}. In practice, thinking of D_{L} as a noisy version of D_{A} can aid the choice of an appropriate data model. The form of the BDR data model could be identical to a traditional BHM data model linking observations to a latent ecological state, though the interpretation would be different. Instead of directly representing the data collection process, the data model in BDR represents the differences between the two collected data sets. This could include factors such as differences in detection between D_{A} and D_{L}, temporal change in the ecological process between when D_{A} and D_{L} were observed, or variation in spatial resolution from data collected at differing scales.

The process model in BDR specifies the ecological process that gives rise to D_{A}, typically relating D_{A} to the coefficients P_E of environmental covariates. Typically these covariates are assumed to be fixed and known, but if a choice must be made between using environmental covariates collected from the more or less accurate survey, use of the covariates from the large, less accurate survey allows for straightforward prediction at locations not present in the small, more accurate survey. If there is reason to assume that these covariates are also less accurate than those collected in the smaller survey, the covariates from the more accurate survey can be used in the process model as long as a statistical model describing the relationship of the inaccurate covariates to these more accurate covariates is specified.

The BDR framework allows us to borrow strength from the reliability of the more accurate set of data, D_{A}, to make predictions through the posterior predictive distribution at any location for which we have D_{L} and any environmental covariates needed for the process model. Thus, within this framework we can use a small, accurate source of data to update and improve predictions made from a less accurate source of data across its range.

Application: Dwarf Mistletoe in Minnesota Black Spruce

To illustrate Bayesian data reconciliation we focus on the infestation of eastern spruce dwarf mistletoe (*Arceuthobium pusillum*), which causes the most serious disease of black spruce (*Picea mariana*) throughout its range (Baker et al. 2006) (see Plate 1). Black spruce is a valuable species used in the manufacture of high-quality paper. Dwarf mistletoe infestation reduces growth, longevity, and quality of host trees (Geils and Hawksworth 2002). Severe infestations of this dwarf mistletoe create large areas of mortality and can affect entire ecosystems. Due to the significant impact of these parasites in forests, information about the dwarf mistletoe infestation often drives management decisions on stand manipulations (e.g., Muir and Geils 2002, Reid and Shamoun 2009).

Using aerial photography, Anderson (1949) estimated that 3–19% of the black spruce in the Big Falls Management Unit in Minnesota was out of production due to dwarf mistletoe. More recently, USDA Forest Service Forest Inventory and Analysis (FIA) found dwarf mistletoe on 5% of plots in northern Minnesota, on the low end of what Anderson estimated nearly 60 years before. Additionally, the Minnesota Department of Natural Resources (DNR) forest inventory shows dwarf mistletoe in 11% of 46 415 black spruce stands (Fig. 2). F. A. Baker, M. Hansen, J. D. Shaw, M. Mielke, and D. N. Shelstad (*unpublished manuscript*) intensively surveyed spruce forests around FIA plots to characterize the ability of these operational inventories to detect dwarf mistletoe. This intensive survey (Fig. 3) was inspired by the low proportion of FIA sites in which mistletoe was found, and thus Baker et al. (F. A. Baker, M. Hansen, J. D. Shaw, M. Mielke, and D. N. Shelstad, *unpublished manuscript*) focused their survey on stands near FIA plots. Confidentiality required that the true FIA plot locations were approximated somewhere within 0.5 miles [1 mile = 1.6 km] of the true location and so Baker et al. surveyed 196 black spruce stands in the DNR forest inventory within 0.5 miles of 31 approximate FIA plot locations. Thus the intensive survey can be thought of as a confirmatory survey on a small subset of the same areal units (stands) in the DNR inventory. Of the 196 stands for which we have both the DNR and intensive survey information, 17% are reported by the DNR inventory to be infested with mistletoe, while the intensive survey found mistletoe in 56% of the same stands.

Forest inventories such as the one conducted by the Minnesota DNR traditionally focus on the forest type and volume present. Forest insects and diseases are often quite cryptic, and while inventory crews may be

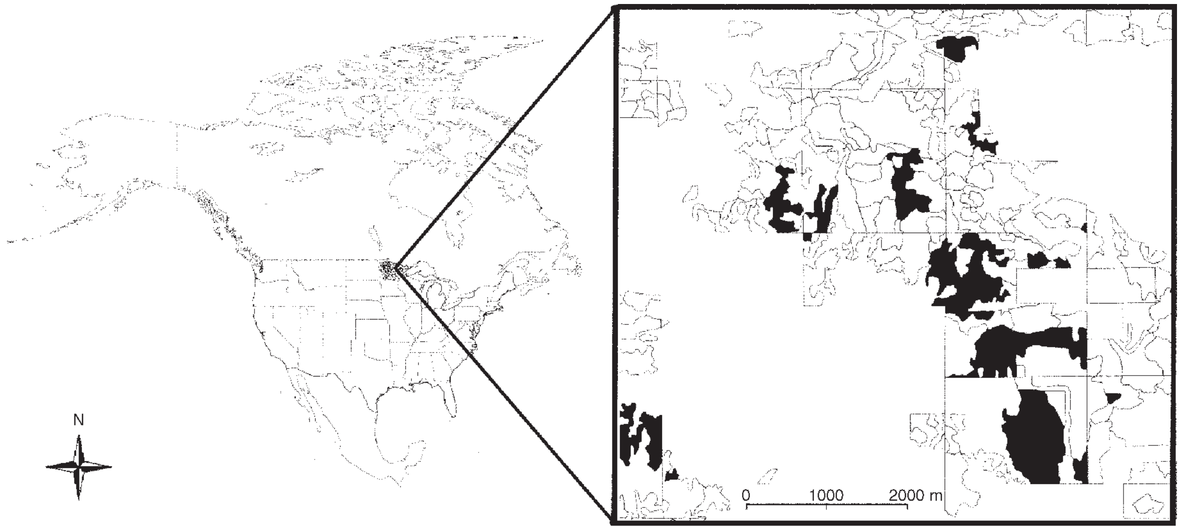

FIG. 2. The Minnesota Department of Natural Resources (DNR) inventory covers 46 415 black spruce (*Picea mariana*) stands across northern Minnesota, USA. At each stand, >40 stand characteristics are recorded. The black stands in the enlarged panel are stands the DNR has reported are infested with dwarf mistletoe (*Arceuthobium pusillum*).

trained to recognize and record parasite incidence, this is typically not the focus of the inventory. The size of the area inventoried (often thousands of square kilometers) and the variation within that area can also limit the reliability of the inventory. This can result in insect and disease information from forest inventories that is often unreliable. Specific pest surveys, such as the intensive survey conducted by Baker et al. (F. A. Baker, M. Hansen, J. D. Shaw, M. Mielke, and D. N. Shelstad, *unpublished manuscript*), are expensive and time consuming, but provide accurate information on the extent of infestation. A comparison of the DNR forest inventory and the intensive survey of Baker et al. shows agreement in only 53% of the 196 stands where we have

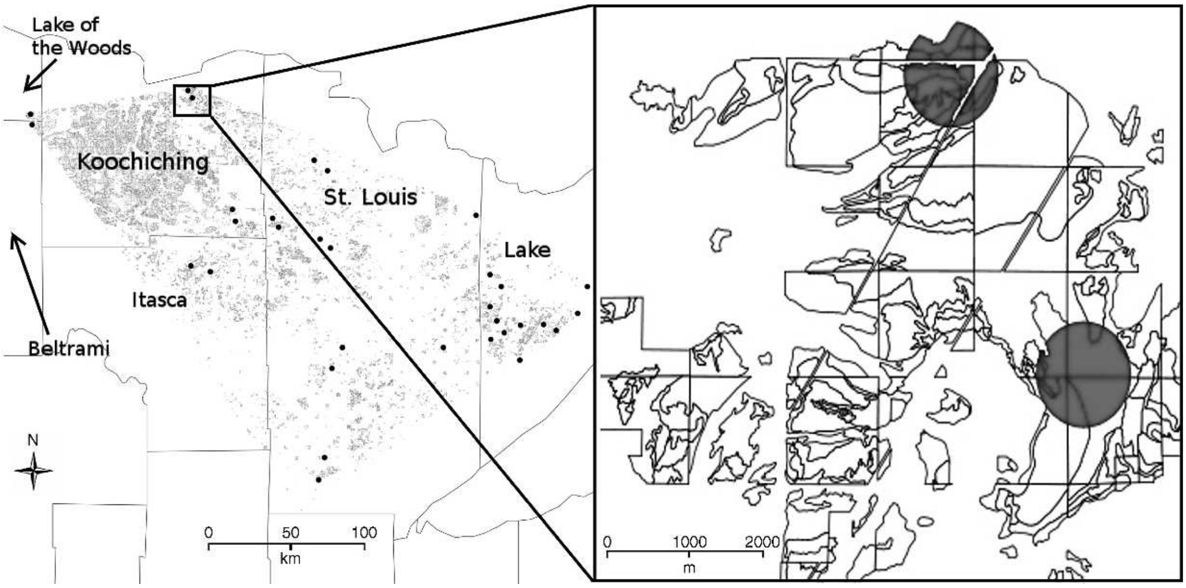

FIG. 3. The Minnesota Department of Natural Resources (DNR) and Federal studies reported dwarf mistletoe (*Arceuthobium pusillum*) present in 5–11% of black spruce (*Picea mariana*) stands in northern Minnesota. Baker et al. (F. A. Baker, M. Hansen, J. D. Shaw, M. Mielke, and D. N. Shelstad, *unpublished manuscript*) conducted a confirmatory study of 196 DNR inventory stands within 0.8 km of 31 different USDA Forest Service Forest Inventory and Analysis plot locations. This intensive survey found dwarf mistletoe in 56% of these 196 stands. We developed a model that uses both the intensive and DNR surveys to predict mistletoe presence or absence at stands in the DNR inventory not surveyed by Baker et al. The stands within the geographic convex hull of this intensive survey are shown on the left.

both observations, little more than would be expected by random chance.

The DNR inventory, while less accurate at detecting dwarf mistletoe, still contains a wealth of valuable information. At each of the DNR stands located across northeastern Minnesota, more than 40 stand attributes were measured, including: presence of different tree species in the stand, cover type, stand density, undergrowth density, age of the stand, basal area, and geographic location (latitude and longitude) of the stand. Many of these stand characteristics could be related to the presence or absence of dwarf mistletoe and could aid in predicting the occurrence of dwarf mistletoe in a particular stand.

The stands in the intensive survey are not a random sample of the stands in the DNR inventory; rather, the survey locations were selected to be near FIA inventory sites. This results in 17% of stands in the intensive survey being labeled by the DNR inventory as having mistletoe present, as opposed to 11% of stands in the entire DNR inventory. Care must be taken in interpreting the results of such nonrandom samples (e.g., Keating and Cherry 2004). Appendix A contains details of a simulation study conducted to examine the effect that such a discrepancy might have on our results. For the level of discrepancy found in this study, simulations show that inferences based on our sample do not appear to be affected.

In what follows, we link the small intensive survey to the large DNR forest inventory through Bayesian data reconciliation. We seek to understand and model the differences between these two data sets, as well as the ecological process driving dwarf mistletoe presence in black spruce and use this understanding to make improved predictions on the extent of the dwarf mistletoe infestation across northern Minnesota.

Data model

To avoid extrapolation, all stands in the DNR inventory that were outside of the geographic convex hull of the 196 stands for which we have both the DNR and intensive surveys were removed from the analysis.

The intensive survey and the DNR data agree quite often (91% of the time) when the intensive survey did not find dwarf mistletoe in a stand. On the other hand, when the intensive survey found dwarf mistletoe, the DNR inventory often did not, agreeing only 23% of the time. From this, it is clear that the DNR inventory contains both "false positive" and "false negative" errors when compared to the accurate intensive survey and that the false negative rate is likely to be higher than the false positive rate. Royle and Link (2006) suggest a generalized site occupancy model that allows for both false positive and false negative errors in the sampling process. Following Royle and Link (2006), if y_i is the presence ($y_i = 1$) or absence ($y_i = 0$) of mistletoe in a stand, as identified in the intensive survey, and w_i is the DNR inventory presence or absence in the same stand, the data model can be written as

$$w_i \mid \phi_i, \psi_i, y_i \sim \begin{cases} \text{Bern}(\phi_i) & \text{if } y_i = 1 \\ \text{Bern}(\psi_i) & \text{if } y_i = 0 \end{cases} \quad (1)$$

where ϕ_i is the probability of the DNR survey finding mistletoe if it is present in the intensive survey ($1 - \phi_i$ is the probability of a false negative) and ψ_i is the probability of the DNR survey reporting mistletoe present if it is not present in the intensive survey (ψ_i is the probability of a false positive).

If we compare this data model to the diagram in Fig. 1c, D_L corresponds to the DNR inventory (w), and D_A corresponds to the intensive survey (y). Thus, in this data model, we have modeled the less accurate DNR inventory as a noisy version of the intensive survey.

Process model

Having specified a data model, we now model the relationship between mistletoe presence or absence, as reported by the intensive survey, and the stand characteristics reported in the DNR inventory.

A generalized linear model (GLM) with a binary response (e.g., logistic or probit regression) is a natural choice for a statistical model of forest pest occupancy. Stand characteristics from the DNR inventory are used as covariates in a GLM with the response variable being the presence or absence of mistletoe as found in the intensive survey. Following Albert and Chib (1993), we employ the probit link function in our GLM to allow the use of a more robust MCMC algorithm; the probit link, denoted as Φ^{-1}, is the inverse cumulative distribution function of the standard normal distribution. Thus, consider the following process model specification:

$$y_i \mid \boldsymbol{\beta} \sim \text{Bern}(\theta_i), \Phi^{-1}(\theta_i) = \mathbf{x}_i' \boldsymbol{\beta} \quad (2)$$

where again y_i is the presence or absence of mistletoe at the ith stand as measured by the intensive survey, and θ_i is the latent probability that stand i is infected. This latent probability of presence, θ_i, depends on the DNR stand characteristics in $\mathbf{x}_i$ through the corresponding regression coefficients $\boldsymbol{\beta}$.

The parameters $\boldsymbol{\phi}$ and $\boldsymbol{\psi}$ could also vary with stand characteristics recorded in the DNR inventory. The processes resulting in false positive and false negative surveys are different concerns and could be driven by different environmental factors. We thus model these processes individually:

$$\Phi^{-1}(\phi_i) = \mathbf{x}_i' \boldsymbol{\beta}_\phi \quad (3)$$

$$\Phi^{-1}(\psi_i) = \mathbf{x}_i' \boldsymbol{\beta}_\psi. \quad (4)$$

Accounting for spatial autocorrelation

Dwarf mistletoe spreads by shooting seeds an average of 1–2 m (e.g., Hawksworth and Wiens 1972, Baker and French 1986). The short-range nature of this dispersal mechanism suggests that the presence of mistletoe should be spatially autocorrelated, and previous studies have indicated that pines infected with dwarf mistletoe are spatially aggregated (Reich et al. 1991). Thus we seek to account for possible spatial autocorrelation explicitly in our model.

The data used in this study arise from black spruce stands of varying sizes and shapes, some isolated from one another and others contiguous (Fig. 2). We have no information about where in the stand the mistletoe was found, only that mistletoe is present (or absent) there. Data of this type are called "areal data" in spatial statistics (e.g., Schabenberger and Gotway 2004). In contrast to geostatistical studies in which spatial analysis relies on geographic distance between points of interest, areal spatial analysis relies on a neighborhood structure that specifies the spatial relationship of locations to one another. There is no standard method for choosing a neighborhood structure in an arbitrary setting, and thus for this study we tested for autocorrelation using three different neighborhood structures and a variety of ranges of spatial structure. We chose between these various neighborhood structures by comparing the resulting goodness of fit of models with different neighborhood structures. The first neighborhood structure defines a neighborhood as all stands within a kilometer of the stand in question. The second defines a neighborhood as the four stands nearest to each stand. The third neighborhood structure was constructed by first creating triangles with each stand's centroid as a vertex, then defining each stand's neighborhood as all other stands that shared a side of a triangle with the stand in question. All neighborhood structures were created by using the "spdep" package (Bivand et al. 2009) in the R statistical computing environment (R Development Core Team 2009). A traditional probit regression was conducted using the presence of mistletoe as reported by the intensive survey as the response variable and all possible stand characteristics as covariates. The residuals of this analysis showed significant spatial autocorrelation ($P < 0.01$ for all neighborhood structures) when tested using Moran's I test statistic (e.g., Schabenberger and Gotway 2004). Likewise, the residuals of a probit regression on $\boldsymbol{\psi}$, the probability of a "false positive," showed significant residual spatial structure (Moran's I, $P < 0.02$), though the residuals of a probit regression on $\boldsymbol{\phi}$, the probability of a "true positive," showed no latent spatial autocorrelation (Moran's I, $P > 0.7$).

Accommodating spatial autocorrelation is necessary to ensure that further modeling assumptions are met and can improve the predictive power of the model (Hoeting et al. 2000). Dormann et al. (2007) reviewed common methods for accounting for spatial structure in areal data, including, among others, spatial eigenvector mapping, conditionally autoregressive (CAR) models, and autocovariate regression. The large size of the DNR data set makes many of these methods computationally infeasible. Spatial eigenvector mapping, for example, requires finding the eigenvectors of a matrix whose entries are the pairwise distances between all sites in the data set. For the DNR data, this would be a 25 235 by 25 235 matrix, requiring more memory than available in standard computing environments. In contrast, autocovariate regression can easily be applied to large data sets such as the DNR inventory. An extra covariate (predictor variable) is created for each stand that represents the proportion of "neighboring" stands infested with dwarf mistletoe. This covariate is then added to the data set for regression analysis.

Multiple autocovariates were created using the three different neighborhood structures described above, all based on a weighted average of stands within a certain radius of the stand in question. Neighborhood distances from 250 to 3000 m, in increments of 250 m, were used in conjunction with three different weighting schemes: all being weighted equally, stands weighted inverse proportionally to their distance from the stand in question, and stands weighted inverse proportionally to the square of their distance. Each autocovariate created in this way was tested in a traditional (non-Bayesian) probit regression with the other stand characteristics, and the Akaike information criterion (AIC; Akaike 1974) of the resulting models were compared. The autocovariates that resulted in a probit regression model with the best AIC in this a priori analysis were used in the full BHM.

Variable selection process

The Minnesota DNR inventory has more than 40 stand characteristics that could be used as predictor variables in our model. This large number of variables could make it more difficult to determine which stand characteristics are significantly related to the presence or absence of mistletoe in black spruce, especially if multicollinearity is present between some of the variables. We employed three traditional probit regression models with $\mathbf{y}$, $\boldsymbol{\phi}$, and $\boldsymbol{\psi}$ as response variables and all stand characteristics, as well as the spatial autocovariates for $\mathbf{y}$ and $\boldsymbol{\psi}$, as predictor variables. A stepwise model selection, based on AIC, was conducted on each full set of models, and the resulting stand characteristics were used in the process model of the BHM: Eqs. 2, 3, and 4. Supplement 1 includes R code to reproduce the variable selection process, and a full list and explanation of DNR stand characteristics is obtainable from the Minnesota DNR (*available online*).[5]

Parameter model

Bayesian statistical techniques require specification of prior distributions on all parameters of interest in the data and process models. In our model, we need to specify priors for the regression coefficients $\boldsymbol{\beta}$, $\boldsymbol{\beta}_{\phi}$, and

[5] ⟨http://www.dnr.state.mn.us/maps/forestview/csa_defs.html⟩

$\boldsymbol{\beta}_\psi$, from the process model. In the absence of specific a priori scientific knowledge of these parameters, we used vague priors. Specifically, each regression coefficient was given a normally distributed prior distribution with mean zero and standard deviation 10^6. Sensitivity to the choice of prior distributions was assessed by varying the means of the prior distributions in four separate MCMC runs.

Posterior distribution

Having specified data, process, and parameter models, we can now consider the joint posterior distribution:

$$[\boldsymbol{\beta}, \boldsymbol{\phi}, \boldsymbol{\psi} \mid \mathbf{y}, \mathbf{w}] \propto \prod_{i=1}^{N} [w_i \mid y_i, \phi_i, \psi_i] \times \prod_{i=1}^{N} [y_i \mid \boldsymbol{\beta}] \prod_{i_1=1}^{N_1} [\phi_{i_1} \mid \boldsymbol{\beta}_\phi] \prod_{i_0=1}^{N_0} [\psi_{i_0} \mid \boldsymbol{\beta}_\psi][\boldsymbol{\beta}][\boldsymbol{\beta}_\phi][\boldsymbol{\beta}_\psi] \tag{5}$$

where N is the number of sites for which we have observations from both the DNR inventory and the intensive survey, N_1 is the number of sites where mistletoe was present in the intensive survey, and N_0 is the number of sites where mistletoe was absent in the intensive survey.

The full-conditional distributions for the parameters of this model were found analytically, allowing us to use the computationally efficient Gibbs sampler (e.g., Gelman et al. 2004). The derivation of these full conditionals is presented in Appendix B, and code used to obtain the results that follow is presented in Supplement 1.

Prediction

In the BDR framework, prediction is accomplished through finding the posterior predictive distribution of D_A, given D_L and all other parameters in the model. In this situation, we seek the probability of mistletoe presence in a stand surveyed by the DNR, but not by the intensive survey:

$$\text{P(mistletoe is present} \mid \text{DNR data)}.$$

Conditioning on all parameters, Bayes' theorem of conditional probability allows for the calculation of the desired predictive distribution. For a stand not examined in the intensive survey (denoted by the "u" subscript), we first take the case in which the DNR has found mistletoe ($w_u = 1$). The posterior predictive probability of mistletoe presence can be written as follows:

$$\begin{aligned} P(y_u = 1 \mid w_u = 1) &= [P(w_u = 1 \mid y_u = 1)P(y_u = 1)] \\ &\quad \div [P(w_u = 1 \mid y_u = 1)P(y_u = 1) \\ &\qquad + P(w_u = 1 \mid y_u = 0)P(y_u = 0)]. \end{aligned} \tag{6}$$

All of the probabilities on the right-hand side are terms we have estimated in the model. Specifically, $P(w_u = 1 \mid y_u = 1)$ is ϕ_u from the data model, $P(y_u = 1)$ is the probability, in the intensive survey, of mistletoe being present, θ_u, which can be predicted from the DNR-collected stand characteristics and the regression coefficients in the process model (Eq. 2). Likewise, $P(w_u = 1 \mid y_u = 0)$ is ψ_u and $P(y_u = 0)$ is $1 - \theta_u$. We can then write Eq. 6 as follows:

$$P(y_u = 1 \mid w_u = 1) = \frac{\phi_u\theta_u}{\phi_u\theta_u + \psi_u(1 - \theta_u)}. \tag{7}$$

In a similar fashion, the probability of true presence of dwarf mistletoe in a stand where the DNR did not observe it can be written as follows:

$$P(y_u = 1 \mid w_u = 0) = \frac{(1 - \phi_u)\theta_u}{(1 - \phi_u)\theta_u + (1 - \psi_u)(1 - \theta_u)}. \tag{8}$$

Eqs. 7 and 8 are written as probabilities, but together they are sufficient to specify a full-conditional posterior predictive distribution for the presence of mistletoe in a stand surveyed by the DNR but not by the intensive survey:

$$y_u \mid w_u, \psi_u, \phi_u, \theta_u \sim \text{Bern}(\eta_u)$$

$$\eta_u = \begin{cases} \dfrac{\phi_u\theta_u}{\phi_u\theta_u + \psi_u(1 - \theta_u)} & \text{if } w_u = 1 \\[2ex] \dfrac{(1 - \phi_u)\theta_u}{(1 - \phi_u)\theta_u + (1 - \psi_u)(1 - \theta_u)} & \text{if } w_u = 0. \end{cases} \tag{9}$$

In this equation, η_u is the posterior predictive probability that mistletoe is present in the stand not examined in the intensive survey.

Of the 46 415 stands in the Minnesota DNR inventory, 21 180 were located outside of the area covered by the intensive survey. To avoid extrapolation we did not make predictions for these stands. Predictions using the model developed in this study were made only for the 25 235 stands located within the convex hull of the stands in the intensive survey (see Fig. 3).

Stand characteristics from the DNR inventory are available for all stands where we seek to predict presence or absence of mistletoe, but the spatial autocovariates used in the model are based on a knowledge of the presence or absence of mistletoe in neighboring stands, something known only for stands in the intensive survey. To approximate these missing autocovariates at stands not in the intensive survey, both spatial autocovariates are approximated at each iteration in the Gibbs sampler using the predicted probability of presence of mistletoe at neighboring stands as a surrogate for the missing presence or absence (Augustin et al. 1996, Hoeting et al. 2000). Essentially, the autocovariates can be thought of as spatial random processes with their own distributions. Estimating the autocovariates in this way accounts for the uncertainty we have regarding the presence or absence of mistletoe at stands in the DNR inventory. Instead of using a point estimate, the autocovariates constructed in each iteration of the Gibbs sampler change as the predicted presence of mistletoe at neighboring stands changes. Embedding the estimation of the spatial autocovariates in the Gibbs sampler also allows us to estimate the autocovariates within the same procedure in which we will fit the model and make predictions. In contrast, traditional statistical techniques would require estimating the autocovariates separately from the model-fitting stage. These techniques would result in a point estimate of the autocovariates, and this point estimate would then be treated as the exact autocovariates and used for prediction.

Model implementation

All full-conditional distributions of parameters in the model were found analytically, and an MCMC algorithm was constructed within the R statistical computing environment to iteratively sample from the joint posterior distribution of the parameters by repeatedly sampling from each full-conditional distribution in turn. The necessary R code to accomplish this is available in Supplement 1. In order to assess convergence to a stationary posterior distribution, four separate runs of the iterative MCMC algorithm were conducted using different starting values that were over-dispersed relative to the posterior distribution of the parameters. We conducted 20 000 iterations of each run, and the first half of each chain was discarded as burn-in iterations. The between- and within-chain variances of the four resulting chains were computed for each parameter being estimated and used to calculate a common convergence statistic, the potential scale reduction factor, $\hat{R}$ (Gelman et al. 2004:297).

For comparison, we also fit two similar, but more parsimonious, BDR models. The first assumes homogeneous detection probabilities, in which $\boldsymbol{\phi}$ and $\boldsymbol{\psi}$ are assumed to be constant for all stands, as opposed to spatially varying. This will allow us to consider whether we are overfitting by making inference about two spatially varying parameters (ϕ_i and ψ_i) for each stand where we have observations.

In the second more parsimonious BDR model we removed the spatial autocovariates for $\mathbf{y}$ and $\boldsymbol{\psi}$. The residuals from this model fit, as well as the corresponding residuals from the full model, were examined for spatial dependence using Moran's I test statistic. This allows us to gauge the effectiveness of these autocovariates at accounting for possible spatial structure in the presence or absence of mistletoe.

Assessment through cross-validation

In order to assess the effectiveness of BDR, we employed leave-one-out cross-validation, in which each stand in the intensive survey was dropped, one by one, from the analysis and the remaining stands were used to fit the model and predict mistletoe presence or absence at the dropped stand. We conducted a single run of 10 000 iterations of the Gibbs sampler for each stand, using parameter estimates from a traditional probit regression as initial values for the parameters. The resulting posterior predictive means of mistletoe presence at the dropped stands were recorded and used to set a threshold value, T, which gives the most accurate predictions on the withheld stands (e.g., Hooten et al. 2003). Stands whose posterior predictive mean probability of mistletoe was greater than T were classified as infested, while those with a posterior predictive mean of less than T were classified as uninfested. For comparison, analogous cross-validation tests were conducted for two prediction methods: a traditional probit regression model and random forests. The predictions resulting from these three methods were compared with the presence or absence of mistletoe in the stand as measured by the intensive survey.

In binary classification, the accuracy of a classifier is defined as the proportion of units that are correctly classified (e.g., Taylor 1996):

$$\begin{aligned}\text{accuracy} = {} & [(\text{true positives}) + (\text{true negatives})]\\ & \div [(\text{true positives}) + (\text{false positives})\\ & + (\text{true negatives}) + (\text{false negatives})].\end{aligned}$$

The accuracy of each prediction method was computed and used to measure the ability of each method to predict the presence or absence of mistletoe as measured in the intensive survey, using only the information in the DNR inventory.

Results

All MCMC runs converged to similar posterior distributions, suggesting that the model is fairly robust to variation in the prior distributions and thus the data will be the dominant driver of statistical inference. The potential scale reduction factor ($\hat{R}$) was calculated for each parameter in our model (Tables 1 and 2). The quantity, $\hat{R}$, measures the factor by which the spread (or scale) of the estimated posterior distribution might be diminished by running an infinite number of samples in each MCMC chain. Values close to one indicate convergence, while values higher than one indicate that convergence to the true joint posterior distribution has not yet occurred. Gelman et al. (2004) suggest, as a rule of thumb, that $\hat{R}$ values below 1.1 indicate convergence. In this study, $\hat{R}$ values for all parameters were deemed to be close enough to one that we can be confident our iterative MCMC algorithm has converged to the true joint posterior distribution of the parameters. The four separate chains were combined for each parameter, and the resulting 40 000 samples were used for inference on

Table 1. Probit regression coefficients in the Bayesian data reconciliation model of eastern spruce dwarf mistletoe (*Arceuthobium pusillum*) incidence in black spruce (*Picea mariana*) stands in Minnesota, USA.

		95% credible interval		
Covariate	Median	Lower bound	Upper bound	$\hat{R}$†
Regression coefficients for y				
Intercept	−4.2628	−7.0404	−1.6239	1.0005
Cover type size class	0.3511	0.1028	0.6100	<1.0001
Stand "wetness" class	1.0409	0.5049	1.6102	1.0002
Stand density (1000 board-feet/acre)	0.3903	0.0985	0.7770	1.0003
Height of dominant species	−0.0215	−0.0364	−0.0070	<1.0001
Mortality of dominant species	0.6293	0.2258	1.0590	1.0001
Understory density	−0.2604	−0.4891	−0.0334	1.0001
Presence of tamarack	−0.4096	−0.9518	0.1241	<1.0001
Presence of northern white cedar	1.1044	0.2255	2.0577	1.0003
Presence of lowland black spruce	−0.9994	−1.9956	−0.0593	1.0002
Presence of balsam fir	−1.3029	−2.8215	0.0535	1.0002
Northern white cedar cover type	−2.2702	−3.8039	−0.8033	<1.0001
Stagnant spruce cover type	−1.5354	−2.5737	−0.5478	1.0001
Aspen cover type	1.2018	−0.0580	2.5019	1.0009
Jack pine cover type	2.3823	−0.0549	4.9522	1.0003
Spatial autocovariate (ac_y)	1.7365	1.1196	2.3783	1.0007
Regression coefficients for ϕ				
Intercept	−2.0911	−2.8211	−1.4656	1.0002
Mortality of dominant species	0.9544	0.5971	1.3323	1.0001
Presence of tamarack	0.5449	−0.1605	1.2392	1.0001
Lowland black spruce cover type	0.9909	0.3080	1.7219	1.0003
Regression coefficients for ψ				
Intercept	−5.7777	−11.6221	−1.2673	1.0226
Cover type size class	−0.7722	−1.6764	0.1065	1.0037
Understory size class	2.9167	1.1279	5.3506	1.0221
Mortality of dominant species	3.4327	1.1825	6.4058	1.0215
Understory density	−1.6632	−3.1416	−0.5392	1.0185
Presence of northern white cedar	8.6992	3.6880	15.6561	1.0262
Spatial autocovariate (ac_ψ)	13.2552	5.8397	23.8439	1.0285
Tests for spatial autocorrelation				
Moran's *I P* value				
Without ac_y and ac_ψ	0.0980	0.0004	0.8774	<1.0001
With ac_y and ac_ψ	0.4730	0.0295	0.9690	<1.0001

Notes: The variable y is the presence ($y_i = 1$) or absence ($y_i = 0$) of mistletoe in a stand, as identified in the intensive survey; ϕ is the probability of the DNR survey finding mistletoe if it is present in the intensive survey; and ψ is the probability of the DNR survey reporting mistletoe present if it is not present in the intensive survey. "DNR" is the Minnesota Department of Natural Resources.

† $\hat{R}$ is the potential scale reduction factor (Gelman et al. 2004). Values of $\hat{R} < 1.1$ indicate convergence of the iterative Markov chain Monte Carlo (MCMC) procedure used to fit the model.

model parameters and for prediction of mistletoe presence or absence.

Inferences on the regression coefficients describing the effect of various stand characteristics in the process model: Eqs. 2, 3, and 4, are shown in Table 1. The stand characteristics in Table 1 are those chosen through the stepwise model selection process for each of the three probit regressions in the process model. The credible

Table 2. Additional model parameters.

	P value for spatial autocorrelation test		Homogeneous detection probabilities	
Statistic	Without ac_y and ac_ψ	With ac_y and ac_ψ	DNR true detection, ϕ	DNR false detection, ψ
Median	0.0980	0.4730	0.0413	0.0530
Lower 95% CI	0.0004	0.0295	0.0182	0.0225
Upper 95% CI	0.8774	0.9690	0.0847	0.0809
$\hat{R}$	<1.0001	<1.0001	<1.0001	<1.0001

Notes: "DNR" is the Minnesota Department of Natural Resources; "CI" is the credible interval. Moran's *I* was used to test for spatial autocorrelation. $\hat{R}$ is the potential scale reduction factor (Gelman et al. 2004). Values of $\hat{R} < 1.1$ indicate convergence of the iterative Markov chain Monte Carlo (MCMC) procedure used to fit the model.

TABLE 3. Contingency table of dwarf mistletoe (*Arceuthobium pusillum*) presence in the intensive survey of black spruce (*Picea mariana*) stands with the Minnesota Department of Natural Resources (DNR) inventory and cross-validation predictions from various models.

Statistical approach	Presence/ absence	Intensive survey: Present	Intensive survey: Absent	Threshold, T	Accuracy (%)
DNR inventory	present	25	8	NA	52.55
	absent	85	78		
Random forests	present	82	32	NA	69.39
	absent	28	54		
Bayesian data reconciliation	present	94	34	0.45	73.98
Homogeneous ϕ and ψ	absent	16	52		
Traditional probit regression	present	93	31	0.43	75.51
	absent	17	55		
Bayesian data reconciliation	present	95	32	0.40	76.02
Spatially varying ϕ and ψ	absent	15	54		

Note: The abbreviation "NA" means "not applicable."

intervals of four parameters in the model for mistletoe presence ($\mathbf{y}$) contain zero: presence of tamarack, presence of balsam fir, aspen cover type, and jack pine cover type. Thus the effect of these stand characteristics on the presence or absence of mistletoe is not statistically different from zero. Similarly, neither the presence of tamarack in the model for $\boldsymbol{\phi}$ nor the cover type size class in the model for $\boldsymbol{\psi}$ are statistically different from zero.

The spatial autocovariate related to the presence of mistletoe in neighboring stands is positively correlated with mistletoe presence. When both autocovariates are removed from the analysis, Moran's I test statistic is marginally significant; however, when the spatial autocovariates are included in the model, Moran's I test statistic is no longer significant (Table 2).

The results of the leave-one-out cross-validation for the various predictive methods are shown in Table 3. The full BDR model with spatially varying $\boldsymbol{\phi}$ and $\boldsymbol{\psi}$ performs best at using the DNR inventory to predict mistletoe presence as observed in the intensive survey.

Of the 25 235 black spruce stands inventoried by the Minnesota DNR within the geographical range of the intensive survey, 11% are classified by the DNR as infested. The BHM presented here estimates that 59% of the same stands have a posterior predictive mean probability of dwarf mistletoe presence greater than the threshold value of $T = 0.4$ (Table 4). These stands are more likely to be infested than not. Under the Bayesian framework we can also predict the standard deviation of the posterior predictive distribution of the probability of infestation in each stand, which can then be used to infer the probability of mistletoe being present at a stand. We define a stand as being "highly likely" to have mistletoe present if there is at least a 95% chance that the posterior predictive probability of mistletoe presence (η) is greater than the threshold value, T. Likewise, a stand is "highly unlikely" to have mistletoe present if there is at least a 95% chance that the posterior predictive probability of mistletoe presence (η) is lower than the threshold value. Of the 25 235 stands in our study, we predict that 8883 (35%) are highly likely to

TABLE 4. Comparison of Minnesota Department of Natural Resources (DNR) status with posterior predictive probability η of dwarf mistletoe (*Arceuthobium pusillum*) presence by county.

			Bayesian probability η exceeds T (% of stands)		
County	No. stands in county	DNR infected status (%)	Likely, $\Pr > 0.5$	Highly unlikely, $\Pr < 0.05$	Highly likely, $\Pr > 0.95$
Koochiching	14 159	10.89	59.61	22.44	37.08
Lake of the Woods	66	18.18	21.21	48.48	3.03
Beltrami	46	73.91	73.91	8.70	13.04
St. Louis	6843	8.18	57.11	20.58	30.22
Lake	1906	26.18	74.82	10.86	49.32
Itasca	2215	8.67	53.36	24.38	27.86
Total (all counties)	25 235	11.25	59.46	21.27	35.20

Note: Stands are classified as infested if η exceeds the threshold level T. In the Bayesian model, predictions come in the form of a probability distribution of η, allowing for gradation in the certainty of the predictions.

be infested and 5367 (21%) are highly unlikely to be infested (Table 4).

In the case in which ϕ and ψ are assumed to be homogeneous, the predicted rate of false negative classification (1 − ϕ) in the DNR inventory, relative to the intensive survey, is 0.81 (Table 5), while the predicted rate of a false positive classification (ψ) is 0.21.

Discussion

Dwarf mistletoe

The results of the leave-one-out cross-validation show that our methods have increased the accuracy of predictions made from the DNR inventory, relative to the intensive survey. Before the data reconciliation process, the DNR data only agreed with the more accurate intensive survey 53% of the time (Table 3), just slightly better than would be expected by random chance. Predictions from the spatially varying BDR process agree with the intensive survey 76% of the time, a higher level of accuracy than any of the other methods attempted here. We have successfully updated the extensive but inaccurate DNR forest inventory using the small, reliable intensive survey, and our predictions reflect what might be found were the intensive survey extended to cover the whole range of the DNR inventory.

Our analysis also suggests a disparity in the reliability of the Minnesota DNR survey. When compared to the more accurate intensive survey, the DNR survey is highly accurate at correctly assessing uninfested stands, as seen by the low proportion of false positives (see **ψ** in Table 5), but much less accurate at correctly assessing stands infested with dwarf mistletoe. The high false negative rate indicates that, on average, the probability of an infested stand being correctly classified is only 19%. Taken together, our results for **ϕ** and **ψ** suggest that in comparison to the intensive survey the Minnesota DNR inventory significantly underestimates the extent of dwarf mistletoe in black spruce stands.

We predict that mistletoe is present in roughly 3–5 times as many stands as reported in the DNR inventory (Table 4). This significant increase in the level of dwarf mistletoe infestation is apparent across all counties in the study except for Beltrami and Lake of the Woods. The predictions for each stand in the DNR survey and maps of these stands indicate locations where our predictions are similar to or different from the DNR survey presence or absence of mistletoe (Fig. 4). This knowledge can be used to make forest management decisions such as which stands to survey next for dwarf mistletoe or which areas are currently most threatened by mistletoe infestation. New surveys could be conducted to verify or refute the predictions we make in this study.

We have allowed the detection probabilities **ϕ** and **ψ** to vary across stands, illuminating stand characteristics that might affect the ease or difficulty of correctly surveying spruce stands for mistletoe presence (Table 1). County-level aggregation of results (Table 5) predicts that the false negative rate (1 − **ϕ**) is fairly constant across all counties in the DNR survey. In contrast, the false positive rate (**ψ**) varies more between counties. For example, the false positive rate in Koochiching county is 0.24, much higher than the rate in St. Louis county, 0.13. This may reflect the difficulty in surveying the extensive forests found in Koochiching county, as compared to the relatively sparse and much more accessible forest stands in St. Louis County (Fig. 3).

Table 5. Summary of results for spatially varying detection probabilities ϕ and ψ for dwarf mistletoe (*Arceuthobium pusillum*) in black spruce (*Picea mariana*) stands in northern Minnesota, USA.

County	Average false negative rate (1 − ϕ)	Average false positive rate (ψ)	Stands with $\psi > 0.75$ (%)
Koochiching	0.80	0.24	19.26
Lake of the Woods	0.88	0.17	14.81
Beltrami	0.78	0.40	23.55
St. Louis	0.83	0.13	8.84
Lake	0.77	0.26	19.14
Itasca	0.82	0.19	14.41
Total	0.81	0.21	16.00

Note: For the numbers of stands in each county, see Table 4.

More extensive results, including a full list of stand characteristics and results for all stands in the Minnesota DNR survey are provided in Supplement 2. Large-scale images showing the incidence of mistletoe as reported by the DNR inventory, predicted probability of mistletoe presence, standard deviation of the posterior predictive distribution, and predictions for the detection probabilities **ϕ** and **ψ** are provided in Appendix C.

The BDR approach allows us to make predictions of mistletoe incidence simultaneously with inference about the epidemiological process driving mistletoe in northern Minnesota. The estimated regression coefficients of stand characteristics from the model for **y** (Table 1) give some insight in to what drives dwarf mistletoe infestation. For example, the positive coefficient related to stand mortality and the negative coefficient related to the height of the dominant species reflect the effects of dwarf mistletoe on infested stands.

By modeling the less accurate DNR survey as a noisy version of the more accurate intensive survey, **ϕ** and **ψ** contain information about any effect that is related to discrepancies in the two data sets. Aside from differences in the ability of the two surveys to correctly detect mistletoe, **ϕ** and **ψ** may also model any temporal difference in the mistletoe infestation that has occurred between the observation of both surveys, differences in the spatial domain surveyed in each stand, and other possible factors. Thus, care must be taken in interpreting the results for detection probabilities **ϕ** and **ψ**, as well as their corresponding regression coefficients. For example, the positive coefficient for mortality in the model for **ψ**

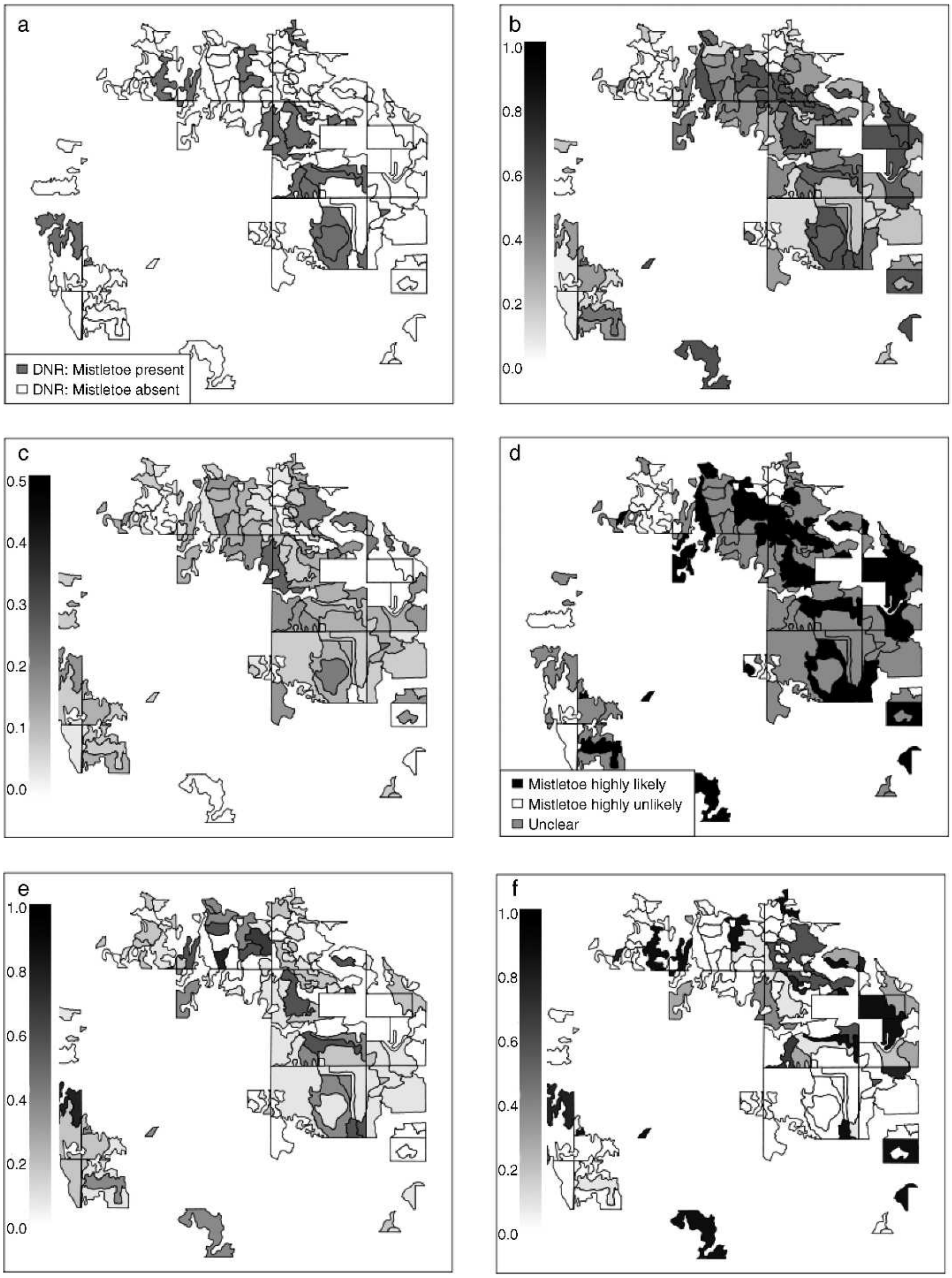

FIG. 4. Plots of Minnesota Department of Natural Resources (DNR) survey and model predictions at a selection of stands in Koochiching County. (a) Presence or absence of dwarf mistletoe (*Arceuthobium pusillum*) as reported in the DNR survey. The DNR survey reports 11% of black spruce (*Picea mariana*) stands are infected with dwarf mistletoe. (b) Posterior predicted mean probability of mistletoe presence in each stand. Darker shades of gray correspond to higher probability of mistletoe in a stand. (c) Standard deviation of the posterior predicted probability of mistletoe presence at each stand. Darker shades of gray correspond to higher variance in the posterior predictive distribution. (d) The mean and standard deviation of the posterior predictive distribution allow us to identify whether stands are "highly likely" (black) or "highly unlikely" (white) to have mistletoe present. Prediction is unclear at stands depicted in gray. Our model predicts that 35% of stands in the DNR survey are "highly likely" to be infected with dwarf mistletoe. (e) Probability of a "false negative" ($1 - \phi_i$) in the DNR inventory, relative to the intensive survey. Darker shades of gray correspond to higher probability of a false negative. The probability of a false negative is high throughout most of the DNR inventory, indicating that the DNR data underrepresent the extent of the mistletoe infestation in northern Minnesota. (f) Probability of a "false positive" (ψ_i) in the DNR inventory, relative to the intensive survey. Darker shades of gray correspond to higher probability of a false positive. This indicates regions where the DNR inventory inflates the extent of the mistletoe infestation.

PLATE 1. Dwarf mistletoe (*Arceuthobium pusillum*) causes the most serious disease of black spruce (*Picea mariana*). This black spruce stand exhibits the thinning and "witches brooms" that are characteristic of a dwarf mistletoe infestation. Photo credit: F. A. Baker.

indicates that false-positive errors in the DNR survey relative to the intensive survey are more likely to occur in stands with high mortality than in stands with low mortality. This result might indicate that survey crews sometimes assume that dwarf mistletoe is the cause of observed mortality in a stand when the mortality is actually caused by some other agent, but, because we have only modeled the differences between the two data sets, making such a conclusion definitively would require a separate study.

The spatial autocovariate ac_y is positively correlated with mistletoe presence, confirming that mistletoe is more likely in stands near other infected stands. Likewise, ac_ψ is positively correlated with $\boldsymbol{\psi}$, indicating spatial structure in the rate of false-positive errors in the DNR data, relative to the intensive survey. The residuals of the full BDR model fit were tested for spatial autocorrelation using Moran's *I*. When the spatial autocovariates are omitted from the model, the median of the Moran's *I P* value (0.0980) indicates significant spatial autocorrelation in the residuals. With the spatial autocovariates in the model, the median (0.4730) and 95% credible interval (0.0295–0.9690) of the posterior predictive distribution of the Moran's *I P* value are close to what would be expected for random, uncorrelated data (mean of 0.5 and 95% credible interval of 0.025–0.975), indicating that the spatial autocovariate successfully accounts for the spatial autocorrelation present in the data and satisfies the model assumption of uncorrelated residuals.

Bayesian data reconciliation

Bayesian data reconciliation provides a flexible and robust framework for integrating multiple sources of data, as illustrated by our study of dwarf mistletoe in Minnesota black spruce stands. The hierarchical Bayesian nature of the modeling framework allows us to implement meaningful models for both the ecological process and for the relationship of the data sources. We chose data and process models that would allow us to make inferences about the ecological process being studied, as well as the relationships between the DNR and intensive surveys, and the Bayesian framework allows us to formally couple these two statistical models, something not easily done using traditional methods.

The hierarchical nature of the data reconciliation process allows for simultaneous inference about ecological process and relationship between the two sources of data. This provides better integration of the two data sets than would be accomplished with a two-step process. For example, based solely on a comparison of the two sources of data (see Table 3), we might empirically assign ψ, the probability of a false positive in the DNR data relative to the intensive survey, to be 0.09. However, when ψ is assumed to be spatially homogeneous, its 95% credible interval is bounded by

0.02 and 0.08. In a similar fashion, the empirical estimate of ϕ, the probability of both the intensive and DNR surveys finding mistletoe in a stand, is higher than the 95% credible interval found in our analysis (see Tables 1 and 5). The simultaneous inference of the parameters in the data and process models through composition sampling allows the ecological process to influence the inference made in the data model. Thus, the hierarchical modeling framework allows us to see that in this case, the DNR inventory is more accurate than would be assumed from a separate empirical comparison of the DNR inventory to the intensive survey.

In analyses involving prediction, there is often an inherent trade-off between predictive power and interpretability of the results. In the application given here, we removed many covariates from the analysis in the variable selection process with the aim of removing collinearity and producing a parsimonious model. Multicollinearity among these covariates could cloud inference on their effects, as well as slow the convergence of the MCMC algorithm, and thus needs to be accounted for in some way. Alternately, all stand characteristics could be used to create a corresponding set of orthogonal covariates (e.g., through principal component analysis), which could then be used, without variable selection, in the model. This would retain the information from all covariates and could improve the predictive ability of the resulting model. Unfortunately, these orthogonal covariates are typically difficult, if not impossible, to interpret in an ecologically meaningful context. This is perfectly acceptable if prediction is the sole aim of the study, but if the researcher is also interested in illuminating ecological processes involved in the natural system, leaving stand characteristics untransformed may be preferred.

Implicit in the data reconciliation approach we have outlined here is the assumption that one data set is more accurate than the other. Outside of controlled experiments, there is typically no way to judge the absolute accuracy of a set of observations of an ecological process, and thus a priori knowledge must guide decisions on the relative accuracy of the data sets being reconciled. In our application, the focus of the small intensive survey was solely on dwarf mistletoe presence, while a wide range of characteristics were recorded for each stand in the DNR forest inventory. This discrepancy in focus makes it likely that the intensive survey more reliably reports dwarf mistletoe presence than does the DNR inventory. We are thus confident that, as our data reconciliation process has produced predictions that are more closely in line with the more accurate intensive survey than were the original DNR data, these predictions are themselves more likely to accurately represent the true extent of the dwarf mistletoe infestation in northern Minnesota.

If no assumption about the relative accuracy of multiple surveys can be made, the disease surveillance literature contains existing methods for estimating the sensitivity and specificity of diagnostic tests in the absence of a "gold standard" test (e.g., Hanson et al. 2003, Engel et al. 2006). Similarly, the occupancy literature contains methods for utilizing repeated observations considered to be of equivalent accuracy to make inference about the true underlying state of nature, as well as the accuracy of the surveys (e.g., Royle and Nichols 2003, Royle 2004, Royle and Link 2006). These methods assume a latent, unobservable, true state. The BHM framework allows for inference on this latent state, based on the observed data.

In this study, we focused on the extent of the infestation of a forest disease. The approach we present has direct application to, and links to existing methods in, disease surveillance in animals (Salman 2003) and humans (Lee et al. 2010). For example, it may be quite difficult or expensive to perform a highly accurate test for a disease on a human or animal (e.g., collecting and analyzing a tissue sample), but much less difficult to obtain a less accurate test for the same disease (e.g., visual examination or interview). Thus, a study in which many animals are tested using the more easily obtained disease test could be augmented by also applying the more accurate test to a small sample of the animals. In this way, more accurate predictions on disease rates can be obtained in a cost-efficient way using the BDR approach presented here. This situation is quite similar to that of "double sampling" (e.g., Tenenbein 1970), though the hierarchical modeling approach we advocate here allows for inference to be made jointly about the underlying process driving the disease, something not typically obtained in double sampling.

The BHM framework is highly flexible and can accommodate a wide variety of designs. For example, three or more sets of related data could be integrated in a similar fashion to the method we have presented here. Likewise, hierarchical models with more levels than the three traditionally included (data, process, and parameter models) could be used to model the relationships between multiple sources of data and highly complex systems. All manner of conditional relationships between data and latent parameters can be written within a hierarchical modeling framework (Cressie et. al. 2009).

The BDR approach could also be used to take advantage of existing data sets that may be outdated. A representative selection of the existing sites could be sampled using more modern and accurate survey methods, even optimally by minimizing a design criterion of choice (e.g., Hooten and Wikle 2009). This could facilitate updating outdated data sets in a cost-efficient manner, as long as a meaningful form of model dependency is used to account for the lag in time between sampling periods.

With the widespread availability of remotely sensed data and advances in geographic information systems (GIS), utilizing multiple sources of data is becoming common in ecological studies. Spatial data are often available at a variety of scales, and the change of scale required to reconcile such data often results in the so-called "modifiable areal unit problem" (e.g., Gotway and Young 2002). While our study focused on two surveys that were conducted on the same spruce stands, and thus on the same spatial scale, a similar approach to what we present here might be employed to reconcile spatial data at differing scales. In this way, BDR could be used to model the differences between the finer and coarser resolution data and predict the finer resolution information at locations where only the coarser resolution data is available.

Using data from multiple sources within a BHM is not a new idea, though the purpose of past studies has typically been to assimilate multiple types of data to make inferences about a latent ecological process (e.g., Clark et al. 2007). In the BDR approach presented here, we model the relationship between two data sets, one more accurate than the other, by placing data at different levels of a hierarchical model. Placing the more accurate data in the process level of a hierarchical model, instead of a latent ecological process as is traditional in BHMs, allows us to model the differences in the data jointly with the ecological process of interest and update predictions across the support of the less accurate data. The BHM framework allows us to accomplish these goals in a way that is statistically rigorous and results in scientifically meaningful information about the observational and ecological processes.

Acknowledgments

Funding for this study was provided by the Utah State University College of Science and the USDA Forest Service Forest Health Monitoring. We also thank three anonymous reviewers for their extremely helpful suggestions. Any use of trade names is for descriptive purposes only and does not imply endorsement by the United States Government.

Literature Cited

Akaike, H. 1974. A new look at the statistical model identification. IEEE Transactions on Automatic Control 19:716–723.

Albert, J. H., and S. Chib. 1993. Bayesian analysis of binary and polychotomous response data. Journal of the American Statistical Association 88:670–679.

Anderson, R. L. 1949. The spruce dwarf mistletoe in Minnesota. Thesis. University of Minnesota, St. Paul, Minnesota, USA.

Augustin, N., M. Mugglestone, and S. Buckland. 1996. An autologistic model for the spatial distribution of wildlife. Journal of Applied Ecology 33:339–347.

Baker, F. A., and D. W. French. 1986. Dispersal of seeds by *Arceuthobium pusillum*. Canadian Journal of Forest Research 16:15.

Baker, F. A., J. G. O'Brien, R. Mathiasen, and M. E. Ostry. 2006. Eastern spruce dwarf mistletoe. USDA Forest Insect and Disease Leaflet NA-PR-04-06. USDA Forest Service, Northeastern Area, State and Private Forestry, Newtown Square, Pennsylvania, USA.

Banerjee, S., B. P. Carlin, and A. E. Gelfand. 2003. Hierarchical modeling and analysis for spatial data. Chapman and Hall/CRC, Boca Raton, Florida, USA.

Berliner, L. M. 1996. Hierarchical Bayesian time series models. Pages 15–22 *in* K. Hanson and R. Silver, editors. Maximum entropy and Bayesian methods. Kluwer Academic, Dordrecht, The Netherlands.

Bivand, R. L., et al. 2009. spdep: Spatial dependence: weighting schemes, statistics and models. R package version 0.4-34. R Foundation for Statistical Computing, Vienna, Austria.

Breiman, L. 2001. Random forests. Machine Learning 45:5–32.

Breiman, L., J. Friedman, C. J. Stone, and R. A. Olshen. 1984. Classification and regression trees. Chapman and Hall/CRC, Boca Raton, Florida, USA.

Clark, J. S., M. Wolosin, M. Dietze, I. Ibanez, S. LaDeau, M. Welsh, and B. Kloeppel. 2007. Tree growth inference and prediction from diameter censuses and ring widths. Ecological Applications 17:1942–1953.

Cressie, N., C. A. Calder, J. S. Clark, J. M. Ver Hoef, and C. K. Wikle. 2009. Accounting for uncertainty in ecological analysis: the strengths and limitations of hierarchical statistical modeling. Ecological Applications 19:553–570.

Dormann, C. F., et al. 2007. Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. Ecography 30:609–628.

Engel, B., B. Swildens, A. Stegeman, W. Burst, and M. de Jong. 2006. Estimation of sensitivity and specificity of three conditionally dependent diagnostic tests in the absence of a gold standard. Journal of Agricultural, Biological, and Environmental Statistics 11:360–380.

Geils, B. W., and F. G. Hawksworth. 2002. Damage, effects, and importance of dwarf mistletoes. Pages 57–65 *in* B. W. Geils, J. C. Tovar, and B. Moody, technical coordinators. Mistletoes of North American conifers. General Technical Report RMRS-GTR-98. USDA Forest Service, Rocky Mountain Research Station, Ogden, Utah, USA.

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin. 2004. Bayesian data analysis. Second edition. Chapman and Hall/CRC, Boca Raton, Florida, USA.

Gotway, C. A., and L. J. Young. 2002. Combining incompatible spatial data. Journal of the American Statistical Association 97:632–648.

Gu, W., and R. K. Swihart. 2004. Absent or undetected? Effects of non-detection of species occurrence on wildlife-habitat models. Biological Conservation 116:195–203.

Hanson, T., W. O. Johnson, I. A. Gardner, and M. Georgiadis. 2003. Determining the infection status of a herd. Journal of Agricultural, Biological, and Environmental Statistics 8(2):469–485.

Hawksworth, F. G., and D. Wiens. 1972. Biology and classification of dwarf mistletoes (*Arceuthobium*). Agricultural Handbook AH401. USDA Forest Service, Washington, D.C., USA.

Hoeting, J. A., M. Leecaster, and D. Bowden. 2000. An improved model for spatially correlated binary responses. Journal of Agricultural, Biological, and Environmental Statistics 5(1):102–114.

Hooten, M. B., D. R. Larsen, and C. K. Wikle. 2003. Predicting the spatial distribution of ground flora on large domains using a hierarchical Bayesian model. Landscape Ecology 18(5):487–502.

Hooten, M. B., and C. K. Wikle. 2009. Optimal spatio-temporal hybrid sampling designs for ecological monitoring. Journal of Vegetation Science 20:639–649.

Keating, K. A., and S. Cherry. 2004. Use and interpretation of logistic regression in habitat-selection studies. Journal of Wildlife Management 68:774–789.

Lee, L. M., S. M. Teutsch, S. B. Thacker, and M. E. St. Louis. 2010. Principles and practice of public health surveillance. Third edition. Oxford University Press, New York, New York, USA.

MacKenzie, D. I., L. L. Bailey, and J. D. Nichols. 2005. Investigating species co-occurrence patterns when species are detected imperfectly. Journal of Animal Ecology 73:546–555.

MacKenzie, D. I., J. D. Nichols, J. E. Hines, M. G. Knutson, and A. D. Franklin. 2003. Estimating site occupancy, colonization and local extinction when a species is detected imperfectly. Ecology 84:2200–2207.

Muir, J. A., and B. W. Geils. 2002. Management strategies for dwarf mistletoe: silviculture. Pages 83–94 *in* B. W. Geils, J. C. Tovar, and B. Moody, technical coordinators. Mistletoes of North American conifers. General Technical Report RMRS-GTR-98. USDA Forest Service, Rocky Mountain Research Station, Ogden, Utah, USA.

Ogle, K. 2009. Hierarchical Bayesian statistics: merging experimental and modeling approaches in ecology. Ecological Applications 19:577–581.

R Development Core Team. 2009. R: A language and environment for statistical computing. Version 2.9.1. R Foundation for Statistical Computing, Vienna, Austria.

Reich, R. M., P. W. Mielke, and F. G. Hawksworth. 1991. Spatial analysis of ponderosa pine trees infected with dwarf mistletoe. Canadian Journal of Forest Research 21:1808–1815.

Reid, N., and S. F. Shamoun. 2009. Contrasting research approaches to managing mistletoes in commercial forests and wooded pastures. Botany 87:1–9.

Rejwan, C., N. C. Collins, J. Brunner, B. J. Shuter, and M. S. Ridgway. 1999. Tree regression analysis on the nesting habitat of smallmouth bass. Ecology 80:341–348.

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

Royle, J. A., and W. A. Link. 2006. Generalized site occupancy models allowing for false positive and false negative errors. Ecology 87:835–841.

Royle, J. A., and J. D. Nichols. 2003. Estimating abundance from repeated presence–absence data or point counts. Ecology 84:777–790.

Salman, M. 2003. Animal disease surveillance and survey systems: methods and applications. Iowa State University Press, Ames, Iowa, USA.

Schabenberger, O., and C. A. Gotway. 2004. Statistical methods for spatial data analysis. Chapman and Hall/CRC, Boca Raton, Florida, USA.

Taylor, J. 1996. An introduction to error analysis: the study of uncertainties in physical measurements. University Science Books, Sausalito, California, USA.

Tenenbein, A. 1970. A double sampling scheme for estimating from binomial data with misclassifications. Journal of the American Statistical Association 65:1350–1361.

Tyre, A. J., B. Tenhumberg, S. A. Field, D. Niejalke, K. Parris, and H. P. Possingham. 2003. Improving precision and reducing bias in biological surveys: estimating false-negative error rates. Ecological Applications 13:1790–1801.

APPENDIX A

Simulation study (*Ecological Archives* A021-053-A1).

APPENDIX B

Derivation of full-conditional distributions (*Ecological Archives* A021-053-A2).

APPENDIX C

Predictions for stands in the Minnesota Department of Natural Resources survey (*Ecological Archives* A021-053-A3).

SUPPLEMENT 1

R package containing annotated code used to conduct the analysis in this paper (*Ecological Archives* A021-053-S1).

SUPPLEMENT 2

Tables of stand characteristics and results for all stands in the Minnesota Department of Natural Resources survey (*Ecological Archives* A021-053-S2).