Ecological Applications, 20(8), 2010, pp. 2131–2147

Spatiotemporal exploratory models for broad-scale survey data

Daniel Fink,[1,6] Wesley M. Hochachka,[1] Benjamin Zuckerberg,[1] David W. Winkler,[2] Ben Shaby,[3] M. Arthur Munson,[4] Giles Hooker,[3] Mirek Riedewald,[5] Daniel Sheldon,[4] and Steve Kelling[1]

[1]*Cornell Lab of Ornithology, Ithaca, New York 14850 USA*
[2]*Department of Ecology and Evolutionary Biology, Cornell University, Ithaca, New York 14850 USA*
[3]*Biological Statistics and Computational Biology, Cornell University, Ithaca, New York 14850 USA*
[4]*Department of Computer Science, Cornell University, Ithaca, New York 14850 USA*
[5]*College of Computer and Information Science, Northeastern University, Boston, Massachusetts 02115 USA*

Abstract. The distributions of animal populations change and evolve through time. Migratory species exploit different habitats at different times of the year. Biotic and abiotic features that determine where a species lives vary due to natural and anthropogenic factors. This spatiotemporal variation needs to be accounted for in any modeling of species' distributions. In this paper we introduce a semiparametric model that provides a flexible framework for analyzing dynamic patterns of species occurrence and abundance from broad-scale survey data. The spatiotemporal exploratory model (STEM) adds essential spatiotemporal structure to existing techniques for developing species distribution models through a simple parametric structure without requiring a detailed understanding of the underlying dynamic processes. STEMs use a multi-scale strategy to differentiate between local and global-scale spatiotemporal structure. A user-specified species distribution model accounts for spatial and temporal patterning at the local level. These local patterns are then allowed to "scale up" via ensemble averaging to larger scales. This makes STEMs especially well suited for exploring distributional dynamics arising from a variety of processes. Using data from eBird, an online citizen science bird-monitoring project, we demonstrate that monthly changes in distribution of a migratory species, the Tree Swallow (*Tachycineta bicolor*), can be more accurately described with a STEM than a conventional bagged decision tree model in which spatiotemporal structure has not been imposed. We also demonstrate that there is no loss of model predictive power when a STEM is used to describe a spatiotemporal distribution with very little spatiotemporal variation; the distribution of a nonmigratory species, the Northern Cardinal (*Cardinalis cardinalis*).

Key words: citizen science; ensemble model; exploratory analysis; multiscale; sample bias; semiparametric; spatiotemporal; survey data.

Introduction

Understanding the determinants of species distributions and why distributions change through time is an important aspect of ecology and is critical for conservation and management. Many animal populations exhibit long-term temporal variation in distribution and abundance due to a variety of mechanisms. For example, distributions respond to changes in the biotic and physical environments (Thomas and Lennon 1999, Brommer 2004, MacLean et al. 2008, Schummer et al. 2008), or colonization of or adaptation to new and suitable environments (Dhondt et al. 2005, Strayer 2009). Other species undergo distribution changes on shorter time scales, such as daily or annual migrations. Describing temporal variation in distributions and identifying the environmental features with which species are associated throughout their movements is essential for developing comprehensive conservation policies.

Recognizing the importance of these issues, there has been an increasing effort to collect, assemble, and organize the ecological data needed to understand how species distributions change through time (e.g., the Global Biodiversity Information Facility; data *available online*).[7] However, data on species occurrences and abundances are, at best, sparsely distributed in space and through time, necessitating the ability to accurately interpolate where data were not collected (Scott 2002). By relating environmental predictors to observed occurrences or abundances, species distribution models can make predictions at unsampled locations and times.

Modeling dynamic species distributions requires that analyses deal with spatiotemporal variation across multiple scales. Systems often exhibit strong homogeneity when viewed at "fine" or "local" scales (Dungan et al. 2002, Beever et al. 2006). There are many processes that induce similarity of nearby observations. For

Manuscript received 27 July 2009; revised 22 January 2010; accepted 25 January 2010. Corresponding Editor: R. L. Knight.
[6] E-mail: df36@cornell.edu

[7] ⟨http://www.gbif.org⟩

example, the fine-scale spatial and temporal patterning of resources induces corresponding local distribution patterns (Smith 1974, Fortin and Dale 2005) and juvenile dispersal limitations help define the extent of "locality" (Waser and Elliott 1991, Anselin 1995). The importance of accounting for spatial (Lichstein et al. 2002, Royle et al. 2002, Banerjee et al. 2004, Latimer et al. 2006) and temporal correlation (Hurlbert 1984) has been broadly recognized. In contrast to fine-scale homogeneity, systems often exhibit strong heterogeneity when viewed at "coarse" or "global" scales. For example, it is known that individuals of the same species often occupy different specialized habitats at the edges of their distributions (Brown et al. 1995) and population dynamics processes such as the Allee effect (Groom 1998, Keitt et al. 2001) and source–sink dynamics (Mouquet and Loreau 2003) can create spatial patterning at relatively large spatial scales. In the temporal domain, large-scale effects like El Niño/La Niña and North Atlantic Oscillation (Lima et al. 2002, Grosbois et al. 2008) create strong, relatively abrupt changes in population size and composition.

Most statistical species distribution models have been developed to estimate static distributions (Latimer et al. 2006, Maggini et al. 2006, Royle et al. 2007, Thogmartin et al. 2007). These methods are based on hierarchical linear models that take into account local-scale spatial patterning via spatial covariates and parametric spatial correlation models. Recent extensions of this framework have been used to analyze dynamic species distributions by adding the structure to link distributions through time. For example, Link and Sauer (2007) included seasonal components of population change to estimate changes between breeding and winter seasons; Wikle (2003) used integro-difference equations to model the range expansion; and vector autoregressive processes have been used to study dynamics in forest growth (Hooten and Wikle 2007). The success of this approach depends strongly on the analyst's ability to correctly specify the important coarse-scale processes that control distribution dynamics.

However, for many applications, the dominant processes are unknown or are not understood in sufficient detail to specify as a parametric statistical model. Nonparametric techniques offer an alternative approach for modeling broad-scale species distributions (De'ath and Fabricius 2000, Elith et al. 2006, 2008, De'ath 2007, Hochachka et al. 2007). The essential feature of these nonparametric models is that they automatically adapt to patterns in data, reducing the amount and detail of information that must be supplied by the user. The ability to automatically discover patterns also makes these methods especially well-suited for exploratory analysis and subsequent hypothesis generation (Hochachka et al. 2007, Kelling et al. 2009). Many nonparametric methods have been successfully used to study static distributions (Phillips et al. 2006, Prasad et al. 2006, Hochachka et al. 2007, Elith et al. 2008).

The challenge with using nonparametric techniques to model spatiotemporally varying distributions is that without any spatiotemporal structure the models are free to share predictor information across regions or seasons in ways that are impossible in nature. Later in this paper we show how strong habitat–occurrence relationships identified in one region can be erroneously applied to distant regions giving rise to biased predictions and inferences. Thus, to conduct successful spatiotemporal exploratory analyses, existing techniques need to be adapted to provide sufficient spatiotemporal structure while not constraining the analysis too narrowly, restricting the utility of the model. To date, no such methods exist.

In this paper, we introduce a novel methodology for adding essential spatiotemporal structure to existing static species distribution models without requiring detailed specification of the underlying dynamic processes. This is achieved by creating an ensemble, or mixture model, from a population of static species distribution models each applied to a spatiotemporally restricted extent. By restricting each model to a local region, we can account for local-scale spatial and temporal patterns and control the risk of extrapolation to distant regions. Partitioning the study area into many local regions gives the ensemble extra flexibility to adapt to coarse-scale heterogeneity. Predictions are made by averaging across local models with shared extents, allowing for local-scale patterns to "scale up" to patterns at the global scale (sensu Levin 1992). We call this method the spatiotemporal exploratory model (STEM).

The motivation for this work was to explore the continent-wide inter-annual migrations of common North American birds using data from the citizen science project, eBird (Sullivan et al. 2009; information *available online*).[8] Understanding temporal variation in species distributions is critical for developing conservation strategies for migratory species (Greenberg and Marra 2005). This is challenging in part because of the great variation in migration dynamics between species. To deal with this, we sought to develop a highly automated STEM capable of producing objective, dynamic species distribution estimates with a minimum of user inputs. Therefore, the implementation presented here focuses on exploratory analysis with decision trees (Breiman et al. 1984, Quinlan 1993) with an ensemble designed specifically for intra-annual migration dynamics.

Developing models that can account for spatial and temporal structuring in species' responses to their environment is one component of an effective modeling framework, but the effectiveness of the modeling process

[8] ⟨http://www.ebird.org⟩

also depends on appropriate evaluation of model accuracy. A common approach to evaluation is to hold out data from the model development, either by splitting the data, k-fold partitioning (Wiens et al. 2008), or bootstrapping (Harrell 2001). Accuracy is assessed through the quality of predictions on the hold out set (Fielding and Bell 1997, Hastie et al. 2001). However, broad-scale survey data collected by volunteers, like eBird, often have strong spatial heterogeneity in sampling intensity, at multiple scales. At a fine resolution, these data tend to be biased toward roadsides and human population centers (Reddy and Davalos 2003), whereas at a continental scale, data density is highest in regions of highest human density. This sampling bias can influence the evaluations by placing too much emphasis in those data regions that are most intensively sampled.

We addressed this problem with a novel model evaluation procedure for controlling spatial sampling bias and then used it to demonstrate the utility of STEM analysis for producing models of distributions of migratory and nonmigratory bird species with eBird data.

Data Used in Examples

Data for our study come from eBird, a citizen science project launched by the Cornell Lab of Ornithology and the National Audubon Society in 2002 (Sullivan et al. 2009): eBird engages a large network of bird watchers, who are highly motivated and self-trained, to report bird observations from around the world. Participants follow a checklist protocol, where time, location, and counts of birds are all reported in a standardized manner. A network of bird distribution experts (more than 400 participants) manage the observations through a series of region- and time-specific checklist filters (more than 800) to ensure data quality. These filters are used to identify species occurrences that are considered unusual, given the date and location of the observation. When unusual sightings are submitted, they are sent to an expert for review and possible acceptance into the database. For this reason, eBird data may be somewhat conservative for measuring changes in timing or location of distributions.

eBird is unique among broad-scale North American bird monitoring projects in that it collects observations made throughout the year. The goal of our application is to use these year-round data to produce continent-wide occurrence maps for all times of the year for a wide range of North American bird species.

Participants record counts of bird species detected visually or acoustically, the location where their search took place, and the time that they initiated the search. eBird reporting rates are highest during the spring and fall migrations (e.g., in May of 2009 eBird gathered more than 1.7 million observations of birds), and are lower during the breeding and winter seasons (approximately 1 million observations of birds in January or June 2009). By asking participants to indicate when they have contributed "complete checklists" of all the species they detected, we can assume species with no detected individuals conveys absence information for that location. This distinguishes eBird as a "presence–absence" data protocol. A subset of eBird participants use standardized protocols designed to collect additional information on search effort. For example, under the "traveling count" protocol participants record the amount of time spent and the transect distance traveled while searching for birds. Together, the reports of absence and effort data add valuable information allowing the analytical control of variable detection rates when inferring absences.

In this paper we used presence–absence data from complete checklists collected under the "traveling count protocol" from 2004 to 2007. Under this protocol all observations are tied to a single location even though birds were recorded continuously along the path that was travelled. Participants are encouraged to record this location at the middle of their transect. However, because some participants record the transect locations differently, there is increasing location uncertainty with increasing transect length. In addition, we were concerned that very long transects would be subject to additional noise because of variation in the mode and speed of transportation. Therefore, we decided to adopt a convenient limit for transect distances, 8.1 km (5 miles). Start times were restricted to daylight hours between 05:00 and 20:00 and the total search time was limited to <3 hours to make observations more comparable. The study area includes all of the conterminous United States. Fig. 1 shows the 9799 unique locations for 57 863 complete traveling count checklists submitted during the study period. There is substantial spatial variation in sampling intensity visible from the large empty map regions in much of the Western and Central United States to the intensively sampled metropolitan areas throughout the country.

We included effort variables in all analyses to account for variation in detection rates. Additionally, variation in availability for detection (Diefenbach et al. 2007) was modeled as a function of the observation time of day and day of year. To account for habitat selectivity of bird species, each eBird location was linked with a set of remotely sensed habitat information. We selected a set of spatial covariates that we believed would be generally useful for capturing occurrence–habitat associations across a wide number of avian species.

The data used to describe habitat are from the U.S. 2001 National Land Cover Database (Homer et al. 2007), which includes a raster GIS layer of classification of vegetation with 30-m cell resolution and a 0.40-ha minimum mapping unit. Land cover was classified into one of 15 classes within our study region. Two additional layers were obtained describing the percentage of canopy cover and the percentage of impervious surface, also with 30-m cell resolutions. Because birds

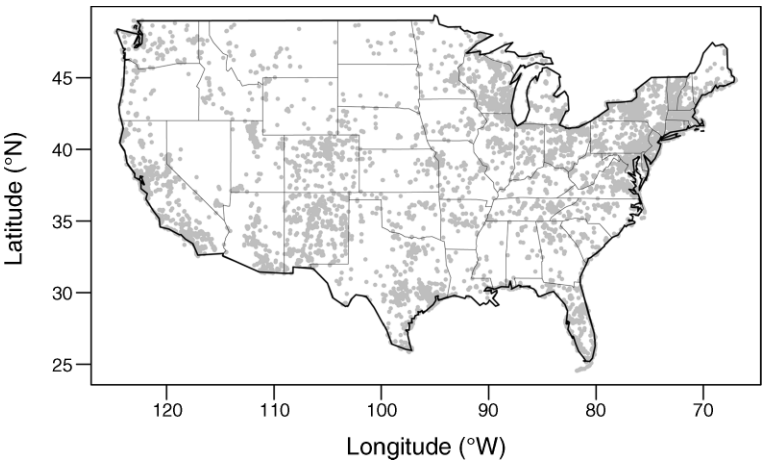

FIG. 1. The eBird data locations, 2004–2007. Presence–absence data were taken from 57 863 complete checklists submitted during the study period. The study area includes 9799 unique data locations across the conterminous United States.

may respond to their habitat over a range of spatial scales, we created predictors describing the percentage of coverage for each of the land-cover classes above (including canopy and impervious coverage) in square neighborhoods at two sizes (2.25 ha and 20.25 ha) centered on each location. We selected these two spatial scales because we have found them to generate informative predictors in previous studies of broad-scale avian distributions for large groups of species.

To account for additional anthropogenic effects we used human population density estimates from the U.S. Census Bureau 2000 census block-level summaries. Elevation measured at 30 arc second (GTOPO30) resolution was also included (data *available online*).[9]

Finally, we included the observation latitude and longitude, in addition to the day of the year and year, to ensure sufficient predictor information to adapt to any spatiotemporal patterns insufficiently represented by the other predictors. There were a total of 48 predictors used in the model (Table 1).

Model Development

Ensembles

Ensemble models have emerged as some of the most powerful nonparametric methods (Breiman 1996, 2001, Freund and Schapire 1996, Friedman 2001, Friedman and Popescu 2008) and are the basis for our new technique. Let y_i, $i = 1, \ldots, N$, be a set of responses each associated with p predictors $\mathbf{x}_i = [x_{1,i}, \ldots, x_{p,i}]$. It is assumed that each observation, y_i, conditioned on $\mathbf{x}_i$, arises as a realization from some true but unknown function, $F^*(\mathbf{x}_i)$ that maps $\mathbf{x}_i$ to y_i. The goal is to use data to estimate $F^*(\mathbf{x}_i)$ while minimizing the expected value of a specified loss function.

[9] ⟨http://eros.usgs.gov/#/Find_Data/Products_and_Data_Available/gtopo30_info⟩

Here we consider ensemble models that are discrete mixture models,

$$F(\mathbf{x}) = \sum_{m=1}^{M} a_m f_m(\mathbf{x})$$

where M is the size of ensemble and each ensemble member or base model $f_m(\mathbf{x})$ is a different function of the predictors $\mathbf{x}$ derived from the data. Ensemble predictions $F(\mathbf{x})$ are taken to be a linear combination of the predictions from each of the ensemble members. The parameters $(a_1, \ldots, a_m)$ determine the mixture weights. An ensemble model is specified by the choice of the particular class of functions used for the base models, a prescription how to construct the M base models and the mixture weights using a specified set of observations and predictors (the training data).

Decision trees

Decision trees (DTs) have several features that make them a good choice as base model, especially for exploratory analysis (Hochachka et al. 2007, Elith et al. 2008). DTs are nonparametric models designed to automatically screen large numbers of predictors to identify the most important ones while determining the shape of the functional relationships between predictors and response, including high order interactions (Breiman et al. 1984, Quinlan 1993). Additionally, most DT implementations automatically impute missing predictors.

These models use a binary recursive partitioning strategy to adaptively search high-dimensional tensor-product predictor spaces. Each split partitions the training data based on values of a single predictor variable. Splits are chosen to maximize information content from among all potential splits among all predictors. Thus, each predictor can appear multiple times within a decision tree in different decision rules, or not at all. Node predictions are taken to be the average

of the training samples at that node. Decision trees have been implemented for many different types of response variables including continuous responses, counts, and nominal categories. In this paper we use "classification" trees to model the binary responses arising from presence–absence data.

There are two key specifications for DTs, the function used to measure information and the strategy used to end the tree-growing process. Often the type of response variable determines which functions are used to measure information. Like any highly flexible model, a single decision tree can over-fit data, producing a model that is too highly tailored to a specific sample of data. Early methods for dealing with over-fitting used various rules to prune back the branches of a decision tree (e.g., Breiman et al. 1984). Current ensemble methods use new strategies to control overfitting by combining information from multiple decision trees.

Bagged decision trees

In order to control the highly variable predictions of overfit DTs, Breiman (1996) suggested averaging predictions from an ensemble of DTs generated by bootstrapping the training data. Each base model DT is deliberately overfit by growing it to its maximal size to produce a low-bias, high-variance estimator. We will refer to these as "saturated" trees. Ensemble averaging is used to control the between-model variance. These "Bootstrap AGgregations" are known as "bagged" decision trees (hereafter BDT) and consistently improve predictive performance compared to single trees.

BDTs are a good, general purpose predictive model with predictive performance competitive with other current exploratory models. They have been used successfully for static species distribution modeling applications (Prasad et al. 2006, Hochachka et al. 2007). Caruana and Niculescu-Mizil (2006) compared 10 binary classification models (including the three tree-based models BDTs, RandomForests [Breiman 2001], and boosted regression trees [Friedman 2001]) using eight performance metrics across 11 different data sets. They concluded that RandomForests and BDTs were among the top overall performers. Boosted regression trees outperformed RandomForests and BDTs only after the predictions went through a post-modeling probability calibration. Elith and Graham (2009) concluded that boosted regression trees outperformed RandomForests, a tree-based method very similar to BDTs, for their static species distribution comparison.

One of the distinguishing features of BDTs is their simplicity; no parameters need to be fit or tuned during model training. Bootstrap samples are generated from the training data by sampling with replacement. Simple ensemble averaging, $a_i = 1/M$, $i = 1, \ldots, M$, controls overfitting, making it feasible to fit relatively complex, saturated trees thereby avoiding any additional parameter tuning or computation to control tree complexity.

TABLE 1. Predictors used in the model.

Predictor	Definition
YEAR	observation year
JDATE	observation day of year
HOUR	observation hour
LONGITUDE	degree longitude
LATITUDE	degree latitude
EFFORT_HRS_ATMOST	effort search time (hr)
EFFORT_DISTANCE_KM	effort distance traveled (km)
POP00_SQMI	human population density (number per square mile)†
GTOPO30_ELEVATION	elevation (m above mean sea level)
NLCD01_LANDCOVER	NLCD class (30-m resolution)
NLCD01_IMPERV	NLCD impervious surface (30-m resolution)
NLCD01_CANOPY	NLCD canopy (30-m resolution)
NLCD 2.25-ha neighborhood	
NLCD01_N11A10R15	open water
NLCD01_N12A10R15	perennial ice/snow
NLCD01_N21A10R15	developed, open space
NLCD01_N22A10R15	developed, low intensity
NLCD01_N23A10R15	developed, medium intensity
NLCD01_N24A10R15	developed, high intensity
NLCD01_N31A10R15	barren land
NLCD01_N41A10R15	deciduous forest
NLCD01_N42A10R15	evergreen forest
NLCD01_N43A10R15	mixed forest
NLCD01_N52A10R15	shrub/scrub
NLCD01_N71A10R15	grassland/herbaceous
NLCD01_N81A10R15	pasture/hay
NLCD01_N82A10R15	cultivated crops
NLCD01_N90A10R15	woody wetlands
NLCD01_N95A10R15	emergent herbaceous wetlands
NLCD01_CMNA10R15	canopy
NLCD01_IMNA10R15	impervious surface
NLCD 20.25-ha neighborhood	
NLCD01_N11A10R5	open water
NLCD01_N12A10R5	perennial ice/snow
NLCD01_N21A10R5	developed, open space
NLCD01_N22A10R5	developed, low intensity
NLCD01_N23A10R5	developed, medium intensity
NLCD01_N24A10R5	developed, high intensity
NLCD01_N31A10R5	barren land
NLCD01_N41A10R5	deciduous forest
NLCD01_N42A10R5	evergreen forest
NLCD01_N43A10R5	mixed forest
NLCD01_N52A10R5	shrub/scrub
NLCD01_N71A10R5	grassland/herbaceous
NLCD01_N81A10R5	pasture/hay
NLCD01_N82A10R5	cultivated crops
NLCD01_N90A10R5	woody wetlands
NLCD01_N95A10R5	emergent herbaceous wetlands
NLCD01_CMNA10R5	canopy
NLCD01_IMNA10R5	impervious surface

Notes: NLCD is the U.S. 2001 National Land Cover Database. We used predictors describing the percentage of coverage for each of the land-cover classes in square neighborhoods at two sizes (2.25 ha and 20.25 ha) centered on each location.

† One square mile = 259 ha.

This approach contrasts with boosted regression trees, which requires the specification of several important tuning parameters (interaction depth, shrinkage rate, and the number of iterations) to control complexity. Because these parameters are almost always unknown, fitting these models requires estimation with indepen-

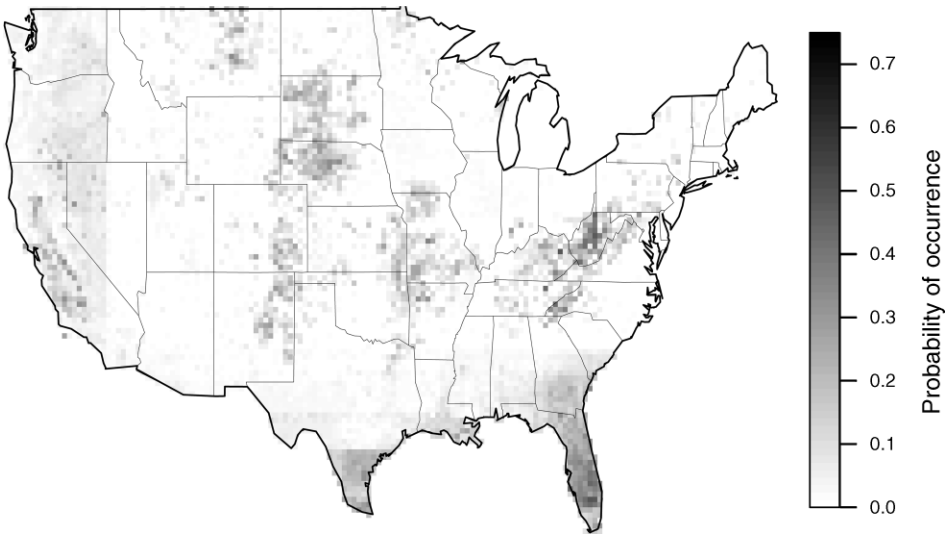

FIG. 2. Tree Swallow winter distribution estimate using bagged decision trees. This surface shows the predicted probability of reported occurrence for Tree Swallows on 9 January 2006. Significant concentrations of Tree Swallows appear in areas where they are known to not occur at this time of year. These "false positives" are strongest in the Appalachians, Missouri, and through the Shortgrass Prairie region, with several smaller high-probability points scattered throughout the western United States.

dent data (Ridgeway 2007, Elith et al. 2008) requiring additional computation (e.g., cross validation). For these reasons, we decided to use BDTs as the basis for this study where we utilize very large ensembles of DTs.

We began by testing the effectiveness of BDTs to study dynamic annual distributions of Tree Swallow (*Tachycineta bicolor*), a common migratory passerine bird. We chose this species because it is an easily identified, conspicuous, social bird with broad distribution across the continental United States, generating data with relatively high detection rates with few misidentifications. Additionally, we selected Tree Swallow because its year-round broad-scale distribution is relatively well understood (Robertson et al. 1992, Winkler 2006).

The BDT was fit using all training data across the entire extent of the study area (see *STEM application and evaluation*). Thus, the BDT had sufficient predictor information to adapt to fine-scale temporal variation (hourly) and fine-scale spatial variation (30 m). On the other hand, because the BDT is so highly automated, it faced the task of finding spatial and temporal signal on its own—without a priori information or analytical "requirements" telling the model what predictors to include or functional forms to fit. Fig. 2 shows the predicted probability of occurrence for Tree Swallow on 9 January 2006. With the land cover predictor variables used in this analysis, the model correctly identified known wintering grounds along the Gulf and Atlantic coasts in the east, and along the central and southern California coasts in the west. However, this map also predicted significant concentrations of Tree Swallows in areas where they are known to not occur in the middle of winter. These "false positives" are strongest in the Appalachians, Missouri, and through the Shortgrass Prairie region, with several smaller high-probability points scattered throughout the western United States. All of these locations are far from suitable food sources and known populations in winter. For ecological and conservation planning, these "false positives" are significant errors that limit the utility of the model (see *STEM application and evaluation*).

Analytically, these errors arise when strong habitat–occurrence associations learned from one region and season are spuriously applied to other regions and/or seasons. Thus, the model has failed to recognize the essential spatiotemporal "scale" of Tree Swallow migration, sharing habitat information across regions where Tree Swallows do not coexist in time. This problem is not specific to this analysis; it is a general problem that can arise for any model whose development involves sharing information across large spatial and temporal extents, regardless of the particular predictors used. The risk of this problem will be greater for more flexible models and predictor sets with greater variation.

The spatiotemporal exploratory model (STEM)

The spatiotemporal exploratory model (STEM) is an ensemble model designed to include essential information about spatial and temporal scales. This is accomplished by restricting each base model to a local spatial and temporal region. Thus there are two key STEM parameters: the spatial scale parameter controls the size of the spatial subregions from which data are analyzed and the temporal scale parameter controls the length of the time period within each temporal sub-region. We use DTs as the base model to conduct local-scale exploratory modeling.

Each ensemble member takes as input predictors $\mathbf{x}$ at location s and time t

$$f_i(\mathbf{x}, s, t)$$

where i indexes the ensemble. Model f_i has an associated parameter, θ_i, that defines the region in space and time where f_i is trained and is allowed to make predictions. In statistics, this is called the support set. For example, in the next section parameter θ_i is defined as the set of all locations within rectangular region S and all times t falling between the two dates, $[\alpha, \beta]$, $\theta = \{(s, t) \mid s \in S, t \in [\alpha, \beta]\}$, STEM predictions are computed as the average prediction taken across the ensemble members with shared support

$$F(\mathbf{x}, s, t) = \frac{1}{n(s,t)} \sum_{i=1}^{M} f_i(\mathbf{x}, s, t) I((s, t) \in \theta_i).$$

$I((s, t) \in \theta_i)$ is the indicator function, taking value 1 when the location and time are within support set θ_i, and zero otherwise. Here the indicator function identifies the subset of ensembles with support that contain location s and time t. Function

$$n(s, t) = \sum_{i=1}^{M} I((s, t) \in \theta_i)$$

calculates the number of ensemble models supporting the prediction at (s, t); we will call this quantity the "prediction support." For simplicity, all supporting predictions are weighted equally. Support sets are randomly sampled from a spatiotemporal distribution $p(\theta)$ in such a way to ensure sufficient spatial and temporal overlap.

Thus, each ensemble prediction is made for a particular date and location and is computed as the average prediction made by all base models that contain that date and time. This local averaging achieves important ecological and statistical modeling objectives. Ecologically, local averaging is the mechanism that allows local-scale patterns to "scale up" to coarse-scale patterns. Statistically, the local averaging procedure is similar to a randomized block design on the support sets where neighboring blocks function like replicates. These neighboring replicates are used to average out inter-model (i.e., block level) variation among the predictions. The locality of the averaging also performs a function similar to that of parametric models of spatial and temporal correlation giving larger weights to nearby predictions. When true large-scale heterogeneity exists, a well designed STEM ensemble will increase the variation between predictions from different base models while reducing the variation among predictions within each model. According to Breiman's (2001) theorem, this will result in improved predictive performance.

STEM design for inter-seasonal dynamics

In order for the model to handle a wide variety of intra-annual migration dynamics, we need to specify a sufficiently flexible sampling strategy for the base model support sets and training data.

Table 2. Pseudo-code for the STEM training algorithm.

A) Parametric specifications
  Temporal scale: specify number and width of temporal windows
  Spatial scale: specify GSRD and dimensions for spatial support rectangles
  Specify support set minimum data requirement.
B) For each temporal window, fit saturated decision trees as follows:
  1) Randomize location of GSRD grid
  2) Sample realization of support set from GSRD
  3) For each support set
    a) Subsample training data within support set
    b) Check ensemble minimum data requirement
    c) Fit model $f_i(\mathbf{x}, s, t)$
  End for
End for

Note: GSRD stands for geographically stratified random design.

First, the temporal windows were defined as evenly spaced, overlapping intervals spread throughout the year. Second, for each temporal window, a randomized set of rectangular spatial support sets was selected from a geographically stratified design to ensure maximal spatial coverage across the study area. Third, a subsample of data within the spatiotemporal support set was used as the training data. Pseudo code for the STEM training algorithm is presented in Table 2 and discussed in more detail in the following subsections.

Temporally, we made a simplifying assumption that there was more day-to-day variation in distributions for a fixed year than the amount of year-to-year variation in distributions observed at a fixed date. Thus, for this design, we decided to model year-to-year variation at the local level by including year as a base model predictor. Then we accounted for day-to-day variation in distribution by specifying a population of temporally restricted support sets. Each temporal support set was defined as an interval on the days of the year. To smoothly model migration dynamics throughout the year we created an evenly spaced set of overlapping temporal windows. By limiting the number of windows and their relative width, we set boundaries on the maximum temporal period over which information was shared (i.e., the longest period over which we assume that relationships between predictors and response are stable). In the examples presented in later sections of the paper, we used 80 temporal windows centered at equally spaced days throughout the year. Each window was 40 days wide for model training. This window width was selected to be small enough to yield informative monthly distributional predictions. For the purpose of making STEM predictions, we further restricted temporal window support to the central 36-day window to limit bias when predicting near support "edges."

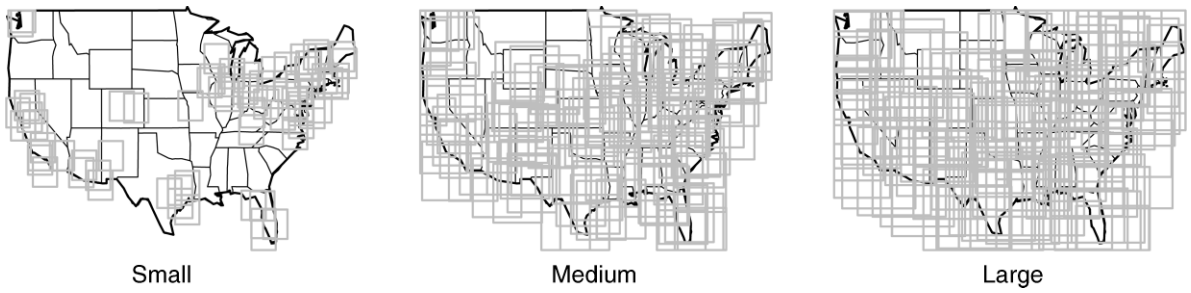

FIG. 3. Typical realizations of spatial support sets at small, medium, and large sizes. All spatial support sets were made up of rectangles with fixed latitudinal height and longitudinal width. The size (area) of the support rectangles controlled the trade-off between the coverage of the study area and the risk of extrapolation due to spurious "long-range" learning.

The goal for designing the STEM spatial support sets was to facilitate adaptation to as wide a variety of intra-annual migrations as possible. Given the variety of plausible migration patterns we located support sets uniformly across the study area. To do this we used a geographically stratified random design (GSRD), where strata are defined by a regular grid of "latitude × longitude" cells overlaying the study area. We define a GSRD by its "latitude × longitude" dimensions. Each cell is a unique "stratum" and is sampled uniformly to create support set "centers." Sampling a fixed number of centers per grid cell gives equal spatial weighting across the ensemble of base models. The "position" of the grid is randomly located for each realization of the GSRD. We will use the same GSRD to construct uniformly distributed test sets and prediction locations, below.

The spatial support sets are rectangles with fixed latitudinal height and longitudinal width centered around sampled "center" coordinates. Varying the area (or scale) of these support rectangles affects the amount and density of data and the level of prediction support. Within each spatiotemporal support set θ_i we subsampled the training data. We randomly sample 63% of the data locations, the percentage of unique data points expected in a bootstrap sample on average. A minimum of 25 unique traveling count training locations within each support set was specified as the minimum data requirement. Support sets that did not meet this requirement were thrown out, reducing ensemble coverage. Together, the minimum data requirement and the scale of the support sets define the maximum extent, minimum data density, and maximum degree of model extrapolation.

In the examples that follow, we consider STEM at three spatial scales: "small," "medium," and "large." Typical realizations of the spatial rectangles at small, medium, and large scales are show in Fig. 3. The "small" scale has the smallest GSRD with 1.50° × 2.0° latitude by longitude cells and the smallest support set rectangles at 3° latitude by 4° longitude. This scale provided the most stringent restriction on long-range extrapolations and ensemble support; however, because the spatial distribution of eBird data is so concentrated large parts of the U.S. study area contained too few data to support predictions at this scale (Fig. 3). The "medium" size was generated from a GSRD with 2.12° × 2.83° cells and 6° × 9° support rectangles. The "large" support set was defined to be large enough to cover approximately the entire coterminous U.S. study area. It was generated from a GSRD with 2.60° × 3.46° cells and 9° × 12° support rectangles. All GSRD and support rectangle dimensions were selected to yield similar predictive support in data rich regions.

Model Evaluation with Spatially Biased Data

Empirical measures of accuracy are useful for model testing, diagnostics, and comparisons. In this section we describe how we measured accuracy for predicted seasonal species distributions when the models were constructed using data from a nonuniform spatial distribution. With strong variation in spatial sampling intensity there will be large areas of low sampling density and smaller regions with very high sampling density. If ignored, this spatial bias will produce model evaluations in which regions with the most data have excessive influence on the overall measure of model accuracy and regions with the least data are under-represented. This is an especially serious problem for many citizen science projects in which spatial bias follows patterns of human population density. Indeed, for conservation applications, we may be most interested in those regions that are the furthest from human population centers. The challenge for evaluating species distribution maps is to generate test set samples that adequately represent the uniform target population using the spatially biased data in hand.

Formally, lack of accuracy is defined by the statistical prediction risk

$$R(F) = E_{XY}L(y, F(\mathbf{x})).$$

Here the loss function $L(y, F(\mathbf{x}))$ describes how to penalize discrepancies between observations y and predictions $F(\mathbf{x})$ and the expectation is taken over the joint distribution of observations y and covariate $\mathbf{x}$. Functions with smaller risk are preferred as more "accurate." In practice one must specify (1) how to estimate the loss expectation and (2) define the loss

function to meet the desired modeling objectives. We describe these specifications in the subsections below.

The loss expectation

Map estimation, by definition, implies a target population that includes *all* locations in the study area. Moreover, we assume that inference quality at all locations is equally important. Thus, the expectation needs to be taken across a uniform distribution of locations. In practice, the goal is to draw Monte Carlo samples from the target population using locations sampled from observed locations. For broad-scale spatial analyses, observational data are at best sparsely distributed across the study area. Thus, map estimation as an inferential goal relies on a model's ability to make accurate "out of sample" predictions at locations where there were no observations. In spatial statistics, this is known as the spatial interpolation problem (Banerjee et al. 2004), and it is an essential consideration for evaluating the quality of a species distribution map.

Here we propose a straightforward subsampling algorithm for generating the test data used to evaluate the model. First, we split the data into training and test sets such that each set contains a distinct set of locations. This avoids location-specific biases in the risk assessment, an important consideration for bird monitoring data for which it is common to repeatedly monitor the same locations. Second, in order to maximize the spatial coverage of the evaluation, test set locations are uniformly subsampled with a GSRD. The grid cell size was specified small enough to measure predictive performance at a finer spatial resolution than STEMs smallest local support size. This we set at approximately 1/16 the area of the "small" scale STEM support size, 0.75° latitude and 1.10° longitude. We also set a per-grid-cell maximum sample size at 2 as a simple way to produce subsamples with more equitable spatial distribution for risk evaluation. For each selected location, all of the observations at that point are included in the test sample. Finally, we evaluate the statistical risk function over 50 Monte Carlo samples of the GSRD. This allows us to use more of the test data and to improve the risk estimate. Pseudo code for the species distribution map evaluation algorithm is presented in Table 3.

Table 3. Species distribution map evaluation algorithm.

```
A) Initialize:
   Split training and test sets such that they have no common
     locations
   Test set spatial scale: specify height and width for test set
     GSRD
B) For each Monte Carlo iteration, subsample test data as
   follows:
   1) Randomize GSRD location
   2) For each test set cell
      a) Randomly sample test locations up to per-grid-cell
         maximum
      b) Evaluate statistical risk & accumulate results across
         MC iterations
   End for
End for
```

Loss functions

Often, species distribution estimates are best displayed as a map of predicted probability of occurrence, or as a continuously varying occurrence score across a set of locations. Our primary inferential goal was to accurately rank occurrence probabilities so that the associated map could be used for the comparison of species prevalence across regions. For this reason we used the area under the receiver operating-characteristic curve (AUC) statistic. This statistic measures a model's ability to discriminate between positive and negative observations (Fielding and Bell 1997). The AUC is equal to the probability that the model will rank a randomly chosen positive observation higher than a randomly chosen negative one. Thus, AUC depends only on the ranking of the predictions. The AUC statistic ranges from 1.0 to zero. In order to preserve the "loss" interpretation and ordering (smaller is better) for evaluating species distribution maps, we use the AUC error defined as 1 − AUC, as our loss function. Thus, when 1 − AUC equals zero it indicates perfect discrimination, a value of 0.5 indicates random discrimination, and values greater than 0.5 are worse than random.

In addition to measuring ranking performance we also calculate the root mean squared error (RMSE) between predicted probabilities and observed outcomes. This measure takes into account the magnitude of the predicted probability, sometimes called calibration information, in addition to ranking. For measuring differences between binary observations and predicted probability scores, the RMSE ranges from 1.0 (worst) to 0.0 for a set of perfect predictions.

STEM Application and Evaluation

To demonstrate the utility of the STEM structure for modeling dynamic species' distributions, we compared STEMs, which constrain each model in the ensemble to meet very specific spatial and temporal requirements, to the unrestricted "global"-scale BDT models that utilize data from across the entire extent of the study area (i.e., the coterminous United States). We investigated if varying the spatial scale of STEM affected the accuracy of the predicted winter distribution for Tree Swallows. Subsequently, we examined the relative success of STEMs in predicting swallow distributions at different seasons, and then evaluated STEM performance for a species with a less dynamics distribution, the nonmigratory Northern Cardinal (*Cardinalis cardinalis*).

All computations were carried out in the R statistical computing language (R Development Core Team 2008) and individual decision trees were computed using the rpart library (Therneau and Atkinson 2008). For each data set we fit all three STEM scales. In addition, BDTs were fit with full temporal and spatial support, the

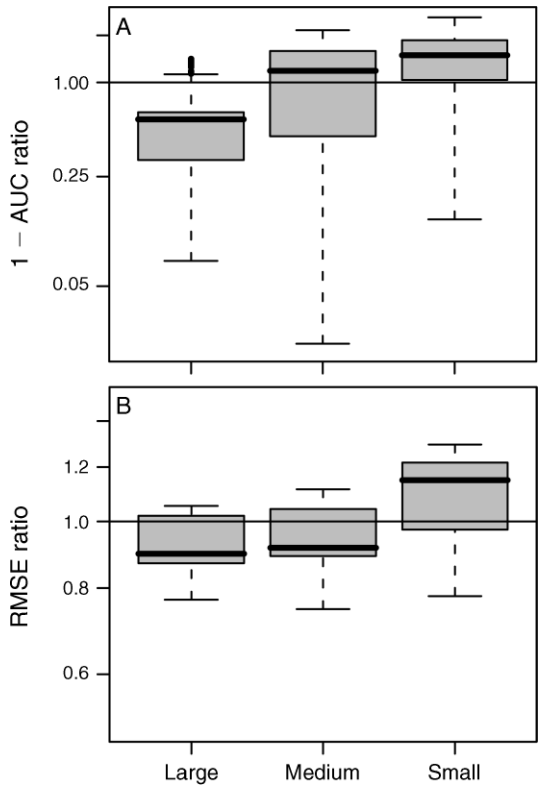

FIG. 4. Performance comparisons between spatiotemporal exploratory model (STEM) and bagged decision tree (BDT) models for Tree Swallow winter distribution. The ratio of STEM error (loss) to BDT model error (loss) is plotted for three STEM spatial scales on a log scale. When STEM has a smaller error than the BDT (i.e., the STEM model is more accurate), the ratio will be less than 1. Errors are calculated by (A) 1 − AUC (the area under the receiver operating-characteristic curve) and (B) root mean-square error. In the box plots, the top of the box shows the 75th percentile, the heavy dark line shows the median, and the bottom box shows the 25th percentile. The "whiskers" above and below the box show 1.5 times the interquartile range, and dots show data points beyond the whiskers.

"global support" case, using the same saturated DTs as base models. We used 125 bootstrap samples with the BDT so that predictions would be supported by a similar number of base models compared to STEM predictions. All models were trained and evaluated with five-fold validation. For each of the five "folds" the data were split into a random sample with 80% of the data locations used for training and the other 20% used for model evaluation, as described in *Model evaluation with spatially biased data*.

Tree Swallow winter distribution comparison

Predicting species distributions is most difficult when populations are most dynamic and exhibit the highest variability in observed occurrences. For migratory species, like the Tree Swallow, we expect the accuracy of estimated distributions to vary among seasons with the best performance during the winter and breeding seasons when populations are relatively static and the worst performance during the migrations. Here we compare STEM and BDT estimates of the winter distribution for Tree Swallow. By late December, if not earlier, most Tree Swallows have returned to their wintering grounds in the southeastern United States. They remain there through the month of January with the earliest spring migrants leaving during February.

To facilitate model comparison, we computed the ratio of STEM error (loss) to BDT model error (loss). When the STEM had a smaller error than the BDT, the ratio was less than one, in proportion to the difference in the error rates. Winter distribution map performance was evaluated using test set observations for a two week window in the middle of January. The STEMs at small, medium, and large scales were compared with the BDT. Each comparison was based on a total of 250 test trials: a single "test trial" was produced for each of the 50 Monte Carlo test set samples across all five data "folds."

The STEM structure improved prediction of this migratory species' winter distribution. Fig. 4 displays the performance comparison results for the Tree Swallow winter distribution. Panel A shows the distributions (box plots) of the AUC error ratio. The performance of STEM improved with increasing spatial scale. STEM outperformed BDT on almost all trials at the largest spatial scale. For this season, at least, STEM does better by pooling data across larger regions. Panel B shows the distributions of the RMSE error ratios. These results are qualitatively similar to the AUC results, STEM performance improved at larger spatial scales and STEM outperformed BDT on most of the trials at the largest spatial scale. However, the percent change in RMSE performance is smaller than for the corresponding AUC comparisons.

Tree Swallow monthly distribution performance

In this test we evaluated the performance of the best-scale STEM and global BDT models during each of 12 monthly distribution tests. Each monthly distribution test is conducted during the central two-week window for that calendar month. We considered a model to be superior if its error rate was smaller on at least half of the test trails.

As expected, distributions were described most accurately during the winter and breeding season months, when Tree Swallow distributions are most stable (Fig. 5). Panel A shows the 1 – AUC distributions for the best-scale STEM arranged by month. The center and lower panels show the 1 – AUC and RMSE ratios for STEM vs. BDT performance arranged by month. We plotted the ratios on the log-scale to facilitate model comparisons. STEM outperformed BDT in terms of AUC error for all months except July. Note that STEM performance was much better for January, February, and November: AUC errors were half as big as BDT errors on half of these test trials. The 1 – AUC relative performance varied widely in December. Averaging

across all months, the STEMs outperformed BDTs in terms of AUC error on 77% of the test trials.

The relative RMSE performance of STEMs followed a similar seasonal pattern to the AUC error. The STEM clearly outperformed BDTs at all times of the year except in July and August, when performance is nearly equal. When the fall migration begins, RMSE performance is similar between a STEM and a BDT. Note that the percent difference in RMSE errors between the STEM and BDT models is smaller than for AUC errors. Averaging across all months, STEMs outperforms BDTs in terms of RMSE on 71% of the test trials.

Tree Swallow seasonal occurrence maps

Royle et al. (2007) noted the importance of controlling for variation in detectability when displaying occurrence surfaces as maps. Here we demonstrated how predictor information can be used to control for sources of observational bias with a nonparametric model. We produced four daily occurrence maps for Tree Swallow, one for each season for both STEM and BDT models. The resulting maps show the improvements due to STEM and highlight the importance of maps as a visual diagnostic for inferring species distributions.

When displaying maps, the inferential target population includes *all* locations in the study area. Therefore, we constructed our maps based on a series of predictions designed to represent this population. We selected 75 000 locations from a GSRD sample designed to cover the entire study area. For each of these locations we had data available from all spatial covariates. Variation in detectability associated with observation effort was controlled by assuming that all three effort predictors (search time, transect length, and time of day) were additively associated with the true occurrence probability and then holding their values fixed for all map predictions.

Using this procedure, we estimated the daily probability of observing Tree Swallows for an eBird participant that searches from 07:00 to 08:00 while traveling 1 km. The daily seasonal occurrence maps for Tree Swallow are shown in Fig. 6. The left column shows the estimates based on the largest scale STEM and the right column shows the corresponding estimates based on the BDT. We chose to display *daily* maps to demonstrate the fine temporal resolution of the eBird data and the models. The seasonal maps were arranged by row: (A) winter (24 January 2006), (B) spring migration (10 April 2006), (C) breeding season (24 June 2006), and (D) the fall migration (24 October 2006).

The STEM maps matched known large-scale patterns in Tree Swallow distributions. Some interesting features of the STEM maps included the limited predicted occurrences through the intermountain West and high plains regions of Montana and Wyoming. The relatively high regional occurrence shown in fall along the Mississippi River valley and along the Atlantic coast were especially interesting. The Mississippi is a plausible migration route with high food availability this time of year and Tree Swallows are known to over-winter and migrate along the Atlantic coast.

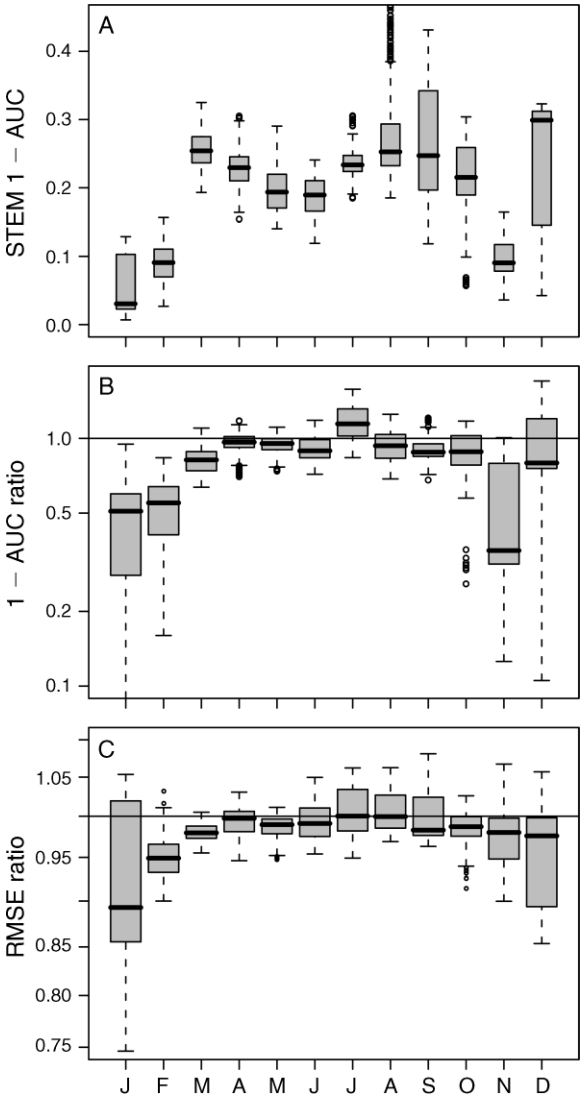

FIG. 5. Predictive performance of monthly Tree Swallow distribution models. (A) The 1 − AUC distributions for the best-scale STEM arranged by month. (B) The 1 − AUC and (C) RMSE ratios on the log scale arranged by month. Each monthly distribution test was conducted during the central two-week window for that calendar month. Averaging across all months, STEM outperformed BDT on 77% of the AUC and 71% of the RMSE test trials.

While the BDT maps show the overall pattern of migration, there were several regional artifacts. These included some of the "false positive" regions during the winter, though these were less pronounced on 24 January compared to the 9 January map shown in Fig.

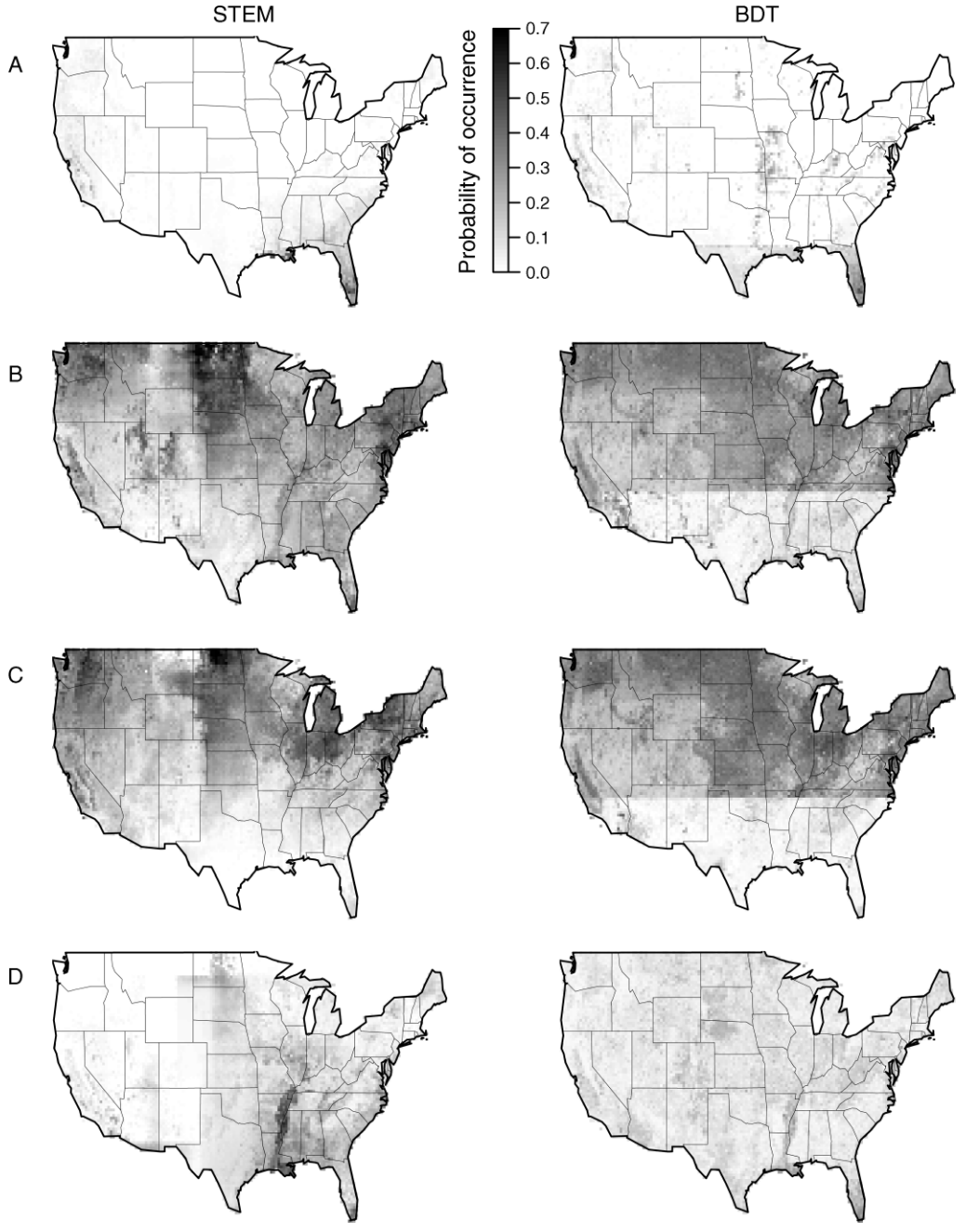

FIG. 6. Single-day estimates of the distribution for Tree Swallow in four seasons. The left column shows the predicted probability of reported occurrence for Tree Swallows based on the largest scale STEM, and the right column shows the corresponding predictions based on the global BDT model. The distributions' dates are arranged by row: (A) winter (24 January 2006), (B) spring migration (10 April 2006), (C) breeding season (24 June 2006), and (D) the fall migration (24 October 2006).

2. The BDT maps for both spring and breeding seasons had a sharp linear southern boundary resulting from DT splits on latitude that were longitudinally extrapolated. The BDT fall map (Fig. 6, row D) failed to predict any of the migrants still known to be remaining in California and the Northeast and revealed in the STEM map. It also retains some false positive predictions in northern Minnesota and the Dakotas at a time when this species is known to be long gone from these regions. Though the STEM structure protects against "long range" learning, it is still possible to produce biased local predictions, see the false positives in the Dakotas on the fall STEM map. Local-scale bias depends on the specific base model and the local predictors.

Northern Cardinal seasonal occurrence performance and maps

While STEMs do a better job at predicting distributions of species whose ranges change through time, the STEM structure will be more generally useful if it can also deal well with species whose distributions do not vary. When a species' distribution does not vary in space or time, data can be pooled across the entire extent of the study to maximize sample size and predictive performance. Thus BDTs built from an ensemble of "global" base models may have an advantage over the STEM built from an ensemble of base models each with restricted support and smaller sample sizes. To test this we compared model performance for a nonmigratory bird, the Northern Cardinal (*Cardinalis cardinalis*). The Northern Cardinal is found throughout eastern and central North America. It is a year-round resident and is common at backyard feeders. It is found in environments with trees and small shrubs throughout its range (Halkin and Linville 1999).

Fig. 7A shows the 1 − AUC distributions for the best-scale STEM arranged by month. Fig. 7B, C shows the 1 − AUC and RMSE ratios, respectively, on the log-scale arranged by month. We used the same *y*-axes for all three panels as in Fig. 5 to facilitate comparison with the Tree Swallow. Compared to the Tree Swallow, there was relatively little month-to-month variation in STEM and BDT performances, as was expected for a nonmigratory species. Moreover, the predictive performance of STEM and BDT were similar to each other indicating that the STEM did not suffer much, if at all, from being unable to pool information from across the entire year. Averaging across all months, the STEMs outperformed BDTs on 45% of the AUC and 48% of the RMSE test trials, also suggesting comparable predictive performance.

While numerically the performance of STEM and BDT models was similar (Fig. 7), visual inspection of the predicted distributions (Fig. 8) suggests that the STEM produced more accurate predictions at the edge of the cardinal's distribution. Both models over-predicted the distribution of cardinals in the Dakotas, but the STEM better described the southwestern U.S. distribution of the species, indicating that range edges are better predicted when information is not globally shared. Interestingly, the western edge of the Midwestern distribution is shown as a straight line in the BDT maps, as these models identified a longitudinal predictor that was effective in an area of high data density and extrapolated it northward; in contrast, the same boundary from the STEM distributions was more diffuse, showing slightly lower probability of occurrence toward the northern part of the eastern range limit. Also, the BDTs shared a problem across seasons of predicting considerable presence of this species in the northern Great Basin where it is absent or extremely rare.

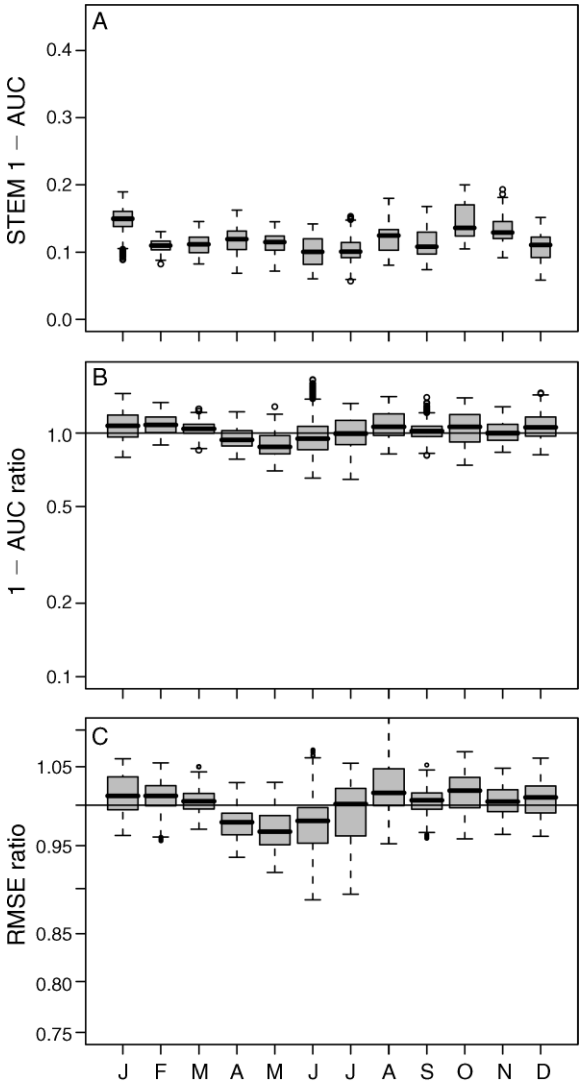

FIG. 7. Predictive performance of monthly Northern Cardinal distribution models. (A) The 1 − AUC distributions for the best-scale STEM arranged by month. (B) The 1 − AUC and (C) RMSE ratios on the log scale arranged by month. Each monthly distribution test was conducted during the central two-week window for that calendar month. Averaging across all months, STEM outperformed BDT on 45% of the AUC and 48% of the RMSE test trials.

DISCUSSION

We have developed a semiparametric model that provides a rigorous and flexible framework for modeling dynamic patterns of species occurrence and abundance from broad-scale survey data. The STEM adds essential spatiotemporal structure to existing species distribution models through a simple parametric structure without requiring a detailed understanding of the underlying dynamic processes. Instead, the approach we take with the STEM is to differentiate between local- and global-scale spatiotemporal structure. At the local scale, we rely on a species distribution model to account for spatio-

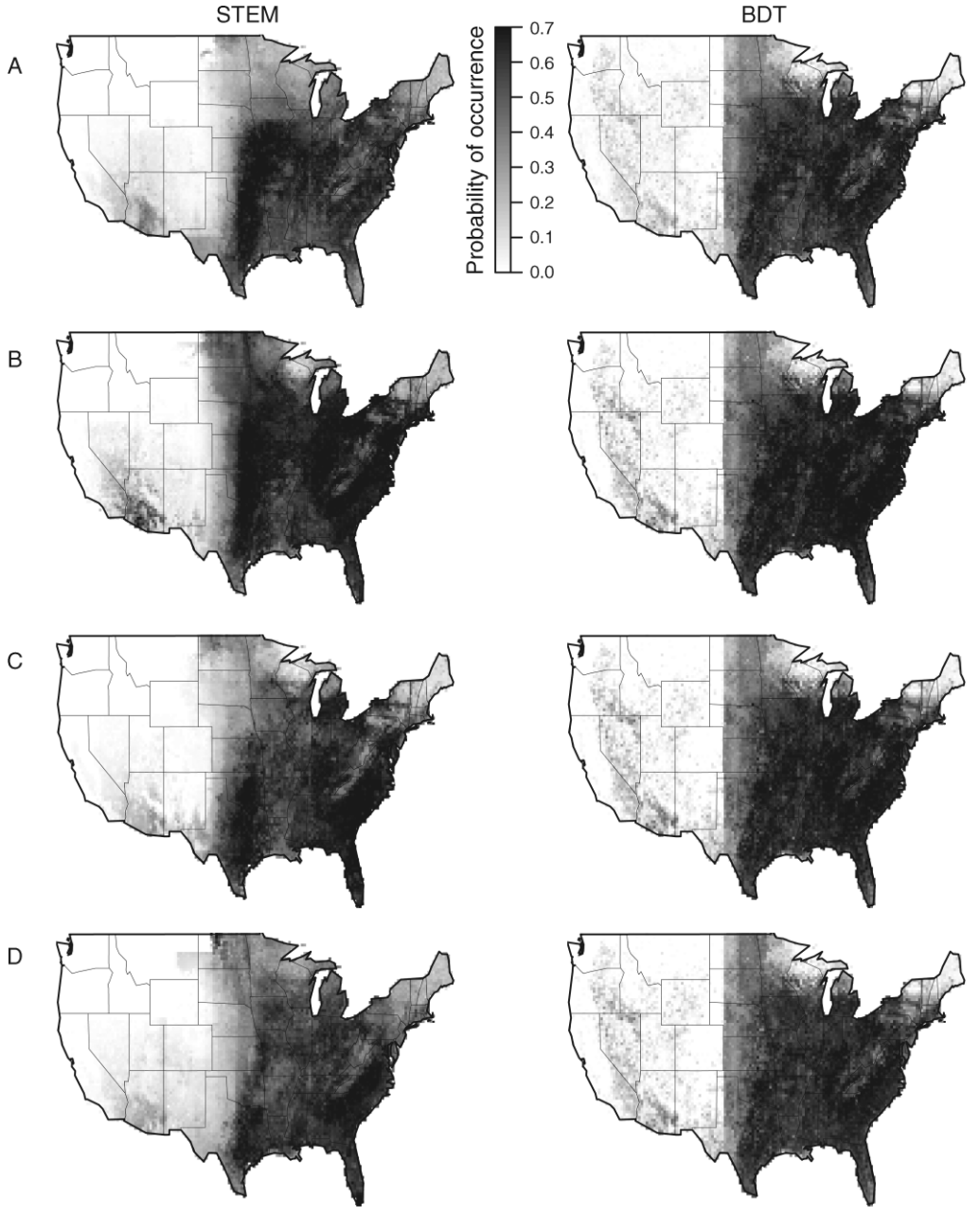

FIG. 8. Single-day estimates of distribution for Northern Cardinal distribution models. The left column shows the predicted probability of reported occurrence for Northern Cardinal based on the largest scale STEM, and the right column shows the corresponding estimates based on the global BDT model. The distributions' dates are arranged by row: (A) winter (24 January 2006), (B) spring migration (10 April 2006), (C) breeding season (24 June 2006), and (D) the fall migration (24 October 2006).

temporal patterning and we let these patterns scale up via ensemble averaging to larger scales. This makes STEMs especially well suited for exploring distributional dynamics arising from a variety of dynamic processes.

The STEM developed in this paper was designed for a specific task: to create a nearly nonparametric dynamic species distribution model for broad-scale intra-annual migrations. More generally, the STEM framework can be customized for a wide range of applications that vary from very exploratory objectives, like the one discussed here, to more confirmatory analysis objectives. There are three key user specifications necessary to implement a STEM: (1) the sizes of the local spatial and temporal scales, (2) the base model type, and (3) the predictors. We briefly discuss theses specifications and note how to adapt STEM to new applications.

We demonstrated that predictive performance varies with spatial scale and across temporal windows (Figs. 3

and 5). In general our experience suggests that the optimal spatial scale varies by season and species. Though we did not address the question of temporal scale directly, we also saw that predictive performance varied by month for the migratory Tree Swallow. Thus, "scale" parameters convey information about the population and it should be possible to fit these parameters to data in future studies, ultimately improving predictive performance and permitting formal inference. Optimal spatial and temporal scale parameters describe the dominant scales at which observations tend to exhibit homogenous phenological patterns and/or predictor associations. Ecologists can use this information to determine the effective range or extent of inference over which small scale studies can be applied. This information may also be useful for determining the "scale" for designing interconnected reserves systems to support migratory species or species experiencing distributional dynamics due to other causes.

Many new and powerful nonparameteric methods have been applied with success to static species distribution problems. However, we have found that these methods can be biased when used to model distribution dynamics. These problems arise from the inability of nonparametric models to take into account regions of space or periods of time where environmental features differ in their importance (i.e., spatiotemporal interactions). As a semi-parametric model, this STEM produced more accurate predictions of species distributions than the associated fully nonparametric BDT model. This result is in keeping with our previous work showing that the addition of well-supported structure to an otherwise nonparametric model improves the accuracy of the model (Sorokina et al. 2008, Fink and Hochachka 2009).

For this eBird analysis we specified a general purpose base model along with a relatively generic set of predictors. This STEM produced improved species distribution maps compared to the associated global BDT models for both the migratory and nonmigratory test species. There is real potential to refine this general-purpose STEM for more specific analysis objectives. For example, the STEM could be easily refined to analyze a single species of conservation concern by carefully selecting the predictors to take into account what is currently known about that species.

When STEM is used to predict patterns from imperfectly detected responses it is important to consider the inclusion of predictor information for important ecological and observational processes. "Observational predictors" can be used to control or mitigate sources of observation bias. To demonstrate this we included effort predictors in the eBird analysis and used these predictors to help control for variation in detection rates. However, there are many other known and potential sources of variation in detection rates and observer biases that were not controlled in the analysis presented here. For example, detection rates are known to vary with habitat types and the observer's level of expertise. We believe that the detection and control of additional types of observational bias within nonparametric and semiparametric models will be an important direction for future research.

We plan to evaluate more sophisticated nonparametric base models like boosted regression trees or RuleFit (see Hochachka et al. 2007) in the future. STEMs can also be built around parametric base models to take advantage of local-scale ecological information. For example, the Bayesian hierarchical models of Royle et al. (2007) may be used within a STEM to explicitly estimate detectability and spatial correlation. The STEM could also be fit to occurrence-only data common in museum and herbarium collections by specifying a model designed for occurrence-only data analysis like MaxEnt (Phillips et al. 2006). Additionally, the ensemble itself may be designed to suit other classes of problems. For example, the STEM ensemble can be designed to analyze non-stationary spatial processes, similar in spirit to varying-coefficient models (Hastie and Tibshirani 1993) and geographically weighted regression models (Fotheringham et al. 2002). The STEM ensemble can also be used to analyze time series, especially those subjected to stochastic change points.

To fully realize the potential of large-scale spatiotemporal data for species distribution modeling it is important to be able to critically evaluate modeling methodologies. We have taken a step toward this goal and developed a method for evaluating the quality of species distribution surface estimates using observational data with heterogeneous, or nonuniform, spatial density. Spatial heterogeneity is a common source of variation in many broad-scale survey data, especially citizen science projects where volunteers decide where and when to make observations. The novelty of our evaluation method is that it explicitly accounts for the spatiotemporal variation in sampling intensity to evaluate the model's ability to make accurate predictions uniformly across the study area.

In conclusion, we see the STEM as a useful addition to the analytical tools available to ecologists. As an important tool for exploratory analysis of dynamic distributions, we see the STEM as a good choice for three general applications. First, when the goal is simply to produce accurate predictions of dynamic distributions. The STEM developed here can exploit large quantities of predictor information with a minimum of user input. Second, this STEM can be used for model-based explorations of data where the goal is to generate hypotheses for further research. This will be especially useful for the initial analysis of populations believed to be simultaneously experiencing several dynamic processes; for example, migration dynamics and range expansion or dispersals. Finally, the STEM can be used as a simple performance benchmark for more detailed parametric analyses of distributional dynamics.

Acknowledgments

This work was funded by the Leon Levy Foundation and the National Science Foundation (grants ITR-0427914 to W. M. Hochachka, S. Kelling, M. Riedewald, and D. Fink; DBI-0542868 to S. Kelling and D. Fink; IIS-0612031 to W. M. Hochachka, S. Kelling, M. Riedewald, and D. Fink; IIS-0920869 to M. Riedewald and D. Fink; CISE-0832782 to D. Fink and S. Kelling; and DEB-0717021 to D. W. Winkler). The authors thank Ken Rosenberg, Rebecca Hutchinson, and the anonymous reviewers for comments on the manuscript. We are grateful to the eBird participants and the Cornell Laboratory of Ornithology Information Sciences unit for providing the data used here, especially Kevin Webb, Tom Fredericks, and Tim Levatich.

Literature Cited

Anselin, L. 1995. Local indicator of spatial association: LISA. Geographical Analysis 27:93–115.

Banerjee, S., B. P. Carlin, and A. E. Gelfand. 2004. Hierarchical modeling and analysis for spatial data. Chapman and Hall/CRC, New York, New York, USA.

Beever, E. A., R. K. Swihart, and B. T. Bestelmeyer. 2006. Linking the concept of scale to studies of biological diversity: evolving approaches and tools. Diversity and Distributions 12:229–235.

Breiman, L. 1996. Bagging predictors. Machine Learning 24: 123–140.

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

Breiman, L., J. H. Friedman, R. A. Olshen, and J. C. Stone. 1984. Classification and regression trees. Chapman and Hall, New York, New York, USA.

Brommer, J. E. 2004. The range margins of northern birds shift poleward. Annales Zoologici Fennici 41:391–397.

Brown, J. H., D. W. Mehlman, and G. C. Stevens. 1995. Spatial variation in abundance. Ecology 76:2028–2043.

Caruana, R., and A. Niculescu-Mizil. 2006. An empirical comparison of supervised learning algorithms. Pages 161–168 *in* W. W. Cohen and A. Moore, editors. Proceedings of the 23rd International Conference on Machine Learning. ACM Press, New York, New York, USA.

De'ath, G. 2007. Boosted trees for ecological modeling and prediction. Ecology 88:243–251.

De'ath, G., and K. E. Fabricius. 2000. Classification and regression trees: a powerful yet simple technique for ecological data analysis. Ecology 81:3178–3192.

Dhondt, A. A., et al. 2005. Dynamics of a novel pathogen in an avian host: mycoplasmal conjunctivitis in House Finches. Acta Tropica 94:77–93.

Diefenbach, D. R., M. R. Marshall, J. A. Mattice, and D. W. Brauning. 2007. Incorporating availability for detection in estimates of bird abundance. Auk 124:96–106.

Dungan, J. L., J. N. Perry, M. R. T. Dale, P. Legendre, S. Citron-Pousty, M.-J. Fortin, A. Jakomulska, M. Miriti, and M. S. Rosenberg. 2002. A balanced view of scale in spatial statistical analysis. Ecography 25:626–640.

Elith, J., and C. H. Graham. 2009. Do they? How do they? WHY do they differ? On finding reasons for differing performances of species distribution models. Ecography 32: 66–77.

Elith, J., et al. 2006. Novel methods improve prediction of species' distributions from occurrence data. Ecography 29: 129–151.

Elith, J., J. R. Leathwick, and T. Hastie. 2008. A working guide to boosted regression trees. Journal of Animal Ecology 77: 802–813.

Fielding, A. H., and J. F. Bell. 1997. A review of methods for the assessment of prediction errors in conservation presence–absence models. Environmental Conservation 24:38–49.

Fink, D., and W. M. Hochachka. 2009. Gaussian semi-parametric analysis using hierarchical predictive models. Environmental and Ecological Statistics 3:1011–1035.

Fortin, M. J., and M. Dale. 2005. Spatial analysis: a guide for ecologists. Cambridge University Press, Cambridge, UK.

Fotheringham, A. S., C. Brunsdon, and M. Charlton. 2002. Geographically weighted regression: the analysis of spatially varying relationships. John Wiley and Sons, New York, New York, USA.

Freund, Y., and R. Schapire. 1996. Experiments with a new boosting algorithm. Pages 148–156 *in* Proceedings of the 13th International Conference on Machine Learning. Morgan Kaufmann, San Francisco, California, USA.

Friedman, J. H. 2001. Greedy function approximation: a gradient boosting machine. Annals of Statistics 29:1189–1232.

Friedman, J. H., and B. E. Popescu. 2008. Predictive learning via rule ensembles. Annals of Applied Statistics 2:916–954.

Greenberg, R., and P. P. Marra. 2005. Birds of two worlds: the ecology and evolution of migration. Johns Hopkins University Press, Baltimore, Maryland, USA.

Groom, M. J. 1998. Allee effects limit population viability of an annual plant. American Naturalist 151:487–496.

Grosbois, V., O. Gimenez, J. M. Gaillard, R. Pradel, C. Barbraud, J. Clobert, A. P. Møller, and H. Weimerskirch. 2008. Assessing the impact of climate variation on survival in vertebrate populations. Biological Reviews 83:357–399.

Halkin, S. L., and S. U. Linville. 1999. Northern Cardinal (*Cardinalis cardinalis*). *In* A. Poole, editor. The birds of North America online. Cornell Lab of Ornithology, Ithaca, New York, USA. ⟨http://bna.birds.cornell.edu/bna/species/440⟩

Harrell, F. E., Jr. 2001. Regression modeling strategies with applications to linear models, logistic regression and survival analysis. Springer Verlag, New York, New York, USA.

Hastie, T., and R. Tibshirani. 1993. Varying-coefficient models. Journal of the Royal Statistical Society, Series B (Methodological) 55:757.

Hastie, T., R. Tibshirani, and J. Friedman. 2001. The elements of statistical learning: data mining, inference, and prediction. Springer Verlag, New York, New York, USA.

Hochachka, W. M., R. Caruana, D. Fink, A. Munson, M. Riedewald, D. Sorokina, and S. Kelling. 2007. Data mining for discovery of pattern and process in ecological systems. Journal of Wildlife Management 71:2427–2437.

Homer, C., J. Dewitz, J. Fry, M. Coan, N. Hossain, C. Larson, N. Herold, A. McKerrow, J. N. VanDriel, and J. Wickham. 2007. Completion of the 2001 National Land Cover Database for the conterminous United States. Photogrammetric Engineering and Remote Sensing 73:337–341.

Hooten, M. B., and C. K. Wikle. 2007. Shifts in the spatio-temporal growth dynamics of shortleaf pine. Environmental and Ecological Statistics 14(3):207–227.

Hurlbert, S. H. 1984. Pseudoreplication and the design of ecological field experiments. Ecological Monographs 54:187–211.

Keitt, T. H., M. A. Lewis, and R. D. Holt. 2001. Allee effects, invasion inning, and species' borders. American Naturalist 157:203–216.

Kelling, S., W. M. Hochachka, D. Fink, M. Riedewald, R. Caruana, G. Ballard, and G. Hooker. 2009. Data-intensive science: a new paradigm for biodiversity studies. BioScience 59:613–620.

Latimer, A. M., S. Wu, A. E. Gelfand, and J. A. Silander. 2006. Building statistical models to analyze species distributions. Ecological Applications 16:33–50.

Levin, S. A. 1992. The problem of pattern and scale in ecology. Ecology 73:1943–1967.

Lichstein, J. W., T. R. Simons, S. A. Shriner, and K. E. Franzreb. 2002. Spatial autocorrelation and autoregressive models in ecology. Ecological Monographs 72:445–463.

Lima, M., N. C. Stenseth, and F. M. Jaksic. 2002. Food web structure and climate effects on the dynamics of small mammals and owls in semi-arid Chile. Ecology Letters 5:273–284.

Link, W. A., and J. R. Sauer. 2007. Seasonal components of avian population change: joint analysis of two large-scale monitoring programs. Ecology 88:49–55.

MacLean, I. M., G. E. Austin, M. M. Rehfisch, J. Blew, O. Crowe, S. Delany, K. Devos, B. Deceuninck, K. Günther, K. Laursen, M. Van Roomen, and J. Wahl. 2008. Climate change causes rapid changes in the distribution and site abundance of birds in winter. Global Change Biology 14(11): 2489–2500.

Maggini, R., A. Lehmann, N. E. Zimmermann, and A. Guisan. 2006. Improving generalized regression analysis for the spatial prediction of forest communities. Journal of Biogeography 33:1729–1749.

Mouquet, N., and M. Loreau. 2003. Community patterns in source–sink metacommunities. American Naturalist 162: 544–557.

Phillips, S. J., R. P. Anderson, and R. E. Schapire. 2006. Maximum entropy modeling of species geographic distributions. Ecological Modelling 190:231–259.

Prasad, A. M., L. R. Iverson, and A. Liaw. 2006. Newer classification and regression tree techniques: bagging and random forests for ecological prediction. Ecosystems 9:181–199.

Quinlan, J. R. 1993. C4.5: programs for machine learning. Morgan Kaufmann Publishers, San Mateo, California, USA.

R Development Core Team. 2008. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ⟨http://www.R-project.org⟩

Reddy, S., and L. M. Davalos. 2003. Geographic sampling bias and its implications for conservation priorities in Africa. Journal of Biogeography 30:1719–1727.

Ridgeway, G. 2007. Generalized boosted regression models. Documentation on the R package "gbm," version 1.5-7. ⟨http://cran.r-project.org/web/packages/gbm/vignettes/gbm.pdf⟩

Robertson, R. J., B. J. Stutchbury, and R. R. Cohen. 1992. Tree Swallow (*Tachycineta bicolor*). *In* A. Poole, editor. The birds of North America online. Cornell Lab of Ornithology, Ithaca, New York, USA. ⟨http://bna.birds.cornell.edu/bna/species/011⟩

Royle, J. A., M. Kéry, R. Gautier, and H. Schmid. 2007. Hierarchical spatial models of abundance and occurrence from imperfect survey data. Ecological Monographs 77:465–481.

Royle, J. A., M. Koneff, and R. Reynolds. 2002. Spatial modeling of wetland condition in the U.S. Prairie Pothole region. Biometrics 8(2):104–113.

Schummer, M. L., S. A. Petrie, and R. C. Bailey. 2008. Interaction between macroinvertebrate abundance and habitat use by diving ducks during winter on northeastern Lake Ontario. Journal of Great Lakes Research 34:54–71.

Scott, J. M. 2002. Predicting species occurrences: issues of accuracy and scale. Island Press, Washington, D.C., USA.

Smith, A. T. 1974. Distribution and dispersal of pikas: consequences of insular population structure. Ecology 55: 1112–1119.

Sorokina, D., R. Caruana, M. Riedewald, and D. Fink. 2008. Detecting statistical interactions with additive groves of trees. Pages 1000–1007 *in* W. W. Cohen, A. McCallum, and S. T. Roweis, editors. Machine learning. Proceedings of the Twenty-Fifth International Conference (ICML 2008), Helsinki, Finland, June 5–9, 2008. ACM International Conference Proceeding Series 307 ACM 2008. Association for Computing Machinery, New York, New York, USA.

Strayer, D. L. 2009. Alien species in fresh waters: ecological effects, interactions with other stressors, and prospects for the future. Freshwater Biology 55:152–174.

Sullivan, B., C. Wood, M. J. Iliff, R. E. Bonney, D. Fink, and S. Kelling. 2009. eBird: a citizen-based bird observation network in the biological sciences. Biological Conservation 142:2282–2292.

Therneau, T. M., and B. Atkinson. 2008. rpart: recursive partitioning. R package version 3.1–42. ⟨http://mayoresearch.mayo.edu/mayo/research/biostat/splusfunctions.cfm⟩

Thogmartin, W. E., J. R. Sauer, and M. G. Knutson. 2007. Modeling and mapping abundance of American Woodcock across the midwestern and northeastern United States. Journal of Wildlife Management 71(2):376–382.

Thomas, C. D., and J. J. Lennon. 1999. Birds extend their ranges northwards. Nature 399:213.

Waser, P. M., and L. F. Elliott. 1991. Dispersal and genetic-structure in kangaroo rats. Evolution 45:935–943.

Wiens, T. S., B. C. Dale, M. S. Boyce, and G. P. Kershaw. 2008. Three way k-fold cross-validation of resource selection functions. Ecological Modelling 212:244–255.

Wikle, C. K. 2003. Hierarchical Bayesian models for predicting the spread of ecological processes. Ecology 84:1382–1394.

Winkler, D. 2006. Roosts and migrations of swallows. Hornero 21:85–97.