UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

PREDICTING BREEDING-BIRD OCCURRENCES IN OKLAHOMA: RELATIONSHIP OF SPECIES DISTRIBUTIONS TO LAND-COVER AND CLIMATIC VARIATION

A DISSERTATION

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

degree of

Doctor of Philosophy

By

DENNIS G SIEGFRIED
Norman, Oklahoma
2007

PREDICTING BREEDING-BIRD OCCURRENCES IN OKLAHOMA: RELATIONSHIP OF SPECIES DISTRIBUTIONS TO LAND-COVER AND CLIMATIC VARIATION

A DISSERTATION APPROVED FOR THE
DEPARTMENT OF ZOOLOGY

BY

_____
Gary D. Schnell, Chair

_____
Jeffrey F. Kelly

_____
William J. Matthews

_____
Thomas S. Ray

_____
May Yuan

PREFACE

This dissertation is presented in three chapters to be submitted for publication to refereed journals. Chapter 1 is formatted for the journal *Landscape Ecology*. Chapters 2 and 3 have been prepared to be submitted to *Ecological Modelling*.

ACKNOWLEDGMENTS

I begin by thanking my advisor Gary Schnell. His patience, criticisms, and good nature are without parallel. I also thank Jeffrey F. Kelly, William J. Matthews, Thomas S. Ray, and May Yuan for serving as committee members and for constructive comments on the earlier drafts of this dissertation.

I also thank my wife and son, Mei-ling and Corey. Without them this project would never have been done. Their encouragement, perspective, balance, and love got me through a lot.

My mom and dad, thank you for your guidance, support and love which have made this possible. Thank you for encouraging me, helping me to make the right decisions, and supporting me. Thank you for everything.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF ILLUSTRATIONS

ABSTRACT

Predictive modeling for species distributions has received considerable attention in the past decade. As more data sets and better satellite images become available it is of interest to know how well modeling techniques can use environmental variables in predictive models of species distributions. I evaluated: how well logistic regression develops predictive models for distributions of the 209 species of the Oklahoma breeding-bird-atlas project; the number of sites needed by logistic regression to develop robust models; how well logistic regression and GARP (genetic algorithm rule-set prediction) predicted species distributions; and how well logistic regression and GARP predicted species richness.

Predictive models for species distributions were developed using logistic regression with 13 land-cover and 21 climatic variables. The 209 species models were then applied to the 562 breeding-bird-atlas sites and 12 independent sites surveyed in 2004. Models correctly predicted occurrences 89.4% of the time for the 562 sites and 81.8% for the 12 sites. Greater species richness was found at sites with more land-cover types and was positively associated with principal component 1 for climatic variables.

It is of practical interest to know the amount of data required to produce reliable models. I developed nine training sets of from 50 to 450 sites for each species to predict occurrence in a 100-site test set. Model performance was determined by the reduction in error of prediction (tau-p) compared to the null. On average, model performance leveled off after 300 sites.

It is of interest to evaluate abilities of techniques to predict species occurrences and species richness. I compared logistic regression and two forms of GARP in predicting

individual species distributions. Logistic regression produced more consistent results, while GARP overpredicted presences. When distribution predictions were summed to estimate species richness for a given site, logistic regression was a better predictor of species richness than GARP, but it did not produce particularly reliable relative measures across sites. GARP routinely overpredicted species richness per site and also did not produce reliable relative measures across sites.

Land-cover and climatic influences on distributions: predicting presence and absence of Oklahoma breeding bird species

Dennis G Siegfried • Gary D. Schnell • Dan L. Reinking

D. G Siegfried
Department of Zoology, University of Oklahoma, Norman, Oklahoma, 73019, USA

G. D. Schnell (corresponding author)
Sam Noble Oklahoma Museum of Natural History and Department of Zoology, 2401 Chautauqua Avenue, University of Oklahoma, Norman, Oklahoma, 73072, USA
e-mail: gschnell@ou.edu
phone: 405-325-5050/fax: 405-325-7699

D. L. Reinking
Sutton Avian Research Center, University of Oklahoma, P.O. Box 2007, Bartlesville, Oklahoma, 74005, USA

Present Address
D. G Siegfried
Department of Biology, Southern Nazarene University, 6729 NW 39th Expressway, Bethany, Oklahoma, 73008, USA
email: dsiegfri@snu.edu

Date of Manuscript: 3 May, 2007

Manuscript word count (including text, references, tables, and captions): 6,028

Abstract Accurately predicting species distributions is a goal of ecologists. We evaluated the efficacy of logistic regression to predict bird distributions in Oklahoma (112 500 km^2), analyzing occurrence data (presences and absences) for 209 species at 562 survey blocks in the *Oklahoma Breeding Bird Atlas*. We studied the relationship of species richness to the number of land-cover types and to a general climatic factor, and assessed accuracy of logistic-regression models relative to the number of sites where individual species were present. We used 13 land-cover and 21 climatic variables to develop models, which then were employed to predict occurrences in original blocks and 12 newly surveyed blocks. Models averaged 3.2 variables, with climatic variables being incorporated most frequently. Models predicted occurrences well for original blocks (89.4% correct) and supplemental sites (81.8%). Species richness was positively associated with the number of land-cover types and climatic PC1, a composite variable with a west-to-east trend of more rain, higher humidity and pressure, less sunlight, lower wind speeds, and lower soil temperatures. Model performance was best for species occurring at few and at many sites. Presences were more accurately predicted at sites having more land-cover types and higher climatic-PC1 values. Percentages of incorrectly predicted occurrences were inversely related to the number of land-cover types, but not to climatic PC1. Atlas projects provide a baseline of species occurrences and, when considered with environmental variables, useful predictions can be made as to where changes in bird distributions will occur if land-cover types or climatic conditions change.

Keywords: Bird distributions • Distribution maps • Land-cover variables • Climatic variables • Logistic regression • Geographic information systems

Introduction

Knowledge of species distributions is central to management of animal populations. Conservation biologists and planners benefit from being able to identify key environmental parameters that can be used as surrogates to represent distribution of species of interest over a broad region (MacKenzie et al. 2006). Therefore, identifying variables that correlate with species occurrences (i.e. presences and absences; Angermeier et al. 2002) is important for management of large areas (Simberloff 1988). Land-cover types represent a suite of possible variables that can be accessed using satellite imagery. Detailed data for other environmental factors also are readily available. Climatic variables, particularly precipitation and temperature measures, may be helpful when attempting to predict species occurrences.

Several analytical techniques can be used to develop prediction models of bird-species distributions based on environmental measures. One approach, GARP (genetic algorithm rule-set predictions; Stockwell 1998), typically has been applied to large presence-only samples generated from museum specimens. This method has been used to predict bird distributions in both Mexico and North America (Peterson 2001; Peterson and Vieglais 2001; Peterson et al. 2002; Peterson and Kluza 2003; Peterson and Robins 2003), as well as those of mammals in North America (Peterson et al. 2002; Peterson et al. 2006) and South America (Anderson et al. 2002; Lim et al. 2002) and reptiles of Madagascar (Raxworthy et al. 2003). Discriminant analysis, correspondence analysis, artificial neural networks, general linear models, and logistic regression have been used with occurrence data (Manel et al. 1999; van Horseen et al. 1999; Brotons et al. 2004; Engler et al. 2004; Venier et al. 2004). Comparisons of predictive abilities of procedures for individual bird

species have suggested, not surprisingly, that occurrence data when available are better predictors of occurrence than presence-only data (Brotons et al. 2004). In predicting species occurrences, logistic regression has been shown to outperform discriminant analysis and artificial neural networks in correctly predicting distributions of six bird species associated with mountain streams (Manel et al. 2004).

Breeding-bird-atlas projects are a valuable source of occurrence data. One of the first projects was completed in the United Kingdom in 1972 (Cox 2006). In North America, such projects were initiated in the 1970s in northeastern United States (Bevier 1994; Brewer et al. 1991) and now have been completed throughout much of the USA and Canada. While numerous breeding-bird-atlas projects have been produced, few quantitative studies (e.g. van Rensburg et al. 2002; Venier et al. 2004; Cox 2006) have been published using such data.

The use of logistic-regression models in conjunction with climatic and land-cover data to predict occurrences of species has become widespread (Murtaugh 1996; van Horseen et al. 1999; Rahbeck and Graves 2001; van Rensburg et al. 2001; Koleff and Gaston 2002; Manel et al. 2004; Venier et al. 2004; Gilbert et al. 2005). Logistic regression is well suited to developing these models as they meet the assumption of a dichotomous dependent variable in the form of present (1) or absent (0), and continuous and/or categorical independent variables. Evaluating regression models usually has been done using one part of the existing data set of occurrences and variable data to predict occurrences in the remaining sites within the data set. Regression models from the original data set also have been applied to new blocks that were not initially surveyed (Pearce and Ferrier 2000). It is also of interest to know how well models actually predict

the values for sites on which they are based, an approach using intrinsic criteria to judge accuracy (Anderson et al. 2003).

Several studies have looked at the efficacy of climatic and land-cover variables to predict species distributions, focusing on either areas of limited extent (i.e. <1000 km^2; Austin et al. 1996*;* van Horseen et al. 1999; Villard et al. 1999; Bustamante and Seoane 2004; Gibson et al. 2004) or on only a few species at a large number of sites (i.e. <10 species; Titeux et al. 2004; Venier et al. 2004). We have evaluated 209 bird distributions in a considerably larger area (i.e. Oklahoma; 112 500 km^2) based on data from the *Oklahoma Breeding Bird Atlas* (Reinking 2004).

Our purposes were two-fold. First, we were interested in determining how species richness was related to the number of land-cover types in a block, as well as to a general climatic factor. Second, we evaluated the efficiency in predicting the occurrences of 209 breeding-bird species at a large number of sites in Oklahoma using models developed from climatic and land-cover variables. We assessed how accuracy was related to the number of land-cover types and a general climatic factor. The results for Oklahoma potentially have broader significance in providing insights as to the predictive success of logistic regression as applied to other data sets, including those from other breeding-bird-atlas programs, as well as from other studies that generate occurrence information.

Methods

Presence/absence bird data

We analyzed occurrence information (both presences and absences) for 209 bird species from the *Oklahoma Breeding Bird Atlas*, including data for 562 complete atlas blocks (Fig. 1), each of equal size (25 km^2), in a species-incidence matrix of occurrence (MacKenzie et al. 2006). Initial block selection for the atlas was based on a stratified-random procedure that ensured broad coverage of the state (Reinking 2004). Over a 5-year period (1997-2001), observers visited each of the 562 blocks during the breeding season, recording all bird species encountered.

Land-cover variables

We defined 13 land-cover variables, 12 of which were based on an initial 19 types found in Oklahoma and described for the 1992 land-cover image in the United States Geological Survey archive (USGS 2002). Four residential types were merged into a single variable, as we did for three cultivated types and for three bare-ground types. Map polygons then were defined as 1 of 12 land-cover variables: (1) deciduous forest; (2) mixed forest; (3) evergreen forest; (4) woody wetland; (5) emergent herbaceous wetland; (6) shrubland; (7) grassland; (8) pasture/hay; (9) cropland; (10) developed; (11) barren; and (12) water. Within each atlas block we determined the proportion of the block covered by each land-cover type. Variable 13 was the distance to water (in meters), calculated as the distance from the center of the block to river or stream as identified in the Oklahoma Digital Elevation Model hydrological network derived from the 1:100,000-scale digital topographic map (USGS 1998). We obtained measurements for this variable

using the Distance Between Points (between layers) procedure of Analyst Tools in ArcGIS (Beyer 2004). All ArcGIS layers were placed in Universal Transverse Mercator projection, zone 14, based on the North American Datum 1983. In addition to using land-cover data in model development, species richnesses were regressed on the number of land-cover types in blocks.

Climatic variables

The climatic data were from the Oklahoma Mesonet provided by the Oklahoma Climatological Service at the University of Oklahoma, which maintains a network of 119 weather stations placed throughout Oklahoma (Brock et al. 1995). The data were for 1997 through 2001, covering May, June, and July, corresponding to the years and months when the birds were surveyed. The five years were averaged for each month to generate point data for each station. For each of the three months (May, June, and July), seven variables were selected: temperature (^{o}C); soil temperature (^{o}C); rainfall (cm); solar radiation (megajoules/m^2); wind speed (kph); barometric pressure (millibars); and humidity (%). The 21 resulting climatic variables (numbered 14-34 in Table 1) were interpolated to estimate values for all locations in the state using the ordinary kriging method (van Horseen et al. 1999) in the Geostatistical Analyst of ArcGIS (ESRI 2004).

The values of the 21 climatic variables for the 562 atlas blocks were analyzed using NTSYSpc (Rohlf 2002) to determine the first principal-component axis (climatic PC1), which represents a general climatic gradient for the state of Oklahoma. Projections of blocks on this component were based on standardized data, where each variable had a

mean of 0 and standard deviation of 1 for the 562 blocks. Species richness was regressed on climatic-PC1 projections.

Individual species models

Bird occurrence data were analyzed based on the 34 land-cover and climatic variables in a stepwise logistic regression (Hosmer and Lemeshow 2000) to determine the best subset of variables that statistically explained presences and absences of a given species. Computations were done in SAS 9 (SAS 2004) with a 0.05 *p*-value for significance. Regression equations for each species gave a response value for each atlas block that was either positive (present) or negative (absent). These predicted occurrences were used to generate predicted distribution maps in ArcGIS.

Model performance

The atlas occurrence data for a given species were compared to the data on predicted occurrences to determine model performance based on the number and percent of correctly predicted occurrences, presences, and absences. We used tau-p (τ_p) as a measure of the proportional change in error rate of the model compared to the expected rate of error without the model for each species. The value of τ_p can vary from 1 to $1 - [N^2/2(N-1)]$, where N is the total number of sites. When τ_p is negative there is proportionally greater error for the model, indicating the model performs worse than chance. A positive τ_p denotes a proportional reduction in error for the model as compared to chance, indicating greater accuracy (Menard 2002).

Additional sites

We further evaluated the predictive performance of the individual regression equations by selecting 12 independent sites for survey during the 2004 breeding season. Since we wanted to prevent overlap or shared boundaries with existing sites, potential sites were identified using the empty outlines of counties and atlas blocks. From these potential sites, 12 were selected, two from each of six counties in central Oklahoma: Canadian, Cleveland, Grady, McClain, Oklahoma and Pottawatomie counties.

Surveys were conducted using the same procedures employed by those recording data for the *Oklahoma Breeding Bird Atlas* as set by the *Oklahoma Breeding Bird Atlas Handbook* (Reinking 2000). From May through July 2004, the first author visited each of the 12 sites once every two weeks. Sites were surveyed for an average of 22.5 hours (range 18 – 28). Species were identified by sight and/or song. After the surveys, we used the logistic-regression models that had been generated for individual species based on atlas data to predict the occurrences of a species in the new blocks. We then compared the surveyed presences and absences to those predicted by the model in the same manner as described above.

Results

Atlas block associations to land-cover types and climatic PC1

Atlas blocks enclosed a variety of the possible land-cover types, with an average of 9.0 land-cover types per block (SD = 1.8; Fig. 1a). The atlas block with the fewest number of land-cover types (3) occurred in the Panhandle region of the state, with the average for the Panhandle being 5.8. The maximum number of types (13) was in one block in south-central Oklahoma along the Red River; relatively high numbers of types also occurred in northeastern Oklahoma. Except along the Red River, which constitutes the southern border with Texas, the southeastern corner of the state was also relatively lower in land-cover types ($\bar{x}$ = 7.1) than the rest of the main body of the state ($\bar{x}$ = 10.0).

For the 562 atlas blocks, observers recorded from 24 to 100 bird species per block ($\bar{x}$ = 57.8, SD = 12.5). Regression of species richness in a block on the number of land-cover types was positive ($p < 0.05$; Fig. 2a). For each additional land-cover type there were, on average, 2.42 more species.

The climatic PC1 had both positive and negative associations with climatic variables (Table 1), explaining 54.9% of the variation in climatic variables. For this component, 14 of the 21 climatic variables had relatively high loadings. Positively associated variables were rain (May and June), pressure (May, June, and July), and humidity (June and July); values for these variables increased from west to east. Negatively associated variables were solar radiation (May, June, and July), wind speed (May, June and July), and soil temperature at 10 cm (June), all of which generally decreased from west to east. Climatic-PC1 projections, which went from negative values in the west to positive values

in the east (Fig. 1b), show a significant association with species richness ($p < 0.05$; Fig. 2b).

Model performance

The 209 species were recorded as present in from 1 to 556 of the 562 atlas blocks, being present on average in 152.8 blocks per species (SD = 175.3). The logistic-regression equations for given species included those with: (1) only a constant (28 of 209 species); (2) a constant and land-cover variables (15 species); (3) a constant and climatic variables (36 species); and (4) a constant and some combination of land-cover and climatic variables (130 species; Appendix 1). On average, equations for individual species had 3.2 variables, with a range of 0 to 10. For 33 species, equations included from 5 to 9 variables. Only the equation for the painted bunting had 10 variables.

The individual models showed a range of variability in performance, as indicated by the percent correctly predicted occurrences for individual species (Fig. 3a). Models for species that had either widespread distributions (present at ≥ 506 of the 562 sites [90%]) or sparse distributions (present at ≤ 56 sites [10%]) had correctly predicted occurrences for most of the blocks ($\bar{x}$ = 97.3%, range 91.1 – 99.8%; Fig. 3a). Models for species that were present in 57 to 505 sites generally did less well at predicting occurrences, and there was notable scatter ($\bar{x}$ = 89.4%, range 56.1 – 98.8% correct; Fig. 3a). Models for this middle group of species generally included more variables than average.

Model performance in terms of correctly predicting presences showed high variability for species that occurred in fewer than 250 sites (left part of Fig. 3b); as actual presences increased above 250 sites, the percentage of correctly predicted presences approached

100% (right half of Fig. 3b). The proportional reduction in error rate (τ_p) for models was also variable, with the lowest values being associated with the species that were present at about one-half of the sites ($\bar{\tau}_p = 0.504$, range 0.09 – 0.97; Fig. 3c). Two curves of τ_p have been superimposed on Fig. 3c, one for the "models" where it is estimated that the species is present everywhere (solid line) and the other for when the species is judged to be absent from all sites (dashed line). In all cases, logistic-regression models resulted in τ_p-values that were higher than for the present-everywhere model. There are a few instances where the τ_p from logistic regression is less than that obtained using the absent-everywhere model. Given that all τ_p-values were positive, all species equations were more accurate in correctly predicting presences and absences than expected by chance alone, a not unexpected result given that predictions were for the same sites on which the models were based.

Correctly predicted presences had a positive association with the number of land-cover types, while correctly predicted absences showed a negative association (both $p < 0.001$; Fig. 4a and b); thus, with more land-cover types in a block there were more correctly predicted presences and fewer correctly predicted absences. Percentages of correctly predicted as presences and absences showed the same patterns (both $p < 0.001$; Fig. 4e and f). Incorrectly predicted presences and absences had the opposite patterns (both $p < 0.001$; Fig. 4c and d), showing that fewer species were incorrectly predicted present when there were more land-cover types in a block. However, when there were more land-cover types, there also were more incorrectly predicted absences.

When presences and absences of individual species were compared separately to climatic PC1, correctly predicted presences had a positive regression slope with climatic

PC1 ($p < 0.001$; Fig. 5a), while the number of correctly predicted absences had a negative regression slope ($p < 0.001$; Fig. 5b). As climatic PC1 went from -2.0 to 1.5, corresponding to a west-to-east trend, more species were correctly predicted present and fewer species were correctly predicted as absent. Regressions of percentages of the correctly predicted presences and absences on climatic PC1 were not significant ($p = 0.102$ and 0.088, respectively; Fig. 5e and f). Thus, even though more species were found at sites with positive values for climatic PC1, the relative number of correctly predicted presences and absences did not change (Fig. 5e and f). As climatic PC1 went from negative to positive, there was a significant increase in the number of incorrectly predicted presences ($p < 0.01$; Fig. 5c). The number of incorrectly predicted absences, however, was not significantly associated with climatic PC1 ($p = 0.717$; Fig. 5d).

Supplemental sites from 2004

When applied to the 12 blocks surveyed in 2004, models for the 111 species that were not found in the blocks correctly predicted their absences in all 12 blocks. Of the remaining 98 species that were found in one or more of these blocks, their respective individual models on average correctly predicted occurrences 81.8% of the time. Considering presences and absences separately, models correctly predicted presences 71.5% of the time and absences 85.3%. When a species occurred in only 1 of the 12 blocks, the model incorrectly predicted the species to be present in several other blocks as well. For the 49 species that occurred in 6 or more of the 12 sites, models of individual species incorrectly predicted that the species were present in all 12 locations.

Discussion

Our findings parallel those of Venier et al. (2004), who found that climatic variables, when compared with land-cover variables, were more often included as predictors of avian species occurrence. They used large-scale climate variables and satellite-derived forest-cover values to predict distributions of 10 forest birds. In our study, more models included only climatic variables (36 species) than only land-cover variables (15 species). We did have more climatic than land-cover variables that could have been included (21 vs. 13 variables), but this likely does not account fully for the discrepancy. Climatic variables, in general, showed clinal variation, while land-cover types often exhibited more abrupt changes between adjacent sites.

From the 36 species models with only climatic variables, two patterns emerged. The first was similar to that found by Lawler et al. (2004), where only climatic variables were needed to develop models in areas where land-cover richness was relatively low and did not add to the explanatory power of the climatic variables. For the 36 species with models that included only climatic variables, 23 were found only in the western half of the state, with several restricted to the Panhandle region during the breeding season (e.g. black-billed magpie [*Pica hudsonia*], gadwall [*Anas strepera*], and lesser prairie-chicken [*Tympanuchus pallidicinctus*]). This region, with 43 sites, had the lowest land-cover richness, being unequally dominated by three types: grasslands (range of proportion 0.136 – 0.965), pasture/hay (0.000 – 0.154), and cropland (0.000 – 0.839). Due to the range of proportional coverage by these land-cover types and their occurrences in other locations east of the Panhandle, models were unable to identify an association that was unique to species that only occurred in the Panhandle with these land-cover types. This

resulted in climatic variables being more instructive in predicting occurrences of species that occur only in this region.

The second pattern was that of 10 species with widespread distributions spanning the entire state (bank swallow [*Riparia riparia*], barred owl [*Strix varia*], blue-winged teal [*Anas discors*], Carolina chickadee [*Poecile carolinensis*], Carolina wren [*Thryothorus ludovicianus*], cedar waxwing [*Bombycilla cedrorum*], Cooper's hawk [*Accipiter cooperii*], great horned owl [*Bubo virginianus*], and northern cardinal [*Cardinalis cardinalis*]); they occurred at many sites that did not share similar land-cover types. The variety of land-cover proportions coupled with the continuous occurrence of the species prevented the logistic regression from identifying land-cover types as having significant predictive power for these species. Rather, climatic variables that showed a clinal variation were more instructive.

Models that performed relatively poorly may have done so, at least in part, because our scale of analysis did not match the scale at which particular species "perceive" the environment. For example, the model for the cattle egret (*Bubulcus ibis*; Appendix 1) had the lowest prediction rating ($\tau_p = 0.09$). On a daily basis this species typically travels distances considerably exceeding the size of an atlas block in search of food (> 20 km; Telfair 1994). Detailed data on block land-cover types probably are not particularly relevant for such highly mobile species. Some models performed exceptionally well (e.g. blue grosbeak, $\tau_p = 0.97$). The blue grosbeak (*Passerina caerulea*) model (Appendix 1) incorporated the deciduous forest land-cover type, which likely was associated with extensive habitat edge, a component of the environment that is important for the species (Ingold 1993). The other variables in the blue grosbeak model — mixed forest, cropland,

developed, and July rain — were negatively associated; the species tended not to be found when there was a substantial portion of mixed forest, where much of the land was cropland or developed, and where July rainfall was relatively meager.

To date, more than 40 breeding-bird atlases have been completed in North America. One aim is to repeat the surveys every 10-15 years (Peterjohn and Rice 1991; Bevier 1994), which would allow comparisons and testing of predictions by models developed with the original data. With no change in environmental conditions, the best prediction is that species will continue to be found at the same sites as in the initial survey. However, climatic changes do occur. For example, from 1910 to 1995, the south-central United States (Texas, Oklahoma, Kansas, Arkansas, and Louisiana) underwent a precipitation increase of 7.7% (Karl and Knight 1998). Based on the Hadley Centre Climate model (Williams et al. 2001), Oklahoma could experience a 1.11^{o}C temperature increase in the next 100 years, which in turn would have a significant affect on bird distributions in the state.

Since the 1890s there also have been seen significant changes in land uses and land-cover types, brought about in part by widespread conversion of grasslands to cropland and pastures (Ramankutty and Foley 1999). In many areas the mixed-forest land-cover type, typical of riparian areas in Oklahoma, has decreased in extent because of man-made reservoirs like Lake Eufala and Lake Texoma. Numerous Oklahoma communities have incorporated reservoirs and the adjacent areas into town/city limits to ensure that they have a significant role in management of the reservoirs, and they are able to make and enforce zoning ordinances in watersheds of these reservoirs. As a result, some cities have experienced rapid growth in area, such as Lawton, which changed from 80 km^2 in 1990

(Clark 1993) to 195 km^2 in 2000 (US Census Bureau 2006). Typically, significant changes in land-cover types occur with increased urbanization.

Changes in land uses and in climate are expected to continue with concomitant changes in bird distributions within the state. With a marked temperature change (as per Williams et al. 2001), areas that currently are marginal for a particular species may become suitable such that the distribution of the species expands. Prediction models based on environmental variables and current atlas data can be valuable in helping us to predict how and where future avifauna changes likely will occur.

Acknowledgments Support for the first author was provided by George Miksch Sutton Scholarships in Ornithology and a Robert E. and Mary B. Sturgis Scholarship, as well as by a grant from the University of Oklahoma Graduate Student Senate. We thank the Oklahoma landowners Melissa and Bob Blevins, the Little River Zoo, Joe McGowan, Todd and Pam Crawford, Sam Kidman, Carey Lane, C. J. Smith, Aaron Hiremine, Mary Rogers, Don Baker, and the El Reno Municipal Airport for access to properties for the 2004 survey blocks. We thank Jeffrey F. Kelly, William J. Matthews, Thomas S. Ray, and May Yuan for serving as committee members of the first author and for helpful comments on earlier drafts of the manuscript. This article is a portion of a dissertation submitted by the first author in partial fulfillment of requirements for Ph.D. in the Department of Zoology, University of Oklahoma.

Table 1. Climatic variables and loadings for climatic PC1.

	Variable	Loading [a]
14	May temperature (^{o}C)	**-**0.245
15	May soil temperature (^{o}C)	**-**0.237
16	May rain (cm)	**0.786**
17	May solar radiation (MJ/m^2)	**-0.866**
18	May wind speed (kph)	**-0.893**
19	May pressure (cm)	**0.867**
20	May humidity (%)	0.650
21	June temperature (^{o}C)	**-**0.238
22	June soil temperature (^{o}C)	**-0.765**
23	June rain (cm)	**0.902**
24	June solar radiation (MJ/m^2)	**-0.870**
25	June wind speed (kph)	**-0.850**
26	June pressure (cm)	**0.870**
27	June humidity (%)	**0.887**
28	July temperature (^{o}C)	**-**0.464
29	July soil temperature (^{o}C)	**-**0.554
30	July rain (cm)	0.083
31	July solar radiation (MJ/m^2)	**-0.858**
32	July wind speed (kph)	**-0.902**
33	July pressure (cm)	**0.869**
34	July humidity (%)	**0.880**

[a] Relatively high loadings (> |0.75|) identified in bold.

Fig. 1 Maps showing locations of 562 Oklahoma breeding bird atlas blocks, as well as (a) number of land cover types in each block, and (b) projections of each atlas block on climatic principal component 1

Fig. 2 Significant regression of number of species in a block on: (a) number of land cover types (X_1); and (b) climatic PC1 (X_2). Equations were (each $N = 562$): (a) $Y = 34.19 + 2.42X_1$ ($r^2 = 0.12$, $p < 0.001$); and (b) $Y = 9.35 + 0.98X_2$ ($r^2 = 0.17$, $p < 0.001$)

Fig. 3 Performance of 209 individual species models relative to number of sites at which each species occurred: (a) percent correctly predicted occurrences (both presence and absence); (b) percent correctly predicted presences; and (c) τ_p, indicating performance of model relative to expected error without the model. In panel C, solid line shows τ_p for model in which species are judged to be present at all sites, while dashed line results from model where species are considered to be absent from all sites

Fig. 4 Regressions on number of land cover types (X) of: (a, b) number of correctly predicted presences and absences; (c, d) number of incorrectly predicted presences and absences; and (e, f) percent of correctly predicted presences and absences (all $p < 0.001$ and $N = 562$). Equations were: (a) $Y = 27.79 + 166X$ ($r^2 = 0.15$); (b) $Y = 161.25 - 1.94X$ ($r^2 = 0.13$); (c) $Y = 13.57 - 0.48X$ ($r^2 = 0.03$); (d) $Y = 6.40 + 0.76X$ ($r^2 = 0.03$); (e) $Y = 70.53 + 1.30X$ ($r^2 = 0.08$); and (f) $Y = 96.41 - 0.54X$ ($r^2 = 0.04$)

Fig. 5 Regression on climatic PC1 (X) of: (a, b) number of correctly predicted presences and absences; (c, d) number of incorrectly predicted presences and absences; and (e, f) percent of correctly predicted presences and absences. Equations were (all $N = 562$): (a) $Y = 43.32 + 6.22X$ ($r^2 = 0.35$, $p < 0.01$); (b) $Y = 143.15 - 7.25X$ ($r^2 = 0.31$, $p < 0.001$); (c) $Y = 9.04 + 0.87X$ ($r^2 = 0.10$, $p < 0.01$); (d) $Y = 13.49 + 0.15X$ ($r^2 = 0.00$, $p > 0.50$); (e) $Y = 82.67 + 0.95X$ ($r^2 = 0.01$, $p > 0.05$); (f) $Y = 91.38 - 0.45X$ ($r^2 = 0.01$, $p > 0.05$)

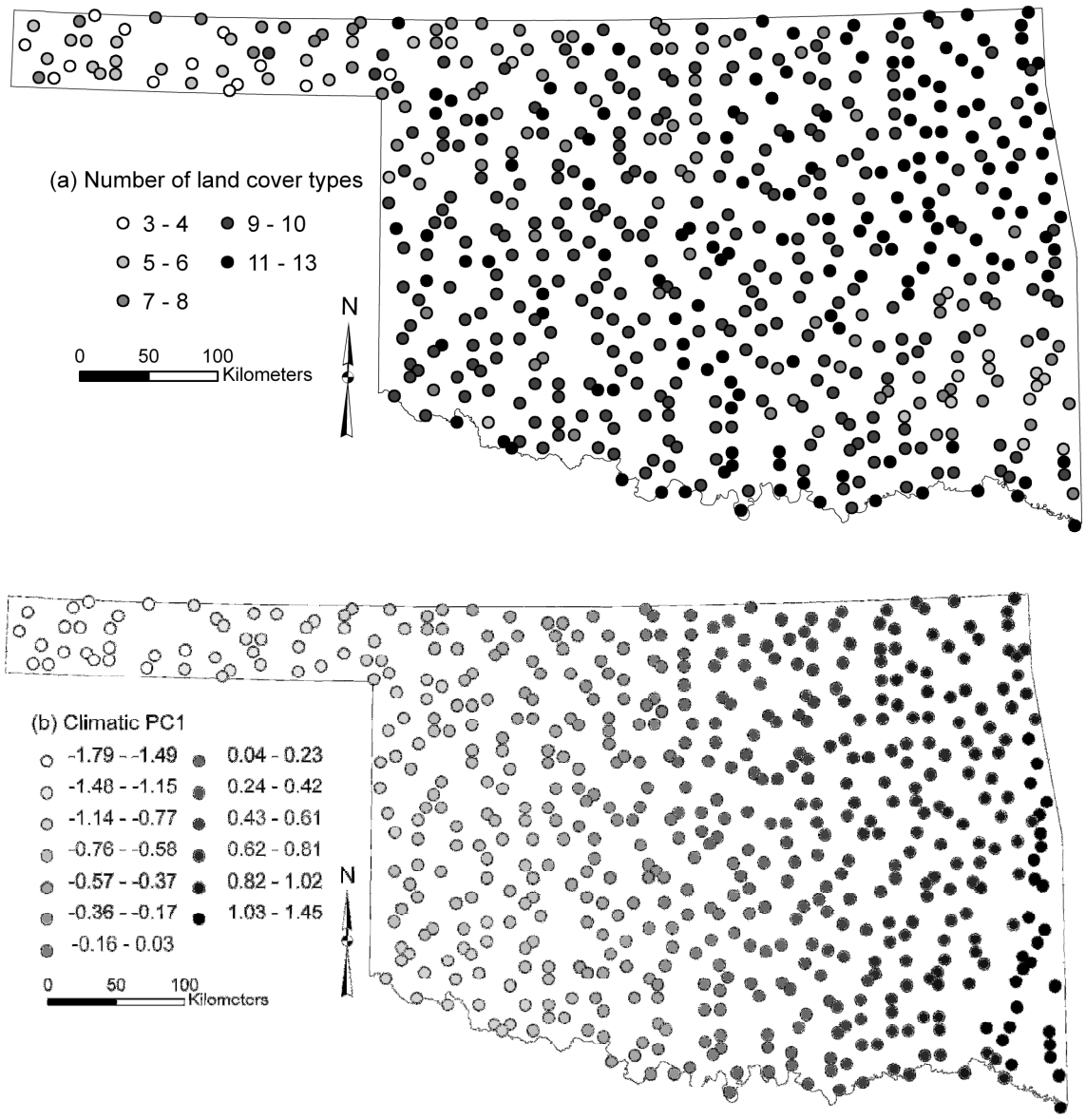
(a) Number of land cover types
3 - 4
5 - 6
7 - 8
9 - 10
11 - 13
N
0 50 100
Kilometers
(b) Climatic PC1
-1.79 - -1.49
-1.48 - -1.15
-1.14 - -0.77
-0.76 - -0.58
-0.57 - -0.37
-0.36 - -0.17
-0.16 - 0.03
0.04 - 0.23
0.24 - 0.42
0.43 - 0.61
0.62 - 0.81
0.82 - 1.02
1.03 - 1.45
N
0 50 100
Kilometers

Figure 1

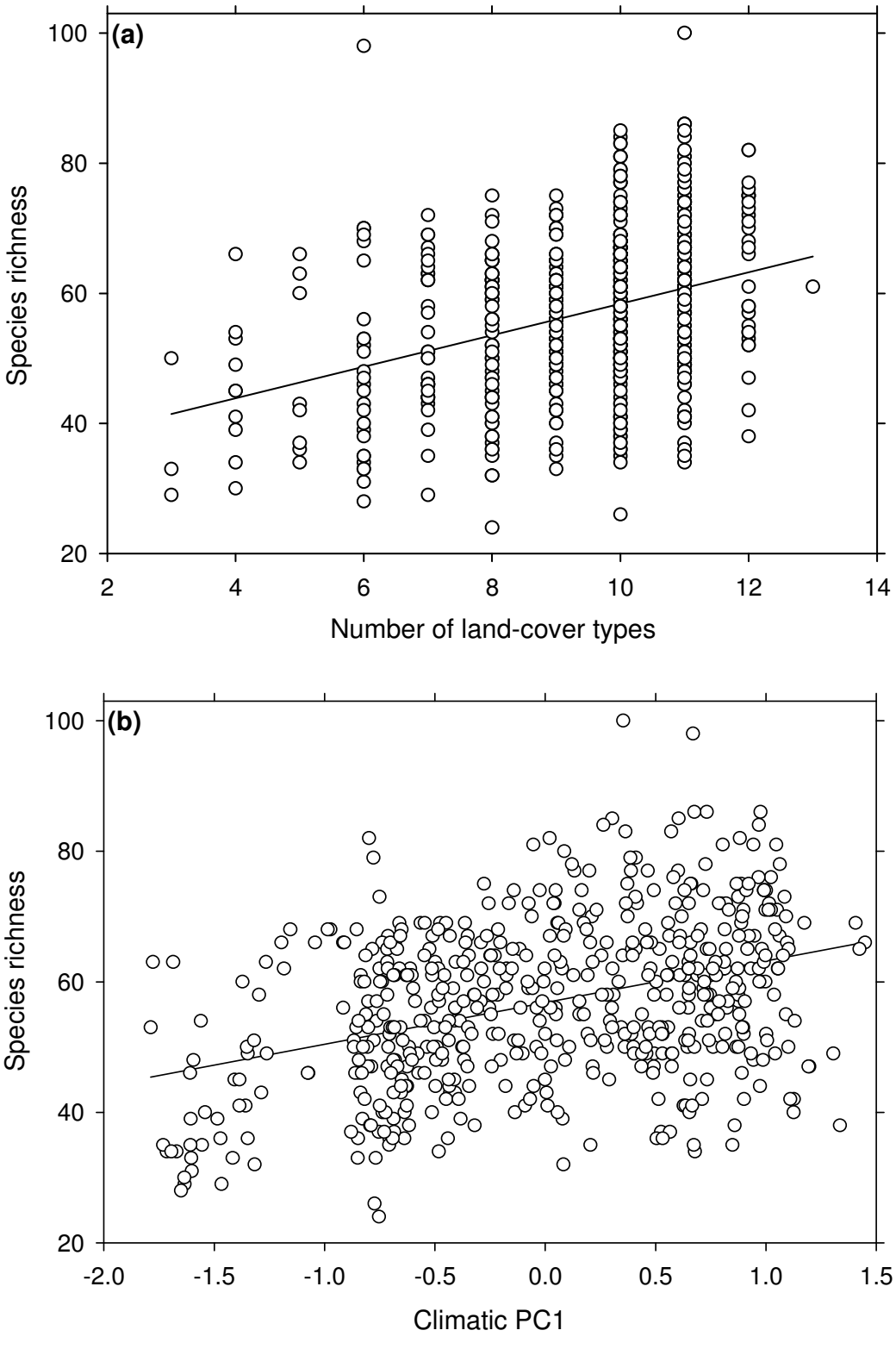
(a)
Species richness
Number of land-cover types
(b)
Species richness
Climatic PC1

Figure 2

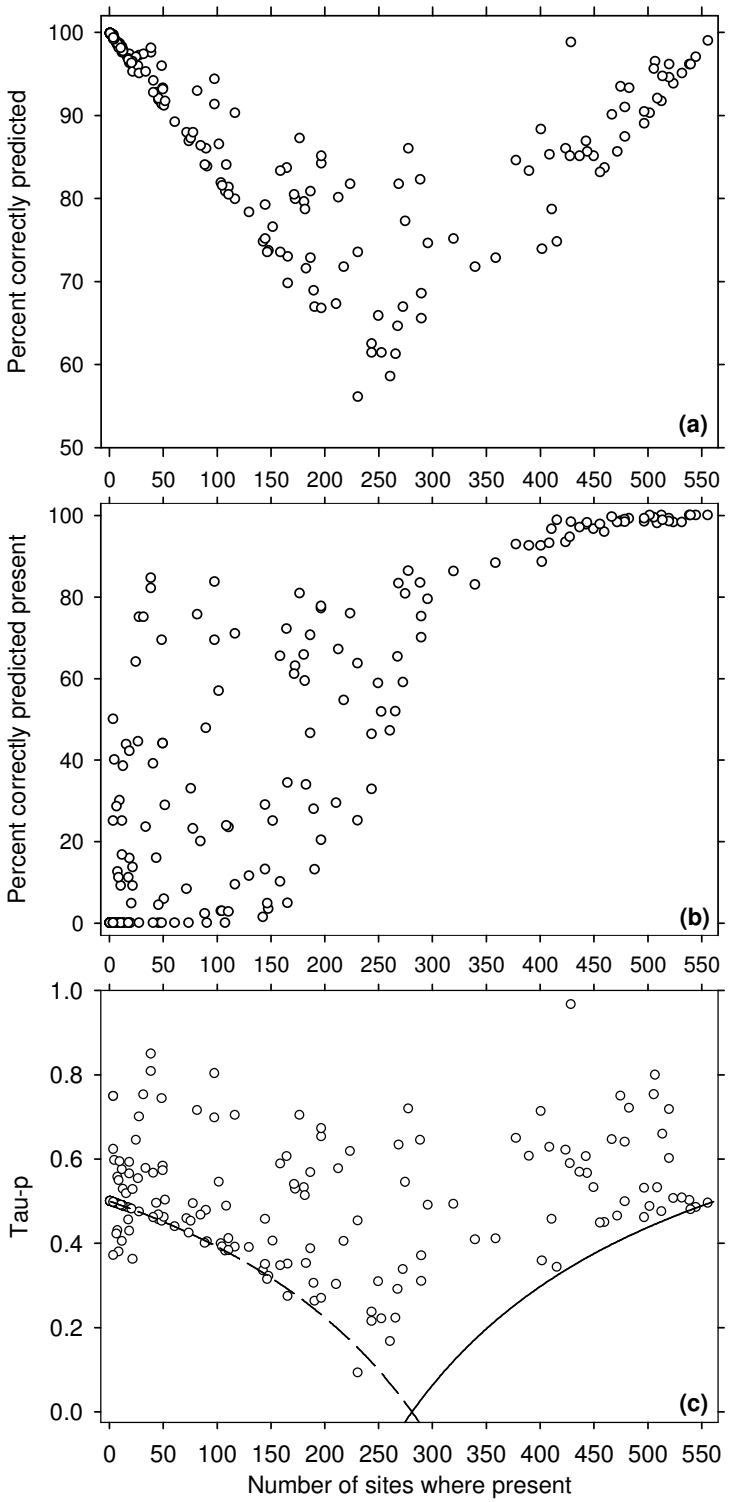
Percent correctly predicted
(a)
Percent correctly predicted present
(b)
Tau-p
(c)
Number of sites where present

Figure 3

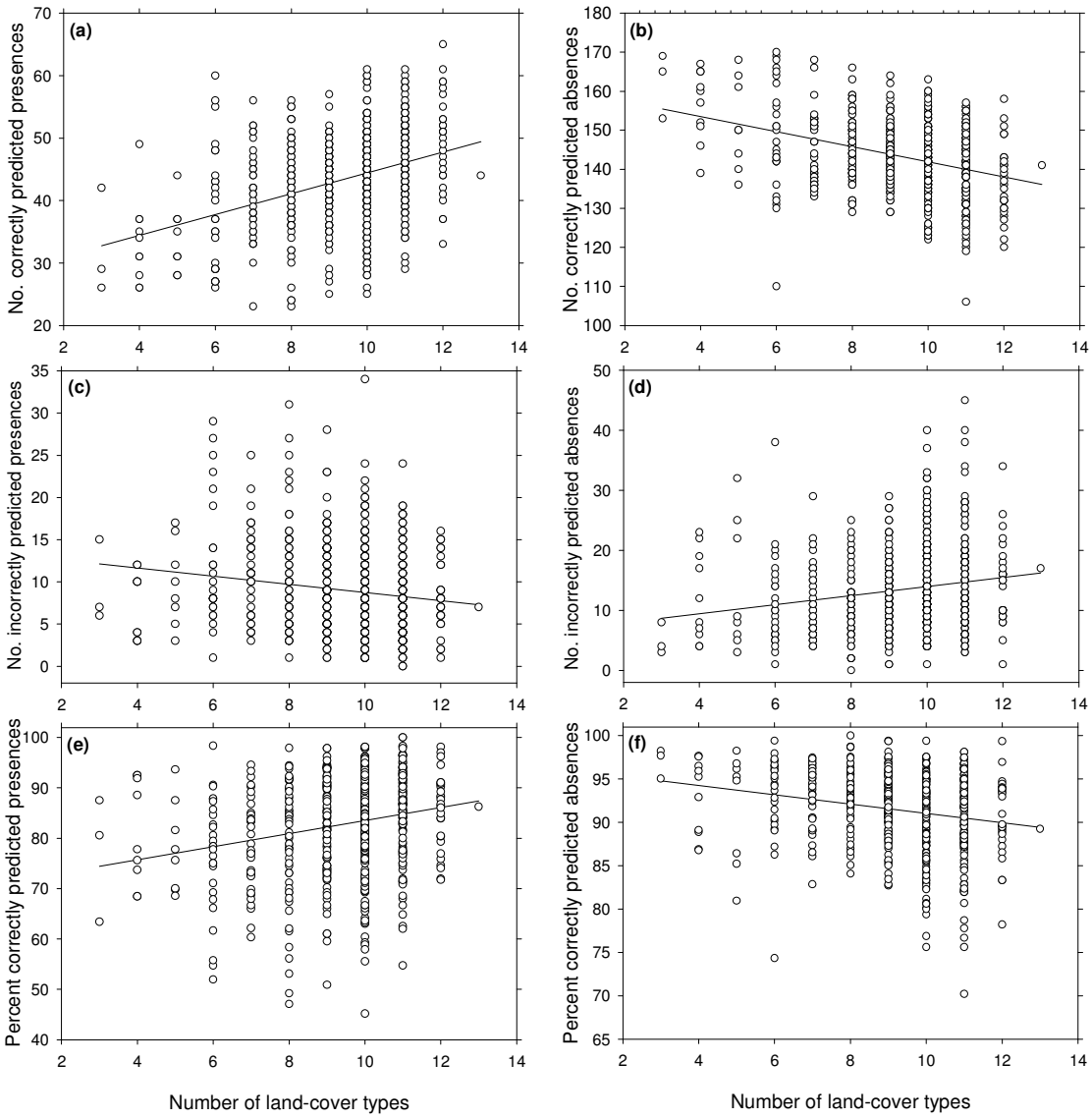
(a)
No. correctly predicted presences
(b)
No. correctly predicted absences
(c)
No. incorrectly predicted presences
(d)
No. incorrectly predicted absences
(e)
Percent correctly predicted presences
(f)
Percent correctly predicted absences
Number of land-cover types
Number of land-cover types

Figure 4

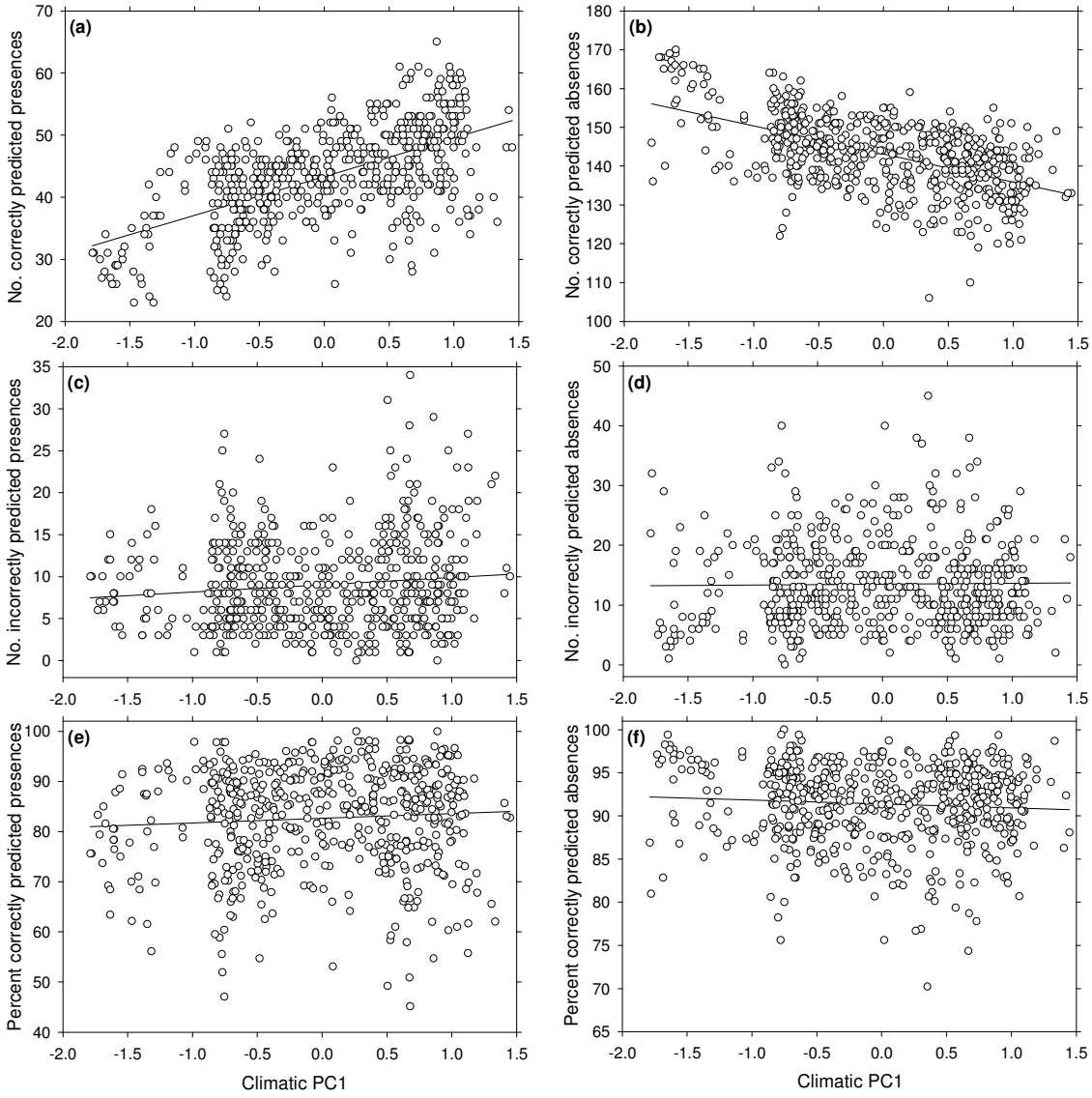
(a)
No. correctly predicted presences
(b)
No. correctly predicted absences
(c)
No. incorrectly predicted presences
(d)
No. incorrectly predicted absences
(e)
Percent correctly predicted presences
(f)
Percent correctly predicted absences
Climatic PC1
Climatic PC1

Figure 5

References

Anderson RP, Gomez-Laverde M, Peterson AT (2002) Geographical distributions of spiny pocket mice in South America: insights from predictive models. Global Ecology and Biogeography 11:131–141

Anderson RP, Lew D, Peterson AT (2003) Evaluating predictive models of species' distributions: criteria for selecting optimal models. Ecological Modelling 162:211–232

Angermeier PL, Krueger KL, Dolloff CA (2002) Discontinuity in stream-fish distributions: implications for assessing and predicting species occurrence. In: Scott JM, Heglund PJ, Morrison ML, Haufler JB, Raphael MG, Wall WA, and Samson FB (eds), Predicting species occurrences: issues of accuracy and scale. Island Press, Washington D.C., USA

Austin GE, Thomas CJ, Houston DC, Thompson DBA (1996) Predicting the spatial distribution of buzzard *Buteo buteo* nesting areas using a geographical information system and remote sensing. Journal of Applied Ecology 33:1541–1550

Bevier LR (ed) (1994) The atlas of breeding birds of Connecticut. State Geological and Natural History Survey of Connecticut, Hartford, Connecticut, USA

Beyer HL (2004) Hawth's analysis tools for ArcGIS. http://www.spatialecology.com/htools. September 23, 2005

Brewer R, McPeek GA, Adams RJ (1991) The atlas of breeding birds of Michigan. Michigan State University Press, East Lansing, Michigan, USA

Bustamante J, Seone J (2004) Predicting the distribution of four species of raptors (Aves: Accipitridae) in southern Spain: statistical models work better than existing maps. Journal of Biogeography 31:295–306

Brock FV, Crawford KC, Elliot RL, Cuperus GW, Stadler SJ, Johnson HL, Eilts MD (1995) The Oklahoma Mesonet: a technical overview. Journal of Atmospheric and Oceanic Technology 12:5–19

Brotons L, Thuiller W, Araujo MB, Hirzel AH (2004) Presence-absence versus presence-only modeling methods for predicting bird habitat suitability. Ecography 27:437–448

Clark RL (1993) Oklahoma almanac. Oklahoma Department of Libraries, Oklahoma City, Oklahoma, USA

Cox J (2006) Trends in breeding distributions based on Florida's breeding bird atlas project. In: Noss RF (ed) The breeding birds of Florida. Special publication No. 7, Florida Ornithological Society, Gainsville, Florida, USA

Engler R, Guisan A, Rechsteiner L (2004) An improved approach to predicting the distribution of rare and endangered species from occurrence and pseudo-absence data. Journal of Applied Ecology 41:263–274

ESRI (2004) ArcGIS version 9.0. Environmental Systems Research Institute, Redlands, California, USA

Gibson LA, Wilson BA, Cahill DM, Hill J (2004) Spatial prediction of rufous bristlebird habitat in a coastal heathland: a GIS-based approach. Journal of Applied Ecology 41:213–223

Gilbert M, Nageleisen L-M, Franklin A, Grégoire J-C (2005) Post-storm surveys reveal large-scale spatial patterns and influences of site factors, forest structure and diversity in endemic bark-beetle populations. Landscape Ecology 20:25–49

Hosmer DW and Lemeshow S (2000) Applied logistic regression. John Wiley and Sons, Hoboken, New Jersey, USA

Ingold JL (1993) Blue grosbeak (*Guiraca caerulea*). In: Poole A, Gill F (eds), Birds of North America No. 79, Academy of Natural Sciences, Philadelphia, Pennsylvania, USA and American Ornithologists' Union, Washington D.C., USA

Karl TR, Knight RW (1998) Secular trends of precipitation amount, frequency, and intensity in the United States. Bulletin of the American Meteorological Society 79:231–241

Koleff P, Gaston KJ (2002) The relationship between local and regional richness and spatial turnover. Global Ecology and Biogeography 11:363–375

Lawler JJ, O'Conner RJ, Hunsaker CT, Jones KB, Loveland TR, White D (2004) The effects of habitat resolution on models of avian diversity and distributions: a comparison of two land-cover classifications. Landscape Ecology 19:515–530

Lim BK, Peterson AT, Engstrom MD (2002) Robustness of ecological modeling algorithms for mammals in Guyana. Biodiversity and Conservation 11:1237–1246

MacKinzie DI, Nichols JD, Royle JA, Polluck KH, Bailey LL, Hines JE (2006) Occupancy estimation and modeling: inferring patterns and dynamics of species occurrence. Elsevier, Boston, Massachusetts, USA

Martaugh PA (1996) The statistical evaluation of ecological indicators. Ecological Applications 6:132–139

Manel S, Dias JM, Buckton ST, Ormerod SJ (1999) Alternative methods for predicting species distribution: an illustration with Himalayan river birds. Journal of Applied Ecology 36:734–747

Menard, S (2002) Applied logistic regression analysis. Sage Publications, Thousand Oaks, California, USA

Pearce J, Ferrier S (2000) Evaluating the predictive performance of habitat models developed using logistic regression. Ecological Modelling 133:225–245

Peterjohn BG, Rice DL (1991) The Ohio breeding bird atlas. Ohio Department of Natural Resources, Columbus, Ohio, USA

Peterson AT (2001) Predicting species' geographical distributions based on ecological niche modelling. Condor 103:599–605

Peterson AT, Vieglais DA (2001) Predicting species invasions using ecological niche modeling: new approaches from bioinformatics attack a pressing problem. BioScience 51:363–371

Peterson AT, Ortega-Huerta O, Bartley J, Sanchez-Cordero V, Soberon J, Buddemeier RH, Stockwell DRB (2002) Future projections for Mexican faunas under global climate change scenarios. Nature 416:626–629

Peterson AT, Kluza DA (2003) New distributional modeling approaches for gap analysis. Animal Conservation 6:47–54

Peterson AT, Robins CR (2003) Using ecological-niche modeling to predict barred owl invasions with implications for spotted owl conservation. Conservation Biology 17:1161–1165

Peterson AT, Papes M, Reynolds MG, Perry ND, Hanson B, Regnery R, Hutson CL, Muizniek B, Damon IK, Carrol DS (2006) Native-range ecology and invasive potential of *Cricetomys* in North America. Journal of Mammalogy 87:427–432

Rahbeck C, Graves GR (2001) Multiscale assessment of patterns of avian species richness. Proceedings of the National Academy of Science USA. 98:4534–4539

Ramankutty N, Foley JA (1999) Estimating historical changes in land cover: North American croplands from 1850 to 1992. Global Ecology and Biogeography 8:381–396

Raxworthy CJ, Martinez-Meyer E, Horning N, Nussbaum RA, Schneider GE, Ortega-Huerta MA, Peterson AT (2003) Predicting distributions of known and unknown reptile species in Madagascar. Nature 426:837-841

Reinking DL (2000) Oklahoma breeding bird atlas handbook. George M. Sutton Avian Research Center, Bartlesville, Oklahoma, USA

Reinking DL (ed) (2004) Oklahoma breeding bird atlas. University of Oklahoma Press, Norman, Oklahoma, USA

Rolf FJ (2002) NTSYSpc: numerical taxonomy system, Release 2.2. Exeter Software, Setauket, New York, USA

SAS (2004) Statistical Analysis Software, release 9. SAS Institute Inc, Cary North Carolina, USA

Simberloff D (1988) The contribution of population and community biology to conservation science. Annual Review of Ecology and Systematics 19:473–511

Stockwell DRB, Peters D (1999) The GARP modeling system: problems and solutions to automated spatial prediction. International Journal of Geographic Information Science 13:143–158

Telfair RC (1994) Cattle egret (*Bubulcus ibis*). In: Poole A, Gill F (eds), Birds of North America No. 113, Academy of Natural Sciences, Philadelphia, Pennsylvania, USA and American Ornithologists' Union, Washington D.C., USA

Titeux N, Dufrene M, Jacob JP, Paquay M, Defourny P (2004) Multivariate analysis of a fine-scale breeding bird atlas using a geographical information system and partial canonical correspondence analysis: environmental and spatial effects. Journal of Biogeography 31:1841–1856

US Census Bureau (2006) Oklahoma quickfacts. http://quickfacts.census.gov/qfd/states/40/4041850.html. August 10, 2006

USGS (1998) US Geological Survey DEM 7.5 Quadrangle. US Geological Survey, Reston, Virginia, USA

USGS (2002) Oklahoma land-cover data set. http://edcftp.cr.usgs.gov/pub/data/landcover/states. September 23, 2005

van Horseen PW, Schot PP, Barendregt A (1999) A GIS-based plant prediction model for wetland ecosystems. Landscape Ecology 14:253–265

van Rensburg BJ, Chown SI, Gaston KJ (2002) Species richness, environmental correlates, and spatial scale: a test using South African birds. American Naturalist 159:567–577

Venier LA, Pearce J, McKee JE, McKenney DW, Niemi GJ (2004) Climate and satellite-derived land cover for predicting breeding bird distribution in the Great Lakes Basin. Journal of Biogeography 31:315–331

Villard M-A, Trzcinski MK, Merriam G (1999) Fragmentation effects on forest birds: relative influence of woodland cover and configuration on landscape occupancy. Conservation Biology 13:774–783

Williams KD, Senior CA, Mitchell JFB (2001) Transient climate change in the Hadley Centre models: the role of physical processes. Journal of Climate 14:2659–2674

Supplementary Material

Appendix 1 Value of τ_p and unstandardized logistic-regression equation for each species[a]

Podicipedidae: pied-billed grebe (*Podilymbus podiceps*; 0.46), $Y = -78.30 + 18.58X_4 + 0.35X_{27} + 1.39X_{28} + 0.41X_{30}$; eared grebe (*Podiceps nigricollis*; 0.50), $Y = -70.35 + 2.38X_{22}$; western grebe (*Aechmophorus occidentalis*; 0.50), $Y = -6.33$. **Pelecanidae**: double-crested coromorant (*Phalacrocorax auritus*; 0.46), $Y = 20.93 + 9.93X_{12} - 0.76X_{24}$. **Ardeidae**: American bittern (*Botaurus lentiginosus*; 0.50), $Y = -5.47 + 36.79X_5$; great blue heron (*Ardea herodias*; 0.46), $Y = -48.70 - 7.27X_3 - 1.08X_9 - 12.71X_{11} + 0.13X_{18} + 0.25X_{27} + 1.18X_{33}$; great egret (*A. alba*; 0.40), $Y = 112.8 + 35.01X_4 + 1.68X_8 + 7.95X_{12} - 1.44X_{14} + 0.17X_{18} + 0.25X_{20} + 3.56X_{26} + 1.45X_{28}$; snowy egret (*Egretta thula*; 0.40), $Y = 38.78 - 2.49X_1 + 4.51X_{12} - 1.03X_{16} - 4.86X_{17} + 0.65X_{20} + 1.81X_{31}$; little blue heron (*E. caerulea*; 0.30), $Y = 23.20 - 1.71X_{17} + 0.33X_{20} - 0.62X_{21}$; cattle egret (*Bubulcus ibis*; 0.09), $Y = 28.8824823 + 7.31X_{10} - 0.99X_{31}$; green heron (*Butorides virescens*; 0.22), $Y = -42.25 + 1.49X_8 - 2.03X_9 + 0.22X_{20} + 0.09X_{25} + 0.93X_{26}$; black-crowned night-heron (*Nycticorax nycticorax*; 0.49), $Y = 144.9 + 48.23X_4 + 11.72X_{10} - 5.45X_{14} - 5.16X_{21} - 15.23X_{24} + 7.20X_{29} + 22.28X_{31}$; yellow-crowned night-heron (*N. violacea*; 0.48), $Y = -26.78 + 6.63X_2 + 14.89X_4 + 5.87X_{10} + 0.85X_{14}$. **Threskiornithidae**: white-faced ibis (*Eudocimus albus*; 0.49), $Y = -45.29 + 47.04X_5 + 0.98X_{28} + 0.38X_{32}$. **Cathartidae**: black vulture (*Coragyps atratus*; 0.48), $Y = 233.1 - 39.75X_{13} - 3.65X_{14} + 0.63X_{18} + 2.88X_{23} - 3.80X_{24} - 0.79X_{27} - 2.37X_{30} + 0.73X_{34}$; turkey vulture (*Cathartes aura*; 0.53), $Y = 152.3 - 1.86X_9 - 6.57X_{10} + 0.38X_{20} + 2.11X_{24} + 2.61X_{33}$. **Anatidae**: black-bellied whistling-duck (*Dendrocygna autumnalis*; 0.50), $Y = -6.33$; Canada goose (*Branta canadensis*; 0.31), $Y = 50.13 + 1.55X_7 + 3.80X_{10} - 1.71X_{17}$; wood duck (*Aix sponsa*; 0.39), $Y = 42.06 +$

$10.70X_4 - 2.06X_8 - 1.49X_{17}$; gadwall (*Anas strepera*; 0.75), $Y = -1783.7 + 57.47X_{31}$; American wigeon (*Anas americana*; 0.50), $Y = -5.63$; mallard (*Anas platyhynchos*; 0.39), $Y = -39.82 - 1.66X_1 + 9.59X_4 - 1.09X_7 + 5.09X_{10} + 12.9X_{13} + 0.91X_{29}$; blue-winged teal (*A. discors*; 0.32), $Y = -31.57 + 0.35X_{28} + 0.09X_{32} + 0.20X_{20}$; cinnamon teal (*Anas cyanoptera*; 0.50), $Y = -133.50 + 4.18X_{17}$; northern shoveler (*A. clypeata*; 0.45), $Y = -2.76 + 1.50X_9$; northern pintail (*A. acuta*; 0.48), $Y = -55.13 + 1.77X_9 + 1.68X_{17}$; green-winged teal (*A. crecca*; 0.50), $Y = -138.5 + 4.35X_{17}$; redhead (*Aythya americana*; 0.47), $Y = -129.8 + 1.49X_{14} + 2.75X_{24} + 0.74X_{30}$; ring-necked duck (*A. collaris*; 0.50), $Y = -4.94$; lesser scaup (*Aythya affinis*; 0.50), $Y = -6.33$; hooded merganser (*Lophodytes cucullatus*; 0.49), $Y = -4.24$; ruddy duck (*Oxyura jamaicensis*; 0.49), $Y = -7.0681 + 0.15X_{32}$. **Accipitridae**: osprey (*Pandion haliaetus*; 0.38), $Y = -30.10 + 9.43X_{12} + 0.95X_{14}$; Mississippi kite (*Ictinia mississippiensis*; 0.45), $Y = -69.29 + 6.07X_6 + 1.59X_7 + 8.07X_{10} + 0.53X_{20} - 1.01X_{23} + 0.76X_{28}$; bald eagle (*Haliaeetus leucocephalus*; 0.49), $Y = -12.40 + 1.78X_{23}$; northern harrier (*Circus cyaneus*; 0.35), $Y = -2.84 - 3.44X_1 - 3.71X_6 + 0.09X_{18}$; sharp-shinned hawk (*Accipiter striatus*; 0.50), $Y = -6.33$; Cooper's hawk (*A. cooperii*; 0.42), $Y = 29.40 - 1.03X_{17}$; red-shouldered hawk (*Buteo lineatus*; 0.57), $Y = -36.58 + 3.73X_1 + 5.33X_2 - 9.41X_6 - 3.90X_9 - 0.27X_{27} + 2.75X_{33}$; broad-winged hawk (*B. platypterus*; 0.47), $Y = -46.67 + 3.51X_1 + 0.49X_{20} + 3.94X_{12}$; Swainson's hawk (*B. swainsoni*; 0.54), $Y = -82.02 - 4.48X_1 + 1.72X_9 + 2.67X_{17}$; red-tailed hawk (*B. jamaicensis*; 0.49), $Y = -152.2 + 2.49X_7 - 3.73X_{12} + 2.69X_{24} + 3.32X_{33}$; ferruginous hawk (*B. regalis*; 0.36), $Y = -889.4 - 16.04X_{14} + 6.82X_{15} + 19.34X_{17} + 1.61X_{20} + 12.84X_{28}$; golden eagle (*Aquila chrysaetos*; 0.50), $Y = -6.33$. **Falconidae**: American kestrel (*Falco sparverius*; 0.37), $Y = -36.59 - 2.45X_1 + 1.26X_{31}$; prairie falcon (*F. mexicanus*; 0.49), $Y =$

–92.60 + 0.99X_{14} + 3.24X_{17} – 0.42X_{20}. **Phasianidae**: ring-necked pheasant (*Phasianus colchicus*; 0.80), Y = –114.5 + 4.19X_9 + 39.93X_{13} + 2.28X_{29} + 1.12X_{30}; greater prairie-chicken (*Tympanuchus cupido*; 0.50), Y = –10.11 + 1.56X_{30}; lesser prairie-chicken (*T. pallidicinctus*; 0.42), Y = –1609.0 + 49.74X_{17} + 18.51X_{23}; wild turkey (*Meleagris gallopavo*; 0.30), Y = –21.16 + 2.26X_1 + 5.445X_3 + 5.42X_6 – 2.13X_8 + 0.63X_{28}. **Odontiphoridae**: scaled quail (*Callipepla squamata*; 0.75), Y = –116.1 + 8.39X_6 – 4.86X_{15} + 8.24X_{21}; northern bobwhite (*Colinus virginianus*; 0.60), Y = –204.7 + 3.34X_7 – 6.23X_{10} + 3.53X_{15} – 0.32X_{17} + 1.26X_{20} + 1.34X_{23} + 10.42X_{24}. **Rallidae**: king rail (*Rallus elegans*; 0.50), Y = –6.33; sora (*Porzana carolina*; 0.50), Y = 49.92 – 2.67X_{33}; American coot (*Fulica americana*; 0.44), Y = 15.90 – 0.21X_{27}. **Charadriidae**: killdeer (*Charadrius vociferous*; 0.53), Y = –5.72 + 4.42X_8 + 0.32X_{18}; mountain plover (*C. montanus*; 0.50), Y = –5.64. **Recurvirostridae**: black-necked stilt (*Himantopus mexicanus*; 0.49), Y = 33.73 – 0.45X_{27} + 47.65X_5; American avocet (*Recurvirostra americana*; 0.40), Y = –296.14 + 10.14X_{17} – 3.88X_7 – 1.101X_{27} + 90.04X_5 + 3.56X_{26}. **Scolopacidae**: spotted sandpiper (*Actitis macularia*; 0.47), Y = –2.59 + 5.84X_{12}; upland sandpiper *(Bartramia longicauda*; 0.49), Y = –17.98 – 4.49X_1 – 4.71X_6 + 0.67X_{23} + 0.18X_{25} + 0.12X_{34}; long-billed curlew (*Numenius americanus*; 0.52), Y = 55.39 – 2.81X_{33}; Wilson's phalarope (*Phalaropus tricolor*; 0.37), Y = 49.32 + 6.12X_9 – 2.74X_{19}. **Laridae**: ring-billed gull (*Larus delawarensis*; 0.50), Y = –6.33; least tern (*Sterna antillarum*; 0.43), Y = –201.0 + 51.93X_5 + 26.87X_{11} + 10.81X_{12} + 2.93X_{14} + 1.26X_{20} + 0.19X_{32}. **Columbidae**: rock pigeon (*Columba livia*; 0.30), Y = –2.54 + 2.40X_8 + 16.25X_{10} + 4.60X_{13} – 0.71X_{16}; Eurasian collard-dove (*Streptopelia decaocto*; 0.50), Y = –5.63; white-winged dove (*Zenaida asiatica*; 0.50), Y = –6.33; mourning dove (*Z. macroura*; 0.50), Y = –52.71 + 0.40X_{25} +

$0.59X_{34}$; Inca dove (*Columbina inca*; 0.50), $Y = –6.33$. **Cuculidae**: black-billed cuckoo (*Coccyzus erythropthalmus*; 0.49), $Y = –4.39 + 12.87X_4$; yellow-billed cuckoo (*C. americanus*; 0.66), $Y = –36.91 + 11.00X_6 – 18.17X_{11} + 1.43X_{16} + 1.58X_{33}$; greater roadrunner (*Geococcyx californianus*; 0.35), $Y = 35.80 + 6.63X_3 – 17.88X_4 + 4.76X_6 + 1.41X_7 + 0.34X_{20} – 1.59X_{24} – 0.23X_{34}$. **Tytonidae**: barn owl (*Tyto alba*; 0.50), $Y = –39.35 + 35.22X_5 + 2.37X_9 + 3.03X_{21} – 1.56X_{29}$. **Strigidae**: western screech-owl (*Megascops kennicottii*; 0.50), $Y = –6.33$; eastern screech-owl (*M. asio*; 0.46), $Y = –2.75 + 4.78X_2 + 3.90X_6$; great horned owl (*Bubo virginianus*; 0.26), $Y = –2.40 + 0.08X_{32}$; burrowing owl (*Athene cunicularia*; 0.81), $Y = –232.39 + 7.45X_{17} + 2.35X_9$; barred owl (*Strix varia*; 0.33), $Y = –45.39 + 0.33X_{27} + 0.47X_{28}$; short-eared owl (*Asio flammeus*; 0.62), $Y = –729.1 + 13.69X_{16} + 4.93X_{30} + 21.69X_{31}$. **Caprimulgidae**: common nighthawk (*Chordeiles minor*; 0.54), $Y = 2.72 – 60.94X_1 + 13.51X_7 – 96.84X_3 + 70.46X_2$; common poorwill (*Phalaenoptilus nuttallii*; 0.49), $Y = –100.91 + 5.06X_7 + 3.41X_{22}$; chuck-will's-widow (*Caprimulgus carolinensis*; 0.39), $Y = –26.62 – 1.32X_8 – 3.55X_9 + 0.29X_{20}$; whip-poor-will (*C. vociferous*; 0.49), $Y = –10.47 – 8.34X_7 + 1.71X_{16}$. **Apodidae**: chimney swift (*Chaetura pelagica*; 0.31), $Y = –160.89 + 3.33X_{26} + 20.41X_{10} – 1.38X_7 – 3.75X_{12} + 3.38X_{31} + 0.48X_{20} – 0.90X_{16} – 1.65X_{17}$. **Trochilidae**: ruby-throated hummingbird (*Archilochus colubris*; 0.64), $Y = 63.92 + 3.73X_1 + 8.07X_3 – 20.23X_{11} + 0.30X_{18} – 2.11X_{24} – 0.37X_{32}$; black-chinned hummingbird (*A. alexandri*; 0.49), $Y = 64.75 – 28.88X_{13} – 1.78X_{29} + 0.53X_{32}$. **Alcedinidae**: belted kingfisher (*Ceryle alcyon*; 0.23), $Y = –24.84 + 1.96X_7 + 4.64\ X_{12} + 0.48X_{20} + 2.70X_{24} + 0.88X_{15} – 4.04X_{17}$. **Picidae**: Lewis's woodpecker (*Melanerpes lewis*; 0.50), $Y = –6.33$; red-headed woodpecker (*M. erythrocephalus*; 0.22), $Y = –42.54 + 18.17X_4 + 9.18X_{13} + 0.78X_{15} + 0.17X_{20}$; golden-

fronted woodpecker (*M. aurifrons*; 0.49), $Y = 265.2 + 16.67X_6 – 40.09X_{13} – 3.03X_{34}$; red-bellied woodpecker (*M. carolinus*; 0.65), $Y = –154.6 – 14.02X_2 + 30.54X_3 – 5.46X_{17} + 0.80X_{20} + 5.41X_{24} – 0.94X_{30} + 4.03X_{33}$; ladder-backed woodpecker (*Picoides scalaris*; 0.57), $Y = –5.57 + 8.15X_6 – 0.70X_{20} + 1.36X_{21} + 0.82X_{28}$; downy woodpecker (*P. pubescens*; 0.61), $Y = –94.95 + 2.13X_7 + 0.35X_{27} + 0.14X_{32} + 2.89X_{33}$; hairy woodpecker (*P. villosus*; 0.27), $Y = 7.27 + 2.41X_1 + 0.17X_{18} + 0.21X_{20} – 1.05X_{31}$; northern flicker (*Colaptes auratus*; 0.21), $Y = –25.16 + 1.95X_1 + 1.046X_7 + 11.35X_{10} + 11.57X_{11} + 6.34X_{13} + 0.75X_{15}$; pileated woodpecker (*Dryocopus pileatus*; 0.59), $Y = 35.23 + 5.87X_1 + 9.21X_2 + 28.43X_4 + 2.14X_7 + 0.46X_{18} + 0.89X_{23} – 1.520X_{31} – 0.43X_{32}$. **Tyrannidae**: western wood-pewee (*Contopus sordidulus*; 0.50), $Y = 62.71 – 3.32X_{33}$; eastern wood-pewee (*C. virens*; 0.71), $Y = 42.41 + 4.50X_1 – 1.87X_{24} + 0.19X_{18} + 2.21X_8 – 1.031X_{28} + 1.91X_{26}$; Acadian flycatcher (*Empidonax virescens*; 0.59), $Y = 18.46 + 3.31X_1 + 1.76X_{23} +18.88X_4 – 0.96X_{28}$; willow flycatcher (*E. traillii*; 0.46), $Y = 14.91 + 25.82X_5 – 0.59X_{28}$; eastern phoebe (*Sayornis phoebe*; 0.64), $Y = 130.5 – 6.29X_2 – 33.30X_5 – 3.04X_{17} – 1.23X_{21} – 0.99X_{30}$; Say's phoebe (*S. saya*; 0.70), $Y = 207.5 + 5.42X_{28} – 12.55X_{33} – 1.56X_{34}$; vermilion flycatcher (*Pyrocephalus rubinus*; 0.50), $Y = –6.33$; ash-throated flycatcher (*Myiarchus cinerascens*; 0.56), $Y = 103.31 –0.54X_{27} – 0.78X_{34}$; great crested flycatcher (*M. crinitus*; 0.0.57), $Y = –186.0 + 1.44X_7 + 0.72X_{20} + 1.76X_{24} + 3.18X_{33}$; Cassin's kingbird (*Tyrannus vociferans*; 0.50), $Y = –4.93$; western kingbird (*T. verticalis*; 0.49), $Y = –11.59 – 5.47X_6 + 2.0X_8 + 14.49X_{10} + 4.68X_{11} + 0.86X_{18} + 0.45X_{20} – 0.40X_{25} – 0.47X_{27}$; eastern kingbird (*T. tyrannus*; 0.50), $Y = –57.2538 + 3.62X_8 + 11.58X_{13} + 0.22X_{27} + 2.22X_{33} – 0.23X_{34}$; scissor-tailed flycatcher (*T. forficatus*; 0.72), $Y = –322.1 – 6.63X_1 – 19.14X_4 + 14.14X_7 + 8.03X_9 + 0.69X_{18} + 1.99X_{23} + 0.90X_{27} + 10.22X_{33}$. **Laniidae**:

loggerheaded shrike (*Lanius ludovicianus*; 0.41), $Y = 13.40 – 2.22X_1 – 10.83X_3 + 1.18X_7 + 3.58X_8 – 0.33X_{20} + 0.15X_{28}$. **Vireonidae**: white-eyed vireo (*Vireo griseus*; 0.70), $Y = 60.73 + 2.14X_1 – 3.03X_9 + 0.91X_{23} – 2.07X_{24} – 1.05X_{30}$; Bell's vireo (*V. bellii*; 0.35), $Y = 42.88 + 1.35\ X_{16} + 2.53X_7 + 0.14X_{18} + 2.10X_8 – 1.77X_{24}$; black-capped vireo (*V. atricapilla*; 0.60), $Y = –302.97 + 20.30X_6 + 6.01X_{28} + 1.23X_{34} – 68.84X_5$, yellow-throated vireo (*V. flavifrons*; 0.57), $Y = 21.19 + 3.69X_1 + 5.44X_2 + 15.97X_4 – 2.35X_{17} + 0.51X_{20}$; warbling vireo (*V. gilvus*; 0.35), $Y = –22.53 + 13.26X_{13} + 0.92X_{15} – 0.19X_{32} – 0.11X_{34}$; red-eyed vireo (*V. olivaceus*; 0.58), $Y = –3.59 + 5.10X_1 + 10.37X_2 – 1.73X_9 + 0.60X_{23}$. **Corvidae**: blue jay (*Cyanocitta cristata*; 0.45), $Y = 11.64 – 1.46X_9 – 0.89X_{21} + 0.55X_{23} + 0.39X_{28}$; western scrub-jay (*Aphelocoma californica*; 0.50), $Y = –5.63$; pinyon jay (*Gymnorhinus cyanocephalus*; 0.50), $Y = –6.33$; black-billed magpie (*Pica hudsonia*; 0.49), $Y = 52.20 – 238.36X_{19} + 235.85X_{26}$; American crow (*Corvus brachyrhynchos*; 0.51), $Y = –32.45 + 4.50X_9 + 0.37X_{20} + 0.55X_{25} – 0.56X_{32}$; fish crow (*C. ossifragus*; 0.55), $Y = –219.60 + 2.68X_1 + 20.59X_4 – 11.86X_{13} + 1.75X_{16} + 0.82X_{23} + 3.53X_{31} + 5.20X_{33}$; Chihuahuan raven (*C. cryptoleucus*; 0.53), $Y = 57.50 – 2.93X_{33}$; common raven (*C. corax*; 1.0), $Y = –6.33$. **Alaudidae**: horned lark (*Eremophila alpestris*; 0.53), $Y = –14.70 – 5.27X_1 – 4.49X_6 + 1.85X_9 + 1.93X_{17} – 0.50X_{20}$. **Hirundinidae**: purple martin (*Progne subis*; 0.63), $Y = 97.83 + 3.48X_1 + 1.71X_7 + 5.86X_8 + 29.31X_{10} – 3.20X_{17} + 0.31X_{18} – 0.88X_{23} – 0.38X_{32}$; tree swallow (*Tachycineta bicolor*; 0.53), $Y = –5.11 + 10.45X_{12} + 0.58X_{30}$; northern rough-winged swallow (*Stelgidopteryx serripennis*; 0.17), $Y = –4.02 – 1.95X_9 + 0.48X_{23} + 0.10X_{25}$; bank swallow (*Riparia riparia*; 0.49), $Y = –5.67 + 0.69X_{30}$; cliff swallow (*Petrochelidon pyrrhonota*; 0.31), $Y = –11.89 + 4.059X_{12} + 0.11X_{18} + 0.32X_{22}$; barn swallow (*Hirundo rustica*; 0.50), $Y = 11.10 – 8.07X_{13}$. **Paridae**: Carolina

chickadee (*Poecile carolinensis*; 0.75), $Y = –22.86 + 3.65X_{15} – 0.26X_{18} + 0.89X_{20} – 4.67X_{21} – 2.10X_{23}$; tufted titmouse (*Baeolophus bicolor*; 0.63), $Y = 76.31 + 19.52X_1 – 19.17X_2 + 14.10X_3 – 2.49X_{24}$. **Sittidae**: white-breasted nuthatch (*Sitta carolinensis*; 0.62), $Y = 68.59 + 3.90X_1 – 2.55X_9 – 1.22X_{16} – 2.27X_{17} + 1.11X_{23}$; brown-headed nuthatch (*S. pusilla*; 0.50), $Y = –6.0929 + 19.51X_4$. **Troglodytidae**: rock wren (*Salpinctes obsoletus*; 0.59), $Y = 44.83 – 2.28X_{26}$; canyon wren (*Catherpes mexicanus*; 0.43), $Y = 53.12 + 12.14X_6 – 0.66X_{20}$; Carolina wren (*Thryothorus ludovicianus*; 0.61), $Y = 55.10 + 0.29X_{27} – 0.46X_{30} – 2.65X_{31}$; Bewick's wren (*Thryomanes bewickii*; 0.49), $Y = 39.63 – 12.15X_2 + 2.51X_7 – 1.75X_{17} + 0.30X_{27} + 0.41X_{28} – 0.31X_{34}$; house wren (*Troglodytes aedon*; 0.40), $Y = –7.28 + 5.75X_{10} + 7.81X_{13} – 0.78X_{30}$; sedge wren (*Cistothorus platensis*; 0.50), $Y = –4.71$. **Sylviidae**: blue-gray gnatcatcher (*Polioptila caerulea*; 0.62), $Y = 15.64 + 13.90X_1 + 3.59X_8 – 2.81X_9 – 5.28X_{10} + 0.39X_{20} – 1.82X_{24} + 0.21X_{25}$. **Turdidae**: eastern bluebird (*Sialia sialis*; 0.72), $Y = –156.4 –16.35X_2 – 3.35X_9 – 5.68X_{10} + 1.00X_{20} + 3.35X_{26} – 0.57X_{30}$; mountain bluebird (*S. currucoides*; 0.50), $Y = –5.23$; wood thrush (*Hylocichla mustelina*; 0.58), $Y = 53.40 + 15.37X_4 + 1.97X_{16} – 0.34X_{18} – 2.65X_{19}$; American robin (*Turdus migratorius*; 0.34), $Y = 4.65 – 3.86X_{13} + 12.45\ X_{10}$. **Mimidae**: gray catbird (*Dumetella carolinensis*; 0.40), $Y = –14.34 – 2.08X_9 + 3.92X_{12} + 0.62X_{23} + 0.12X_{34}$; northern mockingbird (*Mimus polyglottos*; 0.48), $Y = 126.6 – 43.59X_5 + 7.32X_7 + 15.22X_8 – 2.36X_{16} – 3.83X_{17}$; brown thrasher (*Toxostoma rufum*; 0.36), $Y = 5.22 – 3.61X_1 – 2.057X_9 – 4.38X_{12} + 0.14X_{18} + 1.64X_{23} + 0.46X_{28} – 0.91X_{29}$; curve-billed thrasher (*T. curvirostre*; 0.57), $Y = 85.38 + 4.69X_{14} + 2.46X_{30} – 0.63X_{32} – 9.75X_{33}$. **Sturnidae**: European starling (*Sturnus vulgaris*; 0.45), $Y = –23.33 – 5.21X_3 – 5.55X_6 + 3.09X_8 + 7.10X_{13} + 0.57X_{28}$. **Bombycillidae**: cedar waxwing (*Bombycilla cedrorum*;

0.38), $Y = 28.73 - 1.02X_{31}$. **Parulide**: northern parula (*Parula americana*; 0.61), $Y = -153.1 + 5.43X_1 + 11.51X_3 + 14.51X_4 + 3.75X_8 - 2.11X_{14} - 3.57X_{24} + 6.15X_{26} + 6.33X_{33}$; yellow warbler (*Dendroica petechia*; 0.38), $Y = -22.64 - 6.69X_6 - 2.30X_9 + 4.78X_{13} + 0.55X_{28}$; yellow-throated warbler (*D. dominica*; 0.50), $Y = -0.66 + 2.37X_1 - 0.25X_{18} + 0.68X_{23}$; pine warbler (*D. pinus*; 0.74), $Y = 5.58 + 3.86X_1 + 15.45X_3 - 0.50X_{18}$; prairie warbler (*D. discolor*; 0.55), $Y = 118.7 - 3.73X_{24} - 0.56X_{25}$; black-and-white warbler (*Mniotilta varia*; 0.53), $Y = 26.18 + 2.37X_{13} - 2.12X_8 - 2.62X_9 + 0.33X_{20} - 1.86X_{24}$; American redstart (*Setophaga ruticilla*; 0.49), $Y = -4.74 + 3.57X_1$; prothonotary warbler (*Protonotaria citrea*; 0.49), $Y = -4.21 + 4.87X_1 + 19.12X_4 + 3.23X_8 + 10.16X_{12}$; worm-eating warbler (*Helmitheros vermivorum*; 0.50), $Y = -15.71 + 9.18X_2 + 2.18X_{23}$; ovenbird (*Seiurus aurocapilla*; 0.49), $Y = -18.48 + 17.19X_1 + 23.36X_2 + 16.42X_8$; Louisiana waterthrush (*S. motacilla*; 0.41), $Y = -2.53 + 4.39X_1 + 7.04X_3$; Kentucky warbler (*Oporornis formosus*; 0.45), $Y = 17.96 + 3.84X_1 + 13.13X_4 + 3.86X_{12} - 0.19X_{18} - 0.63X_{21}$; common yellowthroat (*Geothlypis trichas*; 0.51), $Y = 7.10 + 1.48X_1 + 4.65X_2 + 14.14X_{12} - 0.71X_{21} + 0.13X_{34}$; hooded warbler (*G. nelsoni*; 0.49), $Y = 4.12 - 0.40X_{18}$; yellow-breasted chat (*Icteria virens*; 0.70), $Y = 21.38 - 8.05X_8 - 2.55X_{14} + 1.61X_{16} - 0.20X_{18} + 1.47X_{21} + 1.38X_{23} - 0.91X_{30}$. **Thraupidae**: summer tanager (*Piranga rubra*; 0.65), $Y = 62.06 + 5.32X_1 - 2.58X_9 - 8.50X_{10} - 2.01X_{17} - 0.13X_{32}$; scarlet tanager (*P. olivacea*; 0.64), $Y = 83.64 + 6.46X_1 + 8.77X_2 + 8.77X_3 + 6.21X_{12} - 0.37X_{32} - 3.87X_{33}$; western tanager (*P. ludoviciana*; 0.50), $Y = -6.33$. **Emberizidae**: spotted towhee (*Pipilo maculates*; 0.50), $Y = 32.70 + 7.83X_6 - 1.83X_{33}$; eastern towhee (*P. erythrophthalmus*; 0.48), $Y = -3.47$; canyon towhee (*P. fuscus*; 0.75), $Y = 84.77 - 1.07X_{27}$; Cassin's sparrow (*Aimophila cassinii*; 0.70), $Y = -119.7 + 5.09X_6 + 0.59X_{14} + 4.00X_{24} - 0.25X_{27}$;

Bachman's sparrow (*A. aestivalis*; 0.56), $Y = 0.03 + 16.54X_{11} – 0.21X_{18}$; rufous-crowned sparrow (*A. ruficeps*; 0.48), $Y = –45.5818 – 2.71X_1 – 8.59X_3 – 4.04X_6 + 6.11X_{12} + 0.20X_{20} – 1.51X_{21} + 1.11X_{24} – 0.11X_{25} + 1.20X_{28}$; chipping sparrow (*Spizella passerina*; 0.41), $Y = 6.13 + 8.80X_2 + 0.33X_{20} + 1.20X_{23} – 0.77X_{26} – 0.29X_{27}$; field sparrow (*S. pusilla*; 0.60), $Y = 10.76 – 5.02X_9 – 4.55X_{10} + 0.75X_{15} – 7.23X_{17} + 0.41X_{18} + 0.74X_{20} + 4.16X_{31} – 0.30X_{32}$; lark sparrow (*Chondestes grammacus*; 0.46), $Y = 0.37 – 4.35X_2 – 2.54X_9 – 7.11X_{10} + 0.78X_{28} – 0.26X_{34}$; black-throated sparrow (*Amphispiza bilineata*; 0.50), $Y = –6.33$; lark bunting (*Calamospiza melanocorys*; 0.85), $Y = –224.8 + 8.51X_8 + 3.14X_9 + 7.31X_{31}$; grasshopper sparrow (*Ammodramus savannarum*; 0.65), $Y = –49.84 – 4.24X_1 – 26.86X_2 – 5.10X_6 – 11.53X_{10} + 0.58X_{21} + 0.25X_{25} + 0.35X_{27}$; Henslow's sparrow (*Ammodramus henslowii*; 0.59), $Y = 52.41 + 5.13X_7 – 2.22X_{22}$; Lincoln's sparrow (*Melospiza lincolnii*; 0.50), $Y = –6.33$. **Cardinalidae**: northern cardinal (*Cardinalis cardinalis*; 0.75), $Y = –76.46 + 4.65X_{19} – 0.25X_{34}$; rose-breasted grosbeak (*Pheucticus ludovicianus*; 0.49), $Y = –7.9206 + 12.33X_4 + 0.96X_{23}$; black-headed grosbeak (*P. melanocephalus*; 0.50), $Y =$ -6.33; blue grosbeak (*Passerina caerulea*; 0.97), $Y = 2.17 + 2.95X_1 – 3.73X_2 – 0.99X_9 – 11.50X_{10} – 0.38X_{30}$; lazuli bunting (*P. amoena*; 0.55), $Y = 11.59 + 12.59X_3 – 80.15X_{26} + 79.4X_{33}$; indigo bunting (*P. cyanea*; 0.71), $Y = 83.04 + 20.86X_1 + 5.86X_{13} – 3.00X_{31}$; painted bunting (*P. ciris*; 0.57), $Y = –84.9511 – 10.40X_2 + 9.24X_6 – 2.32X_9 – 6.23X_{10} – 7.92X_{13} – 2.57X_{14} + 3.60X_{19} + 0.42X_{27} + 2.44X_{28} – 0.35X_{34}$; dickcissel (*Spiza americana*; 0.48), $Y = –62.07 + 3.67X_8 – 5.92X_{10} – 0.79X_{14} + 0.39X_{18} + 3.51X_{33}$. **Icteridae:** red-winged blackbird (*Agelaius phoeniceus*; 0.51), $Y = 175.7 – 4.43X_1 – 7.19X_7 – 3.95X_{17} + 0.39X_{18} – 0.65X_{27}$; eastern meadowlark (*Sturnella magna*; 0.80), $Y = –64.6 + 7.61X_8 + 2.38X_{16} + 1.14X_{18} + 0.62X_{20} –0.45X_{32} + 3.80X_{33}$; western meadowlark (*S. neglecta*; 0.67),

$Y = 68.38 – 7.39X_3 – 2.10X_{19} + 1.20X_{23} + 0.24X_{32} – 0.41X_{34}$; yellow-headed blackbird (*Xanthocephalus xanthocephalus*; 0.46), $Y = –44.04 + 1.67X_9 + 1.33X_{24}$; Brewer's blackbird (*Euphagus cyanocephalus*; 0.50), $Y = 38.14 – 2.08X_{33}$; common grackle (*Quiscalus quiscula*; 0.53), $Y = 12.27 – 2.68X_1 + 4.44X_8 + 3.59X_{31} – 3.84X_{17} – 3.25X_6$; great-tailed grackle (*Q. mexicanus*; 0.39), $Y = –5.42 – 5.02X_6 + 3.69X_8 + 8.16X_{10} + 0.17X_{32}$; brown-headed cowbird (*Molothrus ater*; 0.48), $Y = –1.13 – 8.41X_2 + 1.38X_{23}$; orchard oriole (*Icterus spurious*; 0.29), $Y = –19.57 – 5.39X_6 + 0.94X_7 – 8.97X_{10} + 0.80X_{17} + 0.83X_{23} – 0.25X_{28}$; Baltimore oriole (*I. galbula*; 0.34), $Y = –45.5818 – 2.71X_1 – 8.59X_3 – 4.04X_6 + 6.11X_{12} + 0.20X_{20} – 1.51X_{21} + 1.11X_{24} – 0.11X_{25} + 1.20X_{28}$; Bullock's oriole (*I. bullockii*; 0.71), $Y = 140.1 + 3.16X_{14} – 0.51X_{18} – 79.61X_{26} + 71.63X_{33} – 0.54X_{34}$. **Fringillidae:** house finch (*Carpodacus mexicanus*; 0.46), $Y = –6.72 + 1.52X_8 + 50.86X_{10} + 5.15X_{13}$; red crossbill (*Loxia curvirostra*; 0.50), $Y = –6.52 + 20.94X_4 + 22.39X_{11}$; lesser goldfinch (*Carduelis psaltria*; 0.50), $Y = 94.74 + 11.71X_6 + 8.7X_{10} + 9.47X_{12} – 1.15X_{20}$; American goldfinch (*Carduelis tristis*; 0.41), $Y = 19.22 – 3.81X_6 + 5.46X_{13} – 3.00X_{21} + 1.92X_{29}$. **Passeridae:** house sparrow (*Passer domesticus*; 0.46), $Y = –4.32 – 4.47X_6 + 6.27X_8 + 0.38X_{18} – 0.20X_{32}$.

---

[a] Variables are: (X_1) deciduous forest; (X_2) mixed forest; (X_3) evergreen forest; (X_4) woody wetland; (X_5) emergent herbaceous wetland; (X_6) shrubland; (X_7) grassland; (X_8) pasture/hay; (X_9) cropland; (X_{10}) developed; (X_{11}) barren; (X_{12}) water; (X_{13}) distance to water; (X_{14}) May temperature; (X_{15}) May soil temperature at 10 cm; (X_{16}) May rain; (X_{17}) May solar radiation; (X_{18}) May wind speed; (X_{19}) May barometric pressure; (X_{20}) May percent humidity; (X_{21}) June temperature; (X_{22}) June soil temperature at 10 cm; (X_{23})

June rain; (X_{24}) June solar radiation; (X_{25}) June wind speed; (X_{26}) June barometric pressure; (X_{27}) June percent humidity; (X_{28}) July temperature; (X_{29}) July soil temperature at 10 cm; (X_{30}) July rain; (X_{31}) July solar radiation; (X_{32}) July wind speed; (X_{33}) July barometric pressure; and (X_{34}) July percent humidity.

Sampling requirements for predicting occurrence of Oklahoma breeding birds

Dennis G Siegfried[a,1] and Gary D. Schnell[a, b, *]

[a]Department of Zoology, University of Oklahoma, Norman, OK 73019, USA

[b]Sam Noble Museum of Natural History, 2401 Chautauqau Avenue, University of Oklahoma, Norman, OK 73072, USA

ABSTRACT

The continued loss of species has encouraged the development of predictive models using surrogate information; particular attention has focused on use of environmental variables. Large data sets make it possible to estimate the effect of splitting these sets by showing the effect that sample size has on the development of predictive models for species that are both widespread and sparse. The Oklahoma breeding-bird atlas includes a data set for occurrences (presences and absences) of 209 avian species in 562 sites, each 5.00 × 5.88 km in size. We divided this data set into one 100-site test set and nine 50-site training subsets. The nine training subsets were progressively added together to develop training sets of 50, 100, 150, 200, 250, 300, 350, 400, and 450 sites. Using these training sets and 34 environmental variables, we employed stepwise logistic regression to develop predictive models for 179 of the 209 species that occurred in the 100-site test set. Model performance was based on overall accuracy at predicting species occurrences in a 100-site test set based on average τ_p scores. Logistic regression was more accurate (larger τ_p) when sparsely and widely distributed species were included in the analyses. Model accuracy improved as up to 250 sites were added to the initial 50-site training set. Model accuracy remained relatively consistent when 300 or more training sites were used to

develop species models. While additional sites provide actual data that would be useful for other purposes, our results suggest that for an area the size of Oklahoma little is gained in terms of predictive ability by sampling more than 300 sites.

Keywords: Bird-habitat relationships; Geographic information systems; Model performance; Logistic regression; Training-set size

---

* *Corresponding author*. Tel.: +1 405 325 5050; fax: +1 405 325 7690

E-mail addresses: dsiegfri@snu.edu (D.G Siegfried), gschnell@ou.edu (G.D. Schnell).

[1] Present address: Department of Biology, Southern Nazarene University, 6729 NW 39th Expressway, Bethany, OK 73008, USA.

1. Introduction

Prediction models for species distributions have received increased attention in the past decade (Scott et al., 2002; Lomolino and Heany, 2004). The development and application of such models have been enhanced by the advances in biogeographical approaches that incorporate geographic information systems (GIS) to handle the spatial requirements of such models (Yuan, 1999; Burrough, 2001). Recent focus has been on model development from data sets that include occurrences (both presences and absences, as defined by Angermeier et al., 2002; e.g., Austin et al., 1996; van Horseen et al., 1999; Pearce and Ferrier, 2000; Venier et al., 2004) or presence-only data (e.g., Peterson, 2001; Anderson et al., 2002; Peterson et al., 2002; Illoldi-Rangel et al., 2004), an example of the latter being information based on museum specimens. Logistic regression often is employed with occurrence data (Hosmer and Lemeshow, 2000).

Associations of species occurrences with various land-cover types and climatic factors are used to develop models based on a training subset (e.g., 50%) of the overall data set. These models are then applied to a test set or independent data sets and evaluated for their ability to predict the occurrence of a species (Guisan and Zimmerman, 2000).

For conservation efforts, only limited data may be available to comprise a training data set. For example, funding may be unavailable for extensive sampling, there may be insufficient time in which to collect data, and/or the species of interest simply may be so sparsely distributed or secretive that it is not possible to establish an extensive database. This raises the issue of how much data are required in a training set to provide accurate predictions of occurrence of a species in an area of interest. We have addressed this issue using occurrence data from the *Oklahoma Breeding Bird Atlas* (Reinking, 2004). We developed logistic-regression models using nine different-sized training sets to predict species occurrences for 179 species in a 100-site test set. Our goal was to determine when additional data no longer provided substantial improvement in model predictions.

2. Methods

2.1. Occurrence bird data

We initially evaluated occurrence data for 209 bird species from the *Oklahoma Breeding Bird Atlas* for 562 5.00 × 5.88 km blocks (Fig. 1a). Subsequently, given species occurrences at the initial 50-site training set (see below), analyses were restricted to 179 species (Appendix 1). Data for the atlas were gathered over a 5-year period, with observers visiting each of the 562 blocks during the breeding season to record all bird

species present. The atlas blocks initially were selected using a stratified-random procedure to ensure that sites were not unduly clumped geographically (Reinking, 2000).

2.1.1. Test and training sets

From the 562 atlas blocks, a test set of 100 sites (Fig. 1b) was selected using a stratified-random procedure so that the sites selected were not by chance geographically clumped. From the remaining 462 sites, nine stratified-random selections of 50 (without replacement) were placed in subsets used to develop training sets (Fig. 1c). We generated training sets of 50, 100, 150, 200, 250, 300, 350, 400, and 450 sites by starting with the initial 50-site training set and in turn adding 50-site subsets to the previously formed training sets. For example, to the initial subset of 50 sites we added the second subset of 50 sites to form the 100-site training set; a third subset was then added to form the 150-site training set.

2.2. Variables

We used 34 environmental variables (land-cover and climatic variables) to develop prediction models of occurrences for each of the 179 bird species. Twelve land-cover variables were based on an initial 19 land-cover types defined for Oklahoma by United States Geological Survey archive and found in the 1992 land-cover image (USGS, 2002). We combined four human-use types into one variable (light intensity residential, high intensity residential, commercial/industrial/transportation, and urban/recreational grasses were subsumed under “developed”), as we did for three cultivated types (row crops, small grains, and fallow were designated as “crops”), and for three bare-ground types

(bare rock/gravel/clay, quarries/strip mines/gravel pits, and transitional were combined as "barren"). In ArcGIS (ESRI, 2004) the map polygons were reclassified as representing 1 of the 12 land-cover types: (1) deciduous forest; (2) mixed forest; (3) evergreen forest; (4) woody wetland; (5) emergent herbaceous wetland; (6) shrubland; (7) grassland; (8) pasture/hay; (9) cropland; (10) developed; (11) barren; and (12) water. Within each atlas block we calculated the proportion of the block covered by each land-cover type. The 13th variable was the distance in meters from the center of the block to the nearest river or stream as identified in the Oklahoma Digital Elevation Model hydrological network (USGS, 1998). We obtained measurements for this variable using Analyst Tools in ArcGIS (Beyer, 2004). In order to maintain accurate distance measurements, all ArcGIS layers were placed in Universal Transverse Mercator projection, zone 14, based on the North American Datum 1983.

The climatic data for 1997 to 2001 covering May, June, and July were gathered through the Oklahoma Mesonet, a system of 119 weather stations in Oklahoma (Brock et al., 1995). The five years of data were averaged by month to generate point data for each station. For each of the three months (May, June, and July), seven variables were selected: temperature (^{o}C); soil temperature (^{o}C); rainfall (cm); solar radiation (megajoules/m^2); wind speed (kph); barometric pressure (millibars); and humidity (%). Using the ordinary kriging method in the Geostatistical Analyst of ArcGIS (ESRI, 2004), the values for each of the 21 resulting climatic variables were estimated, from the point values, for all points on a surface that covered Oklahoma (van Horseen et al., 1999).

2.3. Species models based on logistic regression

Individual logistic regressions were developed for each of the 209 species using each of the nine training sets. We used step wise procedures with a 0.05 *p*-value (Hosmer and Lemeshow, 2000) to predict occurrences of each species at the 100 sites in the test set. Models were not produced for some species using the 50-site training set because they were either absent from or present in all localities in the set. These species could not be evaluated in all subsets, so they were removed. As a result, 179 of the 209 species (Appendix 1) were evaluated. The response value for each of the species at each of the 100 test sites was counted as either present (> 0.5) or absent (< 0.5).

2.4. Sample evaluations

The logistic-regression models for the 179 species were evaluated based on the extent to which models correctly predicted the occurrence of the species. As a summary of model prediction efficiency, we used a measure of proportional change in error, tau-p (τ_p):

$$\tau_p = \frac{E_{wo} - E_w}{E_{wo}},$$

where E_{wo} refers to error without the model and E_w to error with the model (Menard, 2002). E_w is the sum of incorrectly predicted presences and incorrectly predicted absences. E_{wo} is 2 times the number of presences times the number of absences divided by the size of the sample (N = 100 for our test set). τ_p can vary from 1 to 1-[N^2/2(N-1)], with 0 being no different from chance, a negative value being worse than chance, and a positive value being better than chance (Menard, 2002). The overall performance of each training set was evaluated as the average value of τ_p for the 100 test sites for the 179 species.

In addition to analyzing all 179 species, we also assessed subsets of those species from which sparsely distributed and widely distributed species had been removed. When the 25 species that occurred in only 1 or in 99 of the test sites were removed, the remaining species constituted subgroup A. When those in 5 or less or 95 or more of the test sites (35 species) were removed, the 119 remaining species are referred to subgroup B. We then removed those in 10 or fewer and 90 or more test sites; the remaining 94 species made up subgroup C.

3. Results

As indicated by average τ_p -values (Fig. 2; Table 1), model accuracy increased as training-set size increased from 50 to 300 sites. The curves asymptoted and the standard deviations stabilized (Table 1) when 300 to 350 sites were in the training set. Except for subgroup C, standard deviations of τ_p tended to be higher for training sets with relatively few sites (Table 1). When all sites were considered, as well as for the three site subgroups, there was a slight decrease in accuracy from the 350-site training set to the 400-site training set (Fig. 2). This may be an example of reduced accuracy by increasing information from more variables producing models that are too specific to the sample sites (Stockwell and Peterson, 2002). Another possibility is that the 400-site training set did not improve accuracy by chance because the 50-site subset that was added to the 350-site training set did not provide sufficient new information to improve the average τ_p-value.

For the various-sized training subsets the average τ_p-values for all species, for subgroup A, and for subgroup B were not appreciably different (Fig. 2). However, when

species in 10 or fewer and in 90 or more of the test sites were removed (i.e. subgroup C), there was a substantial drop in average τ_p-values for all sizes of training subsets (Fig. 2). Logistic regression produced more accurate results when relatively sparsely and widely distributed species were included.

4. Discussion

Using the 300-site training subset to predict the 100-site test set provided a balance between model performance and necessary effort for Oklahoma, which has an area of 112 500 km^2. This fits within the training/testing proportions of other studies that were much larger in extent (Stockwell and Peterson, 2002; Venier et al., 2004). Stockwell and Peterson (2002) evaluated Mexico (1 973 000 km^2) for all bird species, while Venier et al. (2004) studied 10 species of warblers in the Great Lakes basin (80 0000 km^2). For Mexico, Stockwell and Peterson (2002) derived training subsets of 2000 points and a testing set of 1000 points from museum specimens. Venier et al. (2004) divided their data set of 1302 points into roughly five equal-sized groups, using four as training subsets and one as the test set.

When many species are involved over such large areas, it is unavoidable that a considerable effort will be needed to provide accurate predictions of species occurrences. Using 300 sites in Oklahoma allowed useful models to be developed for most avian species, while also meeting the needs of surveyors in terms of costs, time, and model applicability. Beyond a 300-site training set, we did not gain appreciably in terms of prediction ability. Prediction of occurrences is only one purpose for which data from the Oklahoma breeding-bird atlas can be used and, for other uses, having actual occurrence

data for more than 300 sites may well prove beneficial. This likely would be the case for breeding-bird atlases for other regions, as well as for other programs that generate occurrence data sets.

Acknowledgements

Access to the Oklahoma breeding bird atlas data set was kindly provided by Dan L. Reinking and the Sutton Avian Research Center. Support for the first author was provided by George Miksch Sutton Scholarships in Ornithology and a Robert E. and Mary B. Sturgis Scholarship, as well as by a grant from the University of Oklahoma Graduate Student Senate. We thank Jeffrey F. Kelly, William J. Matthews, Thomas S. Ray, and May Yuan for serving as committee members of the first author and for helpful comments on earlier drafts of the manuscript. This article is a portion of a dissertation submitted by the first author in partial fulfillment of requirements for a Ph.D. in the Department of Zoology, University of Oklahoma.

References

Anderson, R.P., Gomez-Laverde, M., Peterson, A.T., 2002. Geographical distributions of spiny pocket mice in South America: insights from predictive models. Global Ecology and Biogeography 11, 131–141.

Anderson, R.P., Lew, D., Peterson, A.T., 2003. Evaluating predictive models of species' distributions: criteria for selecting optimal models. Ecological Modelling 162, 211–232.

Angermeier, P.L., Krueger, K.L., Dolloff, C.A., 2002. Discontinuity in stream-fish distributions: implications for assessing and predicting species occurrence. In: Scott, J.M., Heglund, P.J., Morrison, M.L., Haufler, J.B., Raphael, M.G., Wall, W.A., Samson F.B. (Eds.), Predicting Species Occurrences: Issues of Accuracy and Scale. Island Press, Washington DC, pp.519–527.

Austin, G.E., Thomas, C.J., Houston, D.C., Thompson, D.B.A., 1996. Predicting the spatial distribution of buzzard *Buteo buteo* nesting areas using a geographical information system and remote sensing. Journal of Applied Ecology 33, 1541–1550.

Beyer, H. L., 2004. Hawth's analysis tools for ArcGIS. http://www.spatialecology.com/htools. September 23, 2005.

Brock, F.V., Crawford, K.C., Elliot, R.L., Cuperus, G.W., Stadler, S.J., Johnson, H.L., Eilts, M.D., 1995. The Oklahoma Mesonet: a technical overview. Journal of Atmospheric and Oceanic Technology 12, 5–19

Burrough, P.A., 2001. GIS and geostatistics: essential partners for spatial analysis. Environmental and Ecological Statistics 8, 361–377.

ESRI, 2004. ArcGIS version 9.0. Environmental Systems Research Institute, Redlands, California.

Guisan, A., Zimmermann, N.E., 2000. Predictive habitat distribution models in ecology. Ecological Modeling 135, 147–186.

Hosmer, D.W., Lemeshow, S., 2000. Applied Logistic Regression. John Wiley and Sons, Hoboken, New Jersey, 392 pp.

Illoldi-Rangel, P., Sanchez-Cordero, V., Peterson, A.T., 2004. Predicting distributions of Mexican mammals using ecological niche modeling. Journal of Mammalogy 85, 658–662.

Lomolino, M.V., Heany, L.R. (Eds.), 2004. Frontiers in Biogeography: New Directions in the Geography of Nature. Sinauer Associates, Sunderland, Massachusetts, 419 pp.

Menard, S., 2002. Applied Logistic Regression Analysis. Sage Publications, Thousand Oaks, California, 111 pp.

Pearce, J., Ferrier, S., 2000. Evaluating the predictive performance of habitat models developed using logistic regression. Ecological Modelling 133, 225–245.

Peterson, A. T., 2001. Predicting species' geographical distributions based on ecological niche modelling. Condor 103, 599–605.

Peterson, A.T., Ortega-Huerta, O., Bartley, J., Sanchez-Cordero, V., Soberon, J., Buddemeier, R.H., Stockwell, D.R.B., 2002. Future projections for Mexican faunas under global climate change scenarios. Nature 416, 626–629.

Reinking, D.L., 2000 Oklahoma Breeding Bird Atlas Handbook. George M. Sutton Avian Research Center, Bartlesville, Oklahoma, USA.

Reinking, D.L. (Ed.), 2004. Oklahoma Breeding Bird Atlas. University of Oklahoma Press, Norman, Oklahoma, 519 pp.

Scott, J.M., Heglund, P.J., Morrison, M.L., Haufler, J.B., Raphael, M.G., Wall, W.A., Sampson, F.B. (Eds.), 2002, Predicting Species Occurrences: Issues of Accuracy and Scale. Island Press, Washington DC, 868 pp.

Stockwell, D.R.B., Peterson, A.T., 2002. Effects of sample size on accuracy of species distribution models. Ecological Modeling 148, 1–13.

USGS, 1998. US Geological Survey DEM 7.5 Quadrangle. US Geological Survey, Reston, Virginia, USA.

USGS, 2002. Oklahoma land-cover data set. http://edcftp.cr.usgs.gov/pub /data/landcover/states/. September 23, 2005.

van Horseen, P.W., Schot, P.P., Barendregt, A., 1999. A GIS-based plant prediction model for wetland ecosystems. Landscape Ecology 14, 253–265.

Venier, L.A., Pearce, J., McKee, J.E., McKenney, D.W., Niemi, G.J., 2004. Climate and satellite-derived land cover for predicting breeding bird distribution in the Great Lakes Basin. Journal of Biogeography 31, 315–331.

Yuan, M., 1999. Use of a three-domain representation to enhance GIS support for complex spatiotemporal queries. Transactions in GIS 3, 137–159.

Table 1. Average values for τ_p ± SD (range) for different-sized training sets for: all species; subgroup A (species at 1 or 99 sites removed); subgroup B (species at ≤ 5 or ≥ 95 sites removed); and subgroup C (species at ≤ 10 or ≥ 90 sites removed).

No. sites in training set	All species (179 species)	Subgroup		
		A (154 species)	B (119 species)	C (94 species)
50	0.40 ± 0.23 (-1.04 to 1.00)	0.38 ± 0.25 (-1.04 to 1.00)	0.38 ± 0.20 (-0.22 to 1.00)	0.36 ± 0.19 (-0.10 to 0.75)
100	0.41 ± 0.22 (-1.30 to 0.91)	0.39 ± 0.23 (-1.30 to 0.91)	0.39 ± 0.19 (-0.25 to 0.91)	0.37 ± 0.19 (-0.25 to 0.81)
150	0.41 ± 0.25 (-1.04 to 1.00)	0.42 ± 0.23 (-1.04 to 1.00)	0.42 ± 0.18 (-0.01 to 1.00)	0.40 ± 0.19 (-0.01 to 0.86)
200	0.44 ± 0.18 (-0.28 to 0.91)	0.43 ± 0.19 (-0.27 tp 0.91)	0.43 ± 0.19 (-0.01 to 0.91)	0.41 ± 0.19 (-0.01 to 0.91)
250	0.44 ± 0.16 (-0.02 to 0.89)	0.44 ± 0.18 (-0.02 to 0.89)	0.43 ± 0.18 (0.00 to0.82)	0.41 ± 0.18 (0.00 to 0.76)
300	0.45 ± 0.17 (-0.03 to 0.91)	0.45 ± 0.18 (-0.03 to 0.91)	0.44 ± 0.19 (-0.01 to 0.91)	0.41 ± 0.19 (-0.01 to .086)
350	0.45 ± 0.16 (0.02 to 0.91)	0.45 ± 0.17 (0.02 to 0.91)	0.44 ± 0.18 (0.02 to 0.91)	0.42 ± 0.18 (0.01 to 0.86)
400	0.45 ± 0.16 (-0.04 to 0.82)	0.45 ± 0.17 (-0.04 to 0.82)	0.44 ± 0.18 (-0.04 to 0.82)	0.41 ± 0.18 (-0.04 to 0.81)
450	0.46 ± 0.17 (-0.04 to 1.00)	0.46 ± 0.18 (-0.04 to 1.00)	0.45 ± 0.18 (-0.04 to 0.82)	0.42 ± 0.19 (-0.04 to 0.73)

Fig. 1 - Locations across state of Oklahoma of: (a) 562 breeding bird atlas sites; (b) 100 sites used to test the predictions of logistic-regression models; and (c) 450 training sites used to develop the logistic regressions, where first 50 were used to develop models for the 50-site training set, second 50 was added to first 50 to develop models for the 100-site training set, third 50 was added to the 100-site training set to develop models for the 150-site training set, etc.

Fig. 2 - Average values of τ_p for different-sized training subsets: all species; subgroup A (species at 1 or 99 sites removed); subgroup B (species at ≤ 5 or ≥ 95 sites removed); and subgroup C (species at ≤ 10 or ≥ 90 sites removed). Number of species and subgroups indicated in parentheses in legend.

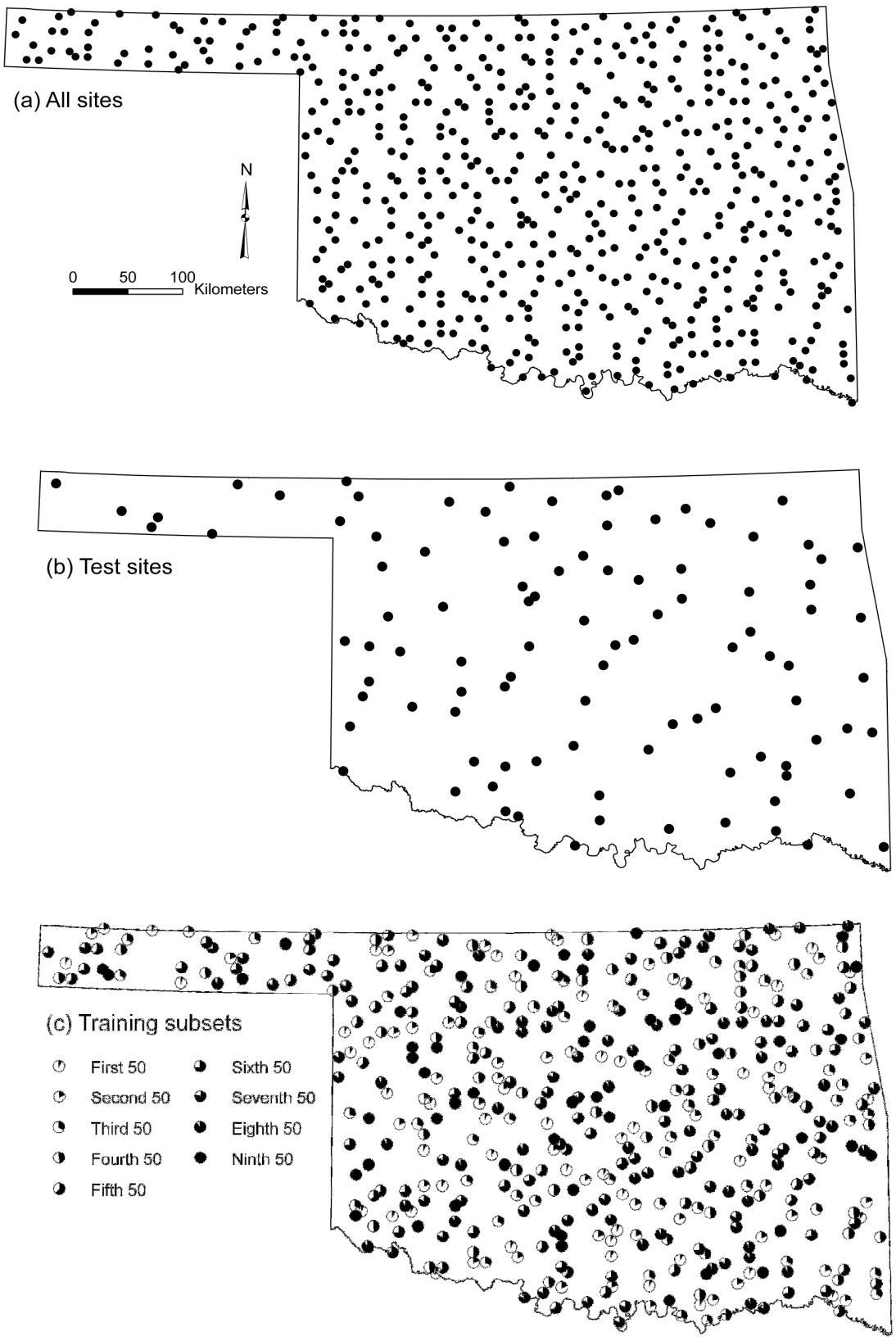
(a) All sites
N
0
50
100
Kilometers
(b) Test sites
(c) Training subsets
First 50
Second 50
Third 50
Fourth 50
Fifth 50
Sixth 50
Seventh 50
Eighth 50
Ninth 50

Figure 1

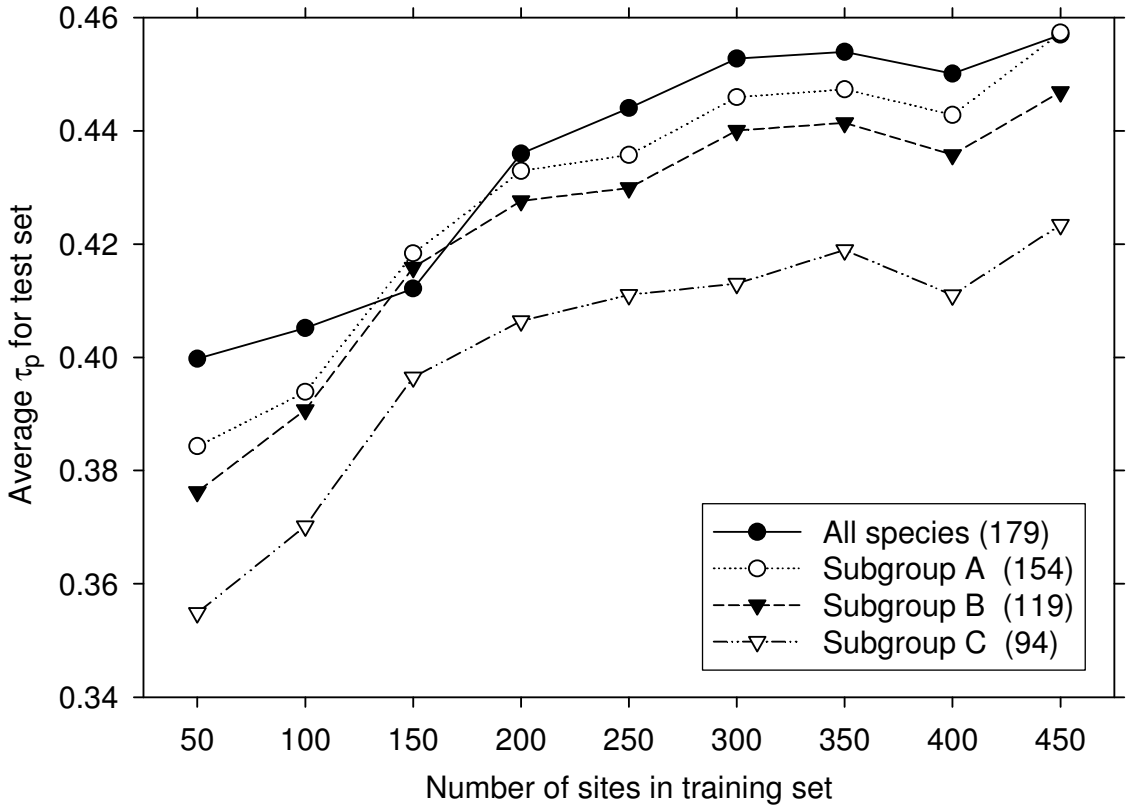
0.46
0.44
0.42
0.40
0.38
0.36
0.34
Average τ_p for test set
50
100
150
200
250
300
350
400
450
Number of sites in training set
All species (179)
Subgroup A (154)
Subgroup B (119)
Subgroup C (94)

Figure 2

Appendix 1 – The 179 species for which logistic-regression models were generated.

Podicipedidae: pied-billed grebe (*Podilymbus podiceps*). **Pelecanidae**: double-crested coromorant (*Phalacrocorax auritus*). **Ardeidae**: American bittern (*Botaurus lentiginosus*); great blue heron (*Ardea herodias*); great egret (*A. alba*); snowy egret (*Egretta thula*); little blue heron (*E. caerulea*); cattle egret (*Bubulcus ibis*); green heron (*Butorides virescens*); black-crowned night-heron (*Nycticorax nycticorax*); yellow-crowned night-heron (*N. violacea*). **Threskiornithidae**: white-faced ibis (*Eudocimus albus*). **Cathartidae**: black vulture (*Coragyps atratus*); turkey vulture (*Cathartes aura*). **Anatidae**: Canada goose (*Branta canadensis*); wood duck (*Aix sponsa*); mallard (*Anas platyhynchos*); blue-winged teal (*A. discors*); northern shoveler (*A. clypeata*); northern pintail (*A. acuta*); green-winged teal (*A. crecca*); redhead (*Aythya americana*); ring-necked duck (*A. collaris*); hooded merganser (*Lophodytes cucullatus*); ruddy duck (*Oxyura jamaicensis*). **Accipitridae**: osprey (*Pandion haliaetus*); Mississippi kite (*Ictinia mississippiensis*); bald eagle (*Haliaeetus leucocephalus*); northern harrier (*Circus cyaneus*); Cooper's hawk (*Accipiter cooperii*); red-shouldered hawk (*Buteo lineatus*); broad-winged hawk (*B. platypterus*); Swainson's hawk (*B. swainsoni*); red-tailed hawk (*B. jamaicensis*); ferruginous hawk (*B. regalis*). **Falconidae**: American kestrel (*Falco sparverius*); prairie falcon (*F. mexicanus*). **Phasianidae**: ring-necked pheasant (*Phasianus colchicus*); greater prairie-chicken (*Tympanuchus cupido*); lesser prairie-chicken (*T. pallidicinctus*); wild turkey (*Meleagris gallopavo*). **Odontiphoridae**: scaled quail (*Callipepla squamata*); northern bobwhite (*Colinus virginianus*). **Rallidae**: sora (*Porzana carolina*); American coot (*Fulica americana*). **Charadriidae**: killdeer

(*Charadrius vociferous*). **Recurvirostridae**: black-necked stilt (*Himantopus mexicanus*); American avocet (*Recurvirostra americana*). **Scolopacidae**: spotted sandpiper (*Actitis macularia*); upland sandpiper (*Bartramia longicauda*); long-billed curlew (*Numenius americanus*). **Laridae**: least tern (*Sterna antillarum*). **Columbidae**: rock pigeon (*Columba livia*); white-winged dove (*Zenaida asiatica*); Inca dove (*Columbina inca*). **Cuculidae**: black-billed cuckoo (*Coccyzus erythropthalmus*); yellow-billed cuckoo (*C. americanus*); greater roadrunner (*Geococcyx californianus*). **Tytonidae**: barn owl (*Tyto alba*). **Strigidae**: eastern screech-owl (*Megascops asio*); great horned owl (*Bubo virginianus*); burrowing owl (*Athene cunicularia*); barred owl (*Strix varia*). **Caprimulgidae**: common nighthawk (*Chordeiles minor*); common poorwill (*Phalaenoptilus nuttallii*); chuck-will's-widow (*Caprimulgus carolinensis*); whip-poor-will (*C. vociferous*). **Apodidae**: chimney swift (*Chaetura pelagica*). **Trochilidae**: ruby-throated hummingbird (*Archilochus colubris*); black-chinned hummingbird (*A. alexandri*). **Alcedinidae**: belted kingfisher (*Ceryle alcyon*). **Picidae**: Lewis's woodpecker (*Melanerpes lewis*); red-headed woodpecker (*M. erythrocephalus*); golden-fronted woodpecker (*M. aurifrons*); red-bellied woodpecker (*M. carolinus*); ladder-backed woodpecker (*Picoides scalaris*); downy woodpecker (*P. pubescens*); hairy woodpecker (*P. villosus*); northern flicker (*Colaptes auratus*); pileated woodpecker (*Dryocopus pileatus*). **Tyrannidae**: western wood-pewee (*Contopus sordidulus*); eastern wood-pewee (*C. virens*); Acadian flycatcher (*Empidonax virescens*); willow flycatcher (*E. traillii*); eastern phoebe (*Sayornis phoebe*); Say's phoebe (*S. saya*); ash-throated flycatcher (*Myiarchus cinerascens*); great crested flycatcher (*M. crinitus*); Cassin's kingbird (*Tyrannus vociferans*); western kingbird (*T. verticalis*); eastern kingbird (*T.*

tyrannus); scissor-tailed flycatcher (*T. forficatus*); **Laniidae**: loggerheaded shrike (*Lanius ludovicianus*). **Vireonidae**: white-eyed vireo (*Vireo griseus*); Bell's vireo (*V. bellii*); yellow-throated vireo (*V. flavifrons*); warbling vireo (*V. gilvus*); red-eyed vireo (*V. olivaceus*). **Corvidae**: blue jay (*Cyanocitta cristata*); western scrub-jay (*Aphelocoma californica*); pinyon jay (*Gymnorhinus cyanocephalus*); black-billed magpie (*Pica hudsonia*); American crow (*Corvus brachyrhynchos*); fish crow (*C. ossifragus*); Chihuahuan raven (*C. cryptoleucus*); common raven (*C. corax*). **Alaudidae**: horned lark (*Eremophila alpestris*). **Hirundinidae**: purple martin (*Progne subis*); tree swallow (*Tachycineta bicolor*); northern rough-winged swallow (*Stelgidopteryx serripennis*); bank swallow (*Riparia riparia*); cliff swallow (*Petrochelidon pyrrhonota*); barn swallow (*Hirundo rustica*). **Paridae**: Carolina chickadee (*Poecile carolinensis*); tufted titmouse (*Baeolophus bicolor*). **Sittidae**: white-breasted nuthatch (*Sitta carolinensis*). **Troglodytidae**: rock wren (*Salpinctes obsoletus*); canyon wren (*Catherpes mexicanus*); Carolina wren (*Thryothorus ludovicianus*); Bewick's wren (*Thryomanes bewickii*); house wren (*Troglodytes aedon*); sedge wren (*Cistothorus platensis*). **Sylviidae**: blue-gray gnatcatcher (*Polioptila caerulea*). **Turdidae**: eastern bluebird (*Sialia sialis*); wood thrush (*Hylocichla mustelina*); American robin (*Turdus migratorius*). **Mimidae**: gray catbird (*Dumetella carolinensis*); northern mockingbird (*Mimus polyglottos*); brown thrasher (*Toxostoma rufum*); curve-billed thrasher (*T. curvirostre*). **Sturnidae**: European starling (*Sturnus vulgaris*). **Bombycillidae**: cedar waxwing (*Bombycilla cedrorum*). **Parulide**: northern parula (*Parula americana*); yellow warbler (*Dendroica petechia*); yellow-throated warbler (*D. dominica*); pine warbler (*D. pinus*); prairie warbler (*D. discolor*); black-and-white warbler (*Mniotilta varia*); American redstart (*Setophaga*

ruticilla); prothonotary warbler (*Protonotaria citrea*); Louisiana waterthrush (*Seiurus motacilla*); Kentucky warbler (*Oporornis formosus*); common yellowthroat (*Geothlypis trichas*); hooded warbler (*G. nelsoni*); yellow-breasted chat (*Icteria virens*). **Thraupidae**: summer tanager (*Piranga rubra*); scarlet tanager (*P. olivacea*). **Emberizidae**: canyon towhee (*Pipilo fuscus*); Cassin's sparrow (*Aimophila cassinii*); Bachman's sparrow (*A. aestivalis*); rufous-crowned sparrow (*A. ruficeps*); chipping sparrow (*Spizella passerina*); field sparrow (*S. pusilla*); lark sparrow (*Chondestes grammacus*); black-throated sparrow (*Amphispiza bilineata*); lark bunting (*Calamospiza melanocorys*); grasshopper sparrow (*Ammodramus savannarum*); Lincoln's sparrow (*Melospiza lincolnii*). **Cardinalidae**: northern cardinal (*Cardinalis cardinalis*); rose-breasted grosbeak (*Pheucticus ludovicianus*); blue grosbeak (*Passerina caerulea*); lazuli bunting (*P. amoena*); indigo bunting (*P. cyanea*); painted bunting (*P. ciris*); dickcissel (*Spiza americana*). **Icteridae:** red-winged blackbird (*Agelaius phoeniceus*); eastern meadowlark (*Sturnella magna*); western meadowlark (*S. neglecta*); yellow-headed blackbird (*Xanthocephalus xanthocephalus*); Brewer's blackbird (*Euphagus cyanocephalus*); common grackle (*Quiscalus quiscula*); great-tailed grackle (*Q. mexicanus*); brown-headed cowbird (*Molothrus ater*); orchard oriole (*Icterus spurious*); Baltimore oriole (*I. galbula*); Bullock's oriole (*I. bullockii*). **Fringillidae:** house finch (*Carpodacus mexicanus*); American goldfinch (*Carduelis tristis*). **Passeridae:** house sparrow (*Passer domesticus*).

Predicting occurrences of individual bird species and species richness using logistic regression and GARP: a comparative analysis

Dennis G Siegfried[a,1] and Gary D. Schnell[a, b, *]

[a]Department of Zoology, University of Oklahoma, Norman, OK 73019, USA

[b]Sam Noble Museum of Natural History, 2401 Chautauqau Avenue, University of Oklahoma, Norman, OK 73072, USA

ABSTRACT

Species-distribution models have been employed with increased frequency to predict species occurrences (both presences and absences) and sometimes species richness. Heuristically, as well as practically, it is useful to compare predictive abilities of commonly used procedures. Using intrinsic criteria, we evaluated the efficiency of stepwise logistic regression and two forms of GARP (genetic algorithm rule-set prediction) in predicting actual occurrences for 209 species included in the *Oklahoma Breeding Bird Atlas* using 34 environmental variables. We also summed predictions to estimate species richness for each of the 562 atlas blocks. Logistic regression developed models by selecting variables that best predicted distributions of each species. $GARP_{50:50}$ used a 0.5 cutoff similar to logistic regression and $GARP_{Best\ subset}$ the summed best subset to develop distribution models for each species, an approach employed by a number of previous investigations. Considering all individual species occurrences, logistic regression correctly predicted species occurrences 89.4% of the time, which was better than $GARP_{50:50}$ (76.6%) or $GARP_{Best\ subset}$ (70.3%). Comparisons were made for subgroups of species based on distribution extents and for subgroups of localities with low, moderate, and high species richnesses. For occurrences, logistic regression was the

better predictor irrespective of the extent of distributions. $GARP_{Best\ subset}$, which consistently overpredicted presences, not surprisingly was a better predictor of presences than either logistic regression or $GARP_{50:50}$ for all but the widespread species, for which it was not significantly different from logistic regression. $GARP_{50:50}$ was intermediate to logistic regression and $GARP_{Best\ subset}$ in predictive ability, better predicting presences than logistic regression and absences than $GARP_{Best\ subset}$. Logistic regression slightly overpredicted species richness for blocks with relatively low species richness and underpredicted it at sites with relatively high species richness, such that the average was close to the actual average species richness. GARP, for both implementations, routinely overpredicted species richness, with $GARP_{Best\ subset}$ on average predicting over twice the actual number of species. Summing results for logistic regression for a given site provided a good estimate of species richness, although results from this technique were not particularly informative when trying to estimate and compare relative species richnesses across localities. GARP substantially overestimated species richness for any given block and also did not produce particularly reliable estimates of relative interlocality differences in species richness. Thus, GARP in these two forms is less than an optimal choice for accurately predicting individual species distributions or site species richness when reputable occurrence data are available.

Keywords: Breeding bird atlas, Presence-absence data, Presence-only data, Occurrence data, Breeding bird atlas, Model comparison, Model performance, Logistic regression, Predictive ability, GARP, Genetic algorithm rule-set prediction

---

* *Corresponding author*. Tel.: +1 405 325 5050; fax: +1 405 325 7690

E-mail addresses: dsiegfri@snu.edu (D.G Siegfried), gschnell@ou.edu (G.D. Schnell).

[1] Present address: Department of Biology, Southern Nazarene University, 6729 NW 39th Expressway, Bethany, OK 73008, USA.

1. Introduction

The development of species-distribution and richness models has received considerable attention over the past decade. Species-distribution models often are used to predict occurrences (both presences and absences; Angermeier et al. 2002) of individual species by developing an equation or rule-set typically based on environmental variables, such as land-cover types or climatic measures. With the increase in types of models now available (Guisan and Thuiller, 2005), one is faced with deciding which models are most appropriately applied to particular data sets and/or for particular purposes. However, few comparisons have been published that can assist researchers in deciding which models are most appropriate for their data.

Manel et al. (1999) showed that logistic regression was more suitable than either artificial neural networks or discriminant analysis in predicting presence/absence of river birds. Other comparisons have assessed models as used with presence-only data. Brotons et al. (2004) and Olivier and Wootherspoon (2006) compared the results of ecological niche factor analysis (EFNA) of Biomapper (Hirzel et al., 2002) and of logistic regression with pseudo-absences. Logistic regression provided more accurate predictions than those from ENFA.

The genetic algorithm rule-set prediction (GARP; Stockwell and Noble, 1992) is another model type typically used in biological applications employed on presence-only data. It has been used extensively and compared to logistic regression, using pseudo-

absences, to determine the effect of sample size on model efficiency (Stockwell and Peterson, 2002). GARP was declared more efficient, using fewer sites to predict presence-only data developed from museum collections; only 50 sites were required by GARP to reach a consistent prediction level, whereas 100 sites were needed to reach the same level as logistic regression using pseudo-absences. However, it is of interest to determine whether GARP would be the technique of choice if one has reputable occurrence information, such as is provided through a number of projects, including statewide programs to develop breeding-bird atlases. Occurrence data from atlas programs also can be employed to evaluate the degree to which GARP and logistic regression accurately predict actual distributions.

In addition, it is of interest to evaluate the degree of concordance between estimates of species richness based on individual species models and actual species richness. Studies have used species-distribution data to sum the number of species that occurred at a site and then, using environmental data, predicted species richness for unsampled sites (e.g. Bohning-Gaese, 1997; van Rensberg et al., 2002; Waldhardt et al., 2004). Others have developed predictions of species richness by summing the predicted presences of individual species at a site and comparing the prediction to the actual site richness (Lehmann et al., 2002; Zaniewski et al., 2002). To be useful, the sum of the unique predictions should provide an estimate that is near the actual richness, or at least provide good relative measure for comparison of species richnesses across sites.

Few comparisons of different species-distribution models have been published that assess model accuracy (Brotons et al., 2004; Olivier and Wootherspoon, 2006). Using data from the *Oklahoma Breeding Bird Atlas* (Reinking, 2004), we have evaluated the

degree to which logistic regression and two forms of GARP correctly predict occurrences, presences, and absences of individual species based on environmental variables. We have made direct comparisons for all species, as well as for: (1) subgroups of species as determined by their distribution extents; and (2) subgroups of localities with low, moderate, and high species richness. Estimates of species richness derived from summations of results for individual species models also were examined.

2. Methods

2.1. Bird data and environmental descriptors

Distribution data for a species can be classified as either occurrence information (presences or absences) or presence-only data. Breeding-bird atlases (e.g., Peterjohn and Rice, 1991; Corman and Wise-Gervais, 2005), where observers visit numerous localities within a given geographic area and record the breeding bird species, can provide reliable occurrence data. For atlas projects observers must meet defined minimum time criteria for a block to be included (e.g. Reinking, 2000); thus, one has presence/absence information for each species at a series of localities. While, of course, a species at a particular site could have been present but went undetected (i.e. an apparent commission error; Anderson et al., 2003); it seems unlikely that commission errors of this type would be widespread. In our comparisons of logistic regression and two forms of GARP, we used occurrence data for 209 bird species at 562 blocks (Fig. 1) as reported in the *Oklahoma Breeding Bird Atlas*. Blocks, initially selected using a stratified-random procedure to ensure no undo clumping spatially, were sampled from 1997 through 2001.

2.2. Environmental variables

We used two types of environmental descriptors in this study – 13 land-cover types and 21 climatic variables. Twelve of the land-cover types were based on an initial 19 land-cover types as described from the 1992 land-cover image in the United States Geological Survey archive (USGS, 2002). We consolidated four variables (low intensity residential, high intensity residential, industrial, and urban grasses) into one variable "developed", three variables (row crops, small grains, and fallow) into "crops", and a further three variables (bare rock, quarries/mines, and transitional) into "barren". With these modifications the 12 resulting land-cover variables were: (1) deciduous forest; (2) mixed forest; (3) evergreen forest; (4) woody wetland; (5) emergent herbaceous wetland; (6) shrubland; (7) grassland; (8) pasture/hay; (9) cropland; (10) developed; (11) barren; and (12) water. A 13th variable, distance to water (in meters), was calculated as the distance from the atlas block center to the nearest river or stream as identified in the Oklahoma Digital Elevation Model hydrological network derived from the 1:100,000-scale digital topographic map (USGS, 1998). We obtained measurements for this variable using the Distance Between Points (between layers) procedure of Analysis Tools (Beyer, 2004) in ArcGIS (ESRI, 2004).

We generated climatic variables using point data provided through the Oklahoma Mesonet (Brock et al., 1995), a series of 119 weather stations with at least one in each of the state's 77 counties. The data we employed were averaged for 1997 through 2001 covering May, June, and July, corresponding to the years and months when the bird surveys were conducted. For each of the three months (May, June, and July), seven variables were selected: temperature (^{o}C); soil temperature (^{o}C); rainfall (cm); solar

radiation (megajoules/m^2); wind speed (kph); barometric pressure (millibars); and humidity (%). The 21 resulting climatic variables were interpolated to generate layers for all locations in the state using the ordinary kriging method (van Horseen et al., 1999) in the Geostatistical Analyst of ArcGIS (ESRI, 2004). The climatic layers were intersected by the atlas-block layer to provide the climatic variable using the Intersect Point Tool of Analysis Tools (Beyer, 2004) in ArcGIS. Using the raster-to-ASCII conversion tool of ArcToolbox, we exported all variables from ArcGIS to ASCII format files for use in GARP.

2.3. Model development

We developed models for each of the 209 species using both logistic regression and GARP. For logistic regression, we employed a stepwise procedure ($p = 0.05$) in SAS 9 (SAS, 2004) to select from the 34 environmental variables the subset that best explained presence/absence for each species. Each logistic-regression equation produced a score for its species that ranged from 0 to 1. For any location with a score for a species of 0.5 or greater, the species was considered present (Pearce and Ferrier, 2000). Predicted presences and absences were compared to actual presences and absences for each species to determine the model's accuracy. A result for an individual block was then categorized as a correctly predicted presence, an incorrectly predicted presence, an incorrectly predicted absence, or a correctly predicted absence (van Horseen et al., 1999). These categories correspond to true positive, false positive, false negative, and true negative of Anderson et al. (2003). Thus, the performance of the technique was judged based on intrinsic measures of error (Anderson et al., 2003).

Two variations of GARP were run: $GARP_{Best\ subset}$ and $GARP_{50:50}$. For $GARP_{Best\ subset}$, we generated 21 maps using 1000 iterations or until convergence was reached for each species (Stockwell and Peters, 1999). Previous studies (e.g. Anderson et al., 2002; Raxworthy et al., 2003; Illoldi-Rengel et al., 2004; Peterson et al., 2006) typically have employed $GARP_{Best\ subset}$, using this technique to generate from 10-100 maps, with model criteria being a low-omission threshold and a moderate commission-error threshold (Anderson et al., 2003). In our study, $GARP_{Best\ subset}$ was the best subset based on optimal combinations of error components, as per Anderson et al. (2003) and Raxworthy et al. (2003). For each species we produced 21 replicate models and a best subset of 10 models; models that predicted less than 90% (i.e. 10% hard omission threshold) of presences were discarded and from among the remaining models the 10 closest to the median predicted area were summed to provide a "best distributional prediction" (Raxworthy et al., 2003). The commission threshold was set at 50% of the distribution. The resulting 10 best maps were exported as Arc/INFO Grids, brought into ArcGIS, and summed using a raster calculator. The values of the sums could then vary from 0 to 10 representing the number of runs in which the species was predicted present for a cell; not all maps had all possible values. We intersected the summed maps with the center points of atlas blocks in ArcGIS. The summed map was considered to represent the potential distribution of a species. Due to the variability in number of values in each map any value greater than zero was judged to be a presence (1), and zero an absence (0). As with logistic regression, predicted presences and absences were compared to actual presences and absences for a species.

For $GARP_{50:50}$, the 21 maps were summed as a potential distribution map, generating 22 possible occurrence categories (0-21) so that a 0.5 cutoff could be established somewhat analogous to that used with the logistic-regression procedure. The 21 maps were also exported as Arc/INFO Grids, brought into ArcGIS, similarly summed using a raster calculator, and intersected. A species having a value of 11-21 for a block was judged to be present (1) in that block; when the sum was from 0-10 for a block, the prediction was that the species was absent (0). Predicted presences and absences then were compared to the actual presences and absences.

2.4. Model comparisons

Sites for which species were correctly predicted as either present or absent were summed to determine the total correctly predicted occurrences. Correctly predicted presences and absences also were considered separately. Correctly predicted occurrences, presences, and absences were the basis of pairwise comparisons of the results for logistic regression, $GARP_{50:50}$, and $GARP_{\text{Best subset}}$. We tabulated which technique was the best predictor for each of the 209 species, with the result compared to the expected null hypothesis using a *G*-test with one degree of freedom (Sokal and Rohlf, 1997).

Similar comparisons were made for subgroups of species determined on the basis of the extent of their actual distributions. Species were partitioned into five approximately equal-range groups: those having sparse distributions (being present in 1-112 of the 562 blocks), moderately sparse distributions (113-225 blocks), intermediate distributions (226-338 blocks), moderately widespread distributions (339-451 blocks), and widespread distributions (452-562 blocks).

In comparisons within each of the 562 blocks, based on the 209 individual species models, we tabulated over species the correctly predicted presences, incorrectly predicted presences, correctly predicted absences, and incorrectly predicted absences; pairwise comparisons were used to determine which technique was the best predictor. Similar comparisons were done for subgroups of blocks – those with relatively low, moderate, and high species richnesses (24-49, 50-75, and 76-100 species, respectively; Fig. 1).

For each technique we then counted the number of species predicted to be present in each block, thus obtaining the predicted species richness. The actual and predicted species richnesses were then compared.

3. Results

3.1. Overall prediction accuracy

Accuracy for predicting species occurrences (both presences and absences) was relatively good for all models. The average correctly predicted occurrences by logistic regression was 89.3%, by $GARP_{50:50}$ 76.7%, and by $GARP_{Best\ subset}$ 70.3%. Logistic regression performed significantly better than either $GARP_{50:50}$ or $GARP_{Best\ subset}$ for predicting occurrences (Table 1). When comparing only presences, $GARP_{50:50}$ did better than logistic regression, while $GARP_{Best\ subset}$ outperformed both logistic regression and $GARP_{50:50}$ (Table 1). However, when only absences were considered, logistic regression performed better than either form of GARP. Comparing GARP procedures for overall performance showed that $GARP_{50:50}$ did better than $GARP_{Best\ subset}$ (Table 1) in predicting occurrences. When GARP procedures were compared for correctly predicting absences, $GARP_{50:50}$ was a better predictor.

3.2. Predictive ability for species relative to distribution extents

When species were separated into five groups based on the extents of their distributions (separation shown by dotted lines in Figs. 2 and 3), logistic regression was best when predicting at the extremes, with the mean percent of correctly predicted occurrences of sparse and widespread distributions being 96.0 and 92.3%, respectively. For distributions intermediate in extent, prediction performance generally was poorer ($\bar{x}$ = 69.0%) and variability among species in model performance was higher (central portion of Fig. 2a).

$GARP_{50:50}$ models showed a range of performance. For sparse distributions (left part of Fig. 2b), predictive ability ranged from 100% to less than 20% correctly predicted occurrences, with the predictions for most sparsely distributed species being highly accurate. Predictive performance by $GARP_{50:50}$ was poorest for several species present at about 100 sites. In general, there was better predictive accuracy for more widespread species, with the average percents of correctly predicted occurrences being 61.3%, 77.0%, and 82.3% for intermediate, moderately widespread, and widespread distributions, respectively (middle section and those to right in Fig. 2b).

$GARP_{Best\ subset}$ showed a pattern similar to $GARP_{50:50}$, but more pronounced (Fig. 2c). Predictions for the sparsest distributions were good, but for several species found at 25 to 60 sites only 20% of occurrences were correctly predicted, although occurrence predictions for a few species in this range were over 80% accurate. For species found at about 100 sites or more, predictive ability of $GARP_{Best\ subset}$ increased in a linear fashion, eventually approaching 100% accuracy.

When presences and absences are considered separately, the sources of variability in the ability to predict occurrences are evident. Logistic regression showed variability in predictive ability for sparse and moderately sparse distributions (Fig. 3a, two leftmost sections), but correct predictions of performance increased markedly from species with intermediate distributions to those that were widespread. The pattern for predictions of absences was essentially a mirror image, ranging from perfect for species found in only a few localities to zero for species found in almost all sites.

For $GARP_{50:50}$, the percent of correctly predicted occurrences showed a range of variation among species found in only a few localities (Fig. 3c), was very high for most species found at from 50 to 300 sites, and was somewhat lower for many of the species found at more localities. $GARP_{50:50}$ predicted absences well for species found at only a few localities (Fig. 3d); for the rest of the species, predictive ability ranged widely (10 to 89%).

$GARP_{Best\ subset}$ showed a similar but more pronounced pattern to $GARP_{50:50}$ for percents of correctly predicted occurrences (Fig. 3e) and correctly predicted presences (Fig. 3f). Except for a few species found at only a few localities, $GARP_{Best\ subset}$ correctly predicted presences at or near 100% of the time (Fig. 3e). This high level of prediction ability for presences was achieved by substantially overpredicting the number of sites where any given species was present, with the result that, except for a few species found infrequently in Oklahoma, predictive ability for where species were not found was notably diminished (Fig. 3f).

When the occurrences of species were compared within their five distribution groups, logistic regression predicted occurrences better than $GARP_{50:50}$. $GARP_{50:50}$ was the better

predictor for 44 of the 120 species with sparse distributions (Table 2). When presences and absences were separated, $GARP_{50:50}$ better predicted presences for the first three distribution groups (Table 2). A comparison of the presence predictions of the moderately widespread species showed no significant difference in predictive ability of logistic regression and $GARP_{50:50}$. When widespread distributions were compared, logistic regression did better (Table 2). For absences, logistic regression consistently did better for the first four distribution groups. $GARP_{50:50}$ better predicted absences for 19 of the 24 species with widespread distributions (Table 2), although neither $GARP_{50:50}$ nor logistic regression did particularly well (left side of Figs. 3b and d).

When logistic regression was compared to $GARP_{Best\ subset}$ for sparse distributions, logistic regression was the better predictor overall; however, 27 species were better predicted by $GARP_{Best\ subset}$. The species in the other four distribution groups were better predicted by logistic regression (Table 2). When only presences were compared, $GARP_{Best\ subset}$ better predicted the first four distribution groups, and there was no significant difference for the widespread distributions (Table 2). For widespread distributions, 12 species were better predicted by logistic regression, with 10 better predicted by $GARP_{Best\ subset}$. Logistic regression better predicted absences in all five distribution groups (Table 2). Since $GARP_{Best\ subset}$ does not use absences directly, this result was expected.

When the two GARP procedures were compared, results were more similar to each other than to those for logistic regression. For the first three groups of species, distributions were better predicted by $GARP_{50:50}$ (Table 2). The comparison of GARP procedures using the moderately widespread group of species distributions was not

significantly different; $GARP_{50:50}$ better predicted 5 species distributions to the 11 better predicted by $GARP_{Best\ subset}$ (Table 2). All species with widespread distributions were better predicted by $GARP_{Best\ subset}$ (Table 2). When presences were compared, $GARP_{Best\ subset}$ better predicted all but one species (yellow-crowned night-heron, *Nycticorax violacea*) in all distribution groups (Table 2). For absences, when compared to $GARP_{Best\ subset,}$ $GARP_{50:50}$ tied or better predicted all species in all distribution groups.

3.3. Predictive ability for blocks relative to species richnesses

Surveyed blocks exhibited a wide range of species richnesses (24-100 species), with an average of 56.8 species. Blocks with the highest species richnesses (Fig. 1, black squares) tended to be associated with large, man-made reservoirs, such as Lake Fort Supply in the northwest, Lake Eufala and Lake Tenkiller in the east, and Lake Texoma in the south. These blocks support shorebirds and waterbirds in addition to land birds.

Logistic regression correctly predicted species occurrences from 78.0 to 95.7% of the time for the 562 blocks. Predictions tended to be less accurate for localities with the most species (Fig. 4a). For $GARP_{50:50,}$ the range was 59.8 to 91.9% (Fig. 4b). When plotted against actual species richness a J-shaped curve resulted, with predictions being relatively good for sites where species richness was very low or very high. Predictions generally were poorest for localities with species richnesses of 30 to 50 species; except for those localities with low species richnesses, prediction percentages tended to increase in a linear fashion with actual species richnesses. A similar pattern to that for $GARP_{50:50}$ was found for $GARP_{Best\ subset}$ (Fig. 4c), but the average percent correct was lower (70.3%, range 54.1 to 90.4%).

We also examined the degree to which each of the techniques correctly predicted species presences only and absences only for each block (Fig. 5). For logistic regression, there was an inverse relationship of percent species correctly predicted present and species richness (Fig. 5a). For percent species correctly predicted absent, logistic regression provided better predictions when more species were present (Fig. 5b). $GARP_{50:50}$, in general, did well at correctly predicting species that were present, although there were a number of blocks where it performed poorly (Fig. 5c). When species actually were absent, $GARP_{50:50}$ made numerous mistakes for most localities (Fig. 5d), although it did better for blocks with very low species richnesses. For $GARP_{Best\ subset}$, the pattern paralleled that of $GARP_{50:50}$, but was more extreme. For most blocks, presences were correctly predicted by $GARP_{Best\ subset}$ most of the time (Fig. 5e), but absences were poorly predicted for almost all localities (Fig. 5f).

When compared to one another, logistic regression better predicted occurrences at localities than either $GARP_{50:50}$ or $GARP_{Best\ subset}$, while $GARP_{50:50}$ performed better than $GARP_{Best\ subset}$ (Table 3). $GARP_{Best\ subset}$ was a better predictor of presences than either logistic regression or $GARP_{50:50}$. However, logistic regression was notably better than either GARP technique at predicting absences at each locality, with $GARP_{50:50}$ always doing better than $GARP_{Best\ subset}$.

Irrespective of whether species richness was low, moderate, or high, logistic regression outperformed both forms of GARP in correctly estimating occurrences of species at localities, with only one comparison (i.e. logistic regression vs. $GARP_{50:50}$ for high species richness) not being statistically significant (Table 4). $GARP_{50:50}$ was a better

predictor of occurrences than $GARP_{Best\ subset}$ for sites with low, moderate, and high species richness (Table 4).

For presences of species at given localities, $GARP_{50:50}$ was a better predictor than logistic regression and $GARP_{Best\ subset}$ better than either of the other techniques for all three subgroups – sites with low, moderate, and high species richness (Table 4). However, when judging whether species were absent at localities, logistic regression was the best performer at all three species-richness levels, and $GARP_{50:50}$ was better than $GARP_{Best\ subset}$ (Table 4).

3.4. Predicted species richness

Predictions of species richness using individual species models based on logistic regression were relatively close to actual species richnesses, but tended to be too high for localities with low species richness and too low for localities with high species richness (Fig. 6a). The product-moment correlation of actual and predicted species richness based on logistic regression was 0.44 ($p < 0.001$; Fig. 6a). The average absolute deviation between actual and predicted species richness was 9.67 species, while the average, taking sign into account, was -4.9 species.

$GARP_{50:50}$ typically overpredicted species richness irrespective of actual species richness, but there were a few sites where species richness was exactly predicted and some where the parameter was underestimated (Fig. 6b). The lowest correlation of actual and predicted species richness occurred using this technique ($r = 0.35$, $p < 0.001$; Fig. 6b). The average absolute deviation was 40.9 species and, taking into account sign, was 37.4 species.

Species richness was overpredicted by $GARP_{Best\ subset}$ for all but one of the blocks (Fig. 6c). The correlation of actual and predicted species richness was 0.39 ($p < 0.001$; Fig. 6c). The average absolute deviation (as well as the deviation considering sign) was 60.7 species; thus, predictions of species richness on average using this technique were more than twice the average actual species richness of 56.8.

4. Discussion

We used intrinsic criteria to evaluate the efficacy of logistic regression and two forms of GARP in correctly estimating species occurrences based on presence-absence data for Oklahoma birds. We have not addressed directly the question of which technique or techniques are best applied to presence-only data, although our analyses indicate that GARP considerably overestimates actual distributions. In our study, logistic regression was better overall in predicting actual occurrences than either of the GARP techniques used. While not documented in detail previously, this is not a particularly surprising result given that GARP as typically implemented in biological studies only makes use of presences, while logistic regression considers both presences and absences.

Logistic regression showed less variability in accuracy at predicting occurrences of species, particularly for species with sparse or widespread distributions. For both forms of GARP, the variability in accuracy was widest for species with sparse distributions, particularly those that occurred in fewer than 50 blocks. This difference in accuracy is consistent with the sample effect reported by Stockwell and Peterson (2002). Using the *Atlas of Mexican Bird Distributions*, they showed GARP to be less accurate when fewer than 50 sites were used to develop distribution models. The inconsistency of the percent

of correctly predicted species occurrences in our analyses reflects the overall reduced accuracy shown by the sample effect.

Breeding-bird atlases employ minimum sampling criteria such that each block is sampled to produce reputable occurrence data of both presences and absences. GARP, using only presence data, is a less than optimal choice for accurately predicting species distributions based on occurrence information. If the purpose is to estimate "potential" distributions, it is an open question as to whether it is advantageous to employ a technique in a way that overestimates where species occur and probably overestimate the range extents of these species as well. Even if one is interested in the "potential" range, such a concept is difficult to define objectively, and is readily open to multiple subjective and often nontestable interpretations.

Occurrences of species at sites with relatively low or moderate actual species richnesses were better predicted by logistic regression than were species at sites of high species richness (Fig. 4a). This difference is due to relatively poor predictions of presence for sparsely distributed species; these typically are "add-on species" to the list of more widespread species, with the result that the given locality is relatively species-rich. In fact, logistic regression predicted that several of these add-on species did not occur in any of the atlas blocks in the state. Similar findings were reported for the Florida breeding-bird-atlas project in that species richness at a site was predominantly comprised of common species that occurred statewide (Cox, 2006).

$GARP_{50:50}$, on average, correctly predicted occurrences of the species found in a block 76.6% of the time. The left portion of the J-shape of the curve (Fig. 4b) was generated because of blocks along the Oklahoma-Kansas border; these sites had relatively low

species richness, yet were well predicted. For the remaining blocks there was a positive linear association of percent correctly predicted occurrences with actual numbers of species in those blocks. This is due to $GARP_{50:50}$ uniformly overpredicting species presence at sites. The fact that, with the exception of the relatively species-poor sites mentioned, occurrences were better predicted for relatively species-rich blocks is simply a function of GARP notably overpredicting presences; when more species actually are present GARP does better on a percentage basis.

Of the three techniques evaluated, $GARP_{Best\ subset}$ had the lowest average percent correctly predicted occurrences at 70.3%. It did reach 90.4% for one block, a relatively species-poor site along the Oklahoma-Kansas border. Aside from these border blocks, $GARP_{Best\ subset}$ also showed near-linear and positive improvement as more species occurred at a site (Fig. 4c); however, it did not reach the performance level of $GARP_{50:50}$ or logistic regression on a site-by-site basis. As with $GARP_{50:50}$, the higher level of performance by $GARP_{Best\ subset}$ for relatively species-rich blocks was due to it markedly overpredicting occurrences on a routine basis, which resulted in a higher percentage of correctly predicted occurrences for species-rich blocks.

In a similar study, Lehmann et al. (2002) used 10 environmental variables and fern occurrence data in GRASP (generalized regression analysis and spatial predictions) to sum predictions for species richness of 43 species of ferns. These predictions were then compared to the actual species richness at sites in New Zealand using a product-moment correlation of predicted to actual species richnesses, with a result that was notably higher ($r = 0.72$) than found in our study. They did have difficulty using GRASP to predict occurrences of species that occurred in only a few sites.

Zaniewski et al. (2002) developed predictions for the same fern data based on environmental predictors using GAM (general additive models) and ENFA (ecological niche factor analysis) to compare the abilities of these techniques to predict species richness by summing individual predictions. On average ENFA, using only species presence data similarly to GARP, predicted higher species richnesses per site than GAM, which uses species presence and absence data similarly to logistic regression. ENFA and GARP both overpredicted the species richnesses for sites, suggesting that using only presence data to develop models may leave out important information by not considering absence data in their analyses.

The development of individual species models to predict occurrences will continue to be a useful tool in identifying areas of interest for the given species. Many of these models were developed using occurrence data or presence-only data, in combination with appropriate predictor variables. However, our investigation suggests that there may be only limited value to summing individual species predictions to estimate species richness at a site. While logistic regression provided good estimates of actual species richnesses, it did not produce consistently reliable relative measures of species richness across sites. GARP did not provide accurate estimates of species richnesses or reliable relative measures of species richness across sites.

Acknowledgements

Access to the Oklahoma breeding bird atlas data set was kindly provided by Dan L. Reinking and the Sutton Avian Research Center. The first author was supported by George Miksch Sutton Scholarships in Ornithology and a Robert E. and Mary B. Sturgis

Scholarship, as well as by a grant from the University of Oklahoma Graduate Student Senate. We thank Jeffrey F. Kelly, William J. Matthews, Thomas S. Ray, and May Yuan for serving as committee members of the first author and for helpful comments on earlier drafts of the manuscript. This is a portion of a dissertation submitted by the first author in partial fulfillment of requirements for Ph.D. in the Department of Zoology, University of Oklahoma.

References

Anderson, R.P., Gomez-Laverde, M., Peterson, A.T., 2002. Geographical distributions of spiny pocket mice in South America: insights from predictive models. Global Ecology and Biogeography 11, 131–141.

Anderson, R.P., Lew, D., Peterson, A.T., 2003. Evaluating predictive models of species' distributions: criteria for selecting optimal models. Ecological Modelling 162, 211–232.

Angermeier, P.L., Krueger, K.L., Dolloff, C.A., 2002. Discontinuity in stream-fish distributions: implications for assessing and predicting species occurrence. In: Scott, J.M., Heglund, P.J., Morrison, M.L., Haufler, J.B., Raphael, M.G., Wall, W.A., Samson F.B. (Eds.), Predicting species occurrences: issues of accuracy and scale. Island Press, Washington DC, pp. 519–527.

Beyer, H. L., 2004. Hawth's analysis tools for ArcGIS. http://www.spatialecology.com/htools. September 23, 2005.

Bohning-Gaese, K., 1997. Determinants of avian species richness at different spatial scales. Journal of Biogeography 24, 49–60.

Brock, F.V., Crawford, K.C., Elliot, R.L., Cuperus, G.W., Stadler, S.J., Johnson, H.L., Eilts, M.D., 1995. The Oklahoma Mesonet: a technical overview. Journal of Atmospheric and Oceanic Technology 12, 5–19.

Brotons L., Thuiller W., Araujo, M.B., Hirzel, A.H. 2004. Presence-absence versus presence-only modeling methods for predicting bird habitat suitability. Ecography 27, 437–448.

Corman, T.E., Wise-Gervais, C. (Eds.), 2005. Arizona breeding bird atlas. University Press of New Mexico, Albuquerque, New Mexico, USA, 636 pp.

Cox, J., 2006. Trends in breeding distributions based on Florida's breeding bird atlas project. In: Noss, R.F. (Ed.) The breeding birds of Florida. Special publication No. 7, Florida Ornithological Society, Gainsville, Florida, pp. 71–126.

ESRI, 2004. ArcGIS version 9.0. Environmental Systems Research Institute, Redlands, California, USA.

Guisan, A., Thuiller, W., 2005. Predicting species distributions: offering more than simple habitat models. Ecology Letters 8, 993–1009.

Hirzel, A.H., Hausser, J., Chessel, D., Perrin, N., 2002. Ecological Niche Factor Analysis: how to compute habitat suitability maps without absence data? Ecology 83: 2027–2036.

Illoldi-Rangel, P., Sanchez-Cordero, V., Peterson, A.T, 2004. Predicting distributions of Mexican mammals using ecological niche modeling. Journal of Mammalogy 85, 658–662.

Lehmann, A., Leathwick, J.R., Overton, J.McC., 2002. Assessing New Zealand fern diversity from spatial predictions of species assemblages. Biodiversity and Conservation 11, 2217–2238.

Manel, S., Dias, J.M., Buckton, S.T., Ormerod, S.J., 1999. Alternative methods for predicting species distribution: an illustration with Himalayan river birds. Journal of Applied Ecology 36, 734–747.

Olivier, F., Wotherspoon, S.J., 2006. Modelling habitat selection using presence-only data: case study of a colonial hollow nesting bird, the snow petrel. Ecological Modelling 195, 187–204.

Pearce, J., Ferrier, S., 2000. Evaluating the predictive performance of habitat models developed using logistic regression. Ecological Modelling 133, 225–245.

Peterjohn, B.G., Rice, D.L., 1991. The Ohio breeding bird atlas. Ohio Department of Natural Resources, Columbus, Ohio, USA, 414 pp.

Peterson, A.T., Kluza, D.A., 2003. New distributional modeling approaches for GAP Analysis. Animal Conservation 6, 47–54.

Peterson, A.T., Papers, M., Reynolds, M.G., Perry, N.D., Hanson, B., Regnery, R.L., Hutson, C.L., Muizniek, B., Damon, I.K., Carrol, D.S., 2006. Native-range ecology and invasive potential of *Cricetomys* in North America. Journal of Mammalogy 87, 427–432.

Raxworthy, C.J., Martinez-Meyer, E., Horning, N., Nussbaum, R.A., Schneider, G.E., Ortega-Huerta, M.A., Peterson, A.T., 2003. Predicting distributions of known and unknown reptile species in Madagascar. Nature 426, 837-841.

Reinking D.L., 2000. Oklahoma breeding bird atlas handbook. George M. Sutton Avian Research Center, Bartlesville, Oklahoma, USA.

Reinking, D.L. (Ed.), 2004. Oklahoma breeding bird atlas. University of Oklahoma Press, Norman, Oklahoma, USA. 519 pp.

SAS, 2004. Statistical Analysis Software, release 9. SAS Institute Inc, Cary North Carolina, USA.

Sokal, R.R., Rohlf, F.J., 1997. Biometry 3rd ed. W.H. Freeman. New York, New York, USA, 887 pp.

Stockwell, D.R.B., Noble, I.R., 1992. Induction of sets of rules from animal distribution data: a robust and informative method of data analysis. Mathmatics and Computers in Simulation 33, 385–390.

Stockwell, D.R.B., Peters, D., 1999. The GARP modeling system: problems and solutions to automated spatial prediction. International Journal of Geographic Information Science 13, 143–158.

Stockwell, D.R.B., Peterson, A.T., 2002. Effects of sample size on accuracy of species distribution models. Ecological Modelling 148, 1–13.

Stockwell D.R.B., Peterson, A.T., 2003. Comparison of resolution of methods used in mapping biodiversity patterns from point-occurrence data. Ecological Indicators 3, 213–221.

USGS, 1998. US Geological Survey DEM 7.5 Quadrangle. US Geological Survey, Reston, Virginia, USA.

USGS, 2002. Oklahoma land-cover data set. http://edcftp.cr.usgs.gov/pub /data/landcover/states/. September 23, 2005.

van Horseen, P.W., Schot, P.P., Barendregt, A., 1999. A GIS-based plant prediction model for wetland ecosystems. Landscape Ecology 14, 253–265.

van Rensburg, B.J., Chown, S.L., Gaston, K.J., 2002. Species richness, environmental correlates, and spatial scale: a test using South African birds. American Naturalist 159, 566–577.

Waldhardt, R., Simmering, D., Otte, A., 2004. Estimation and prediction of plant species richness in a mosaic landscape. Landscape Ecology 19, 211–226.

Zaniewski, A.E., Lehmann, A., Overton, J.McC., 2002. Predicting species spatial distributions using presence-only data: a case study of native New Zealand ferns. Ecological Modelling 157, 261–280.

Table 1 – Paired comparisons of techniques, indicating percent of the 209 species for which particular technique better predicted total occurrences, presences, and absences. Number in parentheses indicates number of species for which technique was better predictor. Statistical significance (*G*-test; ***, $p < 0.001$) of deviation from random expectation for proportion (i.e., 0.5:0.5) of species best predicted by the two techniques compared; ties not considered in statistical comparison

Technique that was best predictor	Correctly predicted		
	Occurences	Presences	Absences
Logistic regression vs. $GARP_{50:50}$			
Logistic regression	73.7 (154)***	16.3 (34)	70.8 (148)***
Tied	4.3 (9)	1.9 (4)	15.8 (33)
$GARP_{50:50}$	22.0 (46)	81.8 (171)***	13.4 (28)
Logistic regression vs. $GARP_{Best\ subset}$			
Logistic regression	83.7 (175)***	6.2 (13)	84.7 (177)***
Tied	2.4 (5)	2.9 (2)	12.4 (26)
$GARP_{Best\ subset}$	13.9 (29)	90.9 (194)***	2.9 (6)
$GARP_{50:50}$ vs. $GARP_{Best\ subset}$			
$GARP_{50:50}$	69.9 (146)***	0.5 (1)	87.6 (183)***
Tied	10.0 (21)	17.2 (35)	12.4 (26)
$GARP_{Best\ subset}$	20.1 (42)	82.3 (173)***	0.0 (0)

Table 2 – For subgroups (based on number of blocks where a species was present) of the 209 species, paired comparisons indicating percent of time particular technique better predicted total occurrences, presences, and absences. Number of species indicated in parentheses. Subgroups defined as: sparse distributions (species present in 1-112 blocks); moderately sparse (113-225); intermediate (226-338); moderately widespread (339-451); and widespread (452-564). Statistical significance (*G*-test; ^{ns}, $p > 0.05$; *, $p < 0.05$; **, $p < 0.01$; ***, $p < 0.001$) of deviation from random expectation that an equal proportion (0.5:0.5) of species would be best predicted by the two techniques. For statistical comparisons, ties not considered

Distribution extent and technique that was best predictor	Correctly predicted		
	Occurrences	Presences	Absences
Logistic regression vs. $GARP_{50:50}$			
Sparse distributions			
Logistic regression	55.0 (66)*	0.0 (0)	69.2 (83)***
Tied	8.3 (10)	2.5 (3)	26.7 (32)
$GARP_{50:50}$	36.7 (44)	97.5 (117)***	4.2 (5)
Moderately sparse distributions			
Logistic regression	100.0 (31)***	3.2 (1)	100.0 (31)***
Tied	0.0 (0)	0.0 (0)	0.0 (0)
$GARP_{50:50}$	0.0 (0)	96.8 (30)***	0.0 (0)

Table 2 (continued).

Distribution extent and technique that was best predictor	Correctly predicted		
	Occurrences	Presences	Absences
Intermediate distributions			
Logistic regression	100.0 (18)***	0.0 (0)	100.0 (18)***
Tied	0.0 (0)	0.0 (0)	0.0 (0)
$GARP_{50:50}$	0.0 (0)	100.0 (18)***	0.0 (0)
Moderately widespread distributions			
Logistic regression	93.8 (15)***	56.3 (9)^{ns}	81.3 (13)**
Tied	0.0 (0)	6.3 (1)	0.0 (0)
$GARP_{50:50}$	6.3 (1)	37.5 (6)	18.8 (3)
Widespread distributions			
Logistic regression	100.0 (24)***	100.0 (24)***	16.7 (4)
Tied	0.0 (0)	0.0 (0)	4.2 (1)
$GARP_{50:50}$	0.0 (0)	0.0 (0)	79.2 (19)**
Logistic regression vs. $GARP_{Best\ subset}$			
Sparse distributions			
Logistic regression	75.0 (90)***	0.0 (0)	80.0 (96)***
Tied	2.5 (3)	0.0 (0)	19.2 (23)
$GARP_{Best\ subset}$	22.5 (27)	100.0 (120)***	0.8 (1)

Table 2 (continued).

Distribution extent and technique that was best predictor	Correctly predicted		
	Occurrences	Presences	Absences
Moderately sparse distributions			
Logistic regression	100.0 (31)***	0.0 (0)	100.0 (31)***
Tied	0.0 (0)	0.0 (0)	0.0 (0)
$GARP_{Best\ subset}$	0.0 (0)	100.0 (31)***	0.0 (0)
Intermediate distributions			
Logistic regression	100.0 (18)***	0.0 (0)	100.0 (18)***
Tied	0.0 (0)	0.0 (0)	0.0 (0)
$GARP_{Best\ subset}$	0.0 (0)	100.0 (18)***	0.0 (0)
Moderately widespread distributions			
Logistic regression	100.0 (16)***	6.3 (1)	100.0 (16)***
Tied	0.0 (0)	0.0 (0)	0.0 (0)
$GARP_{Best\ subset}$	0.0 (0)	93.8 (15)***	0.0 (0)
Widespread distributions			
Logistic regression	87.5 (21)***	50.0 (12)ns	70.8 (17)**
Tied	8.3 (2)	8.3 (2)	12.5 (3)
$GARP_{Best\ subset}$	4.2 (1)	41.7 (10)	16.7 (4)

Table 2 (continued).

Distribution extent and technique that was best predictor	Correctly predicted		
	Occurrences	Presences	Absences
	$GARP_{50:50}$ vs. $GARP_{Best\ subset}$		
Sparse distributions			
$GARP_{50:50}$	77.5 (93)***	0.0 (1)	79.2 (95)***
Tied	17.5 (21)	20.8 (32)	20.8 (25)
$GARP_{Best\ subset}$	5.0 (6)	79.2 (87)***	0.0 (0)
Moderately sparse distributions			
$GARP_{50:50}$	96.8 (30)***	0.0 (0)	100.0 (31)***
Tied	0.0 (0)	9.7 (3)	0.0 (0)
$GARP_{Best\ subset}$	3.2 (1)	90.3 (28)***	0.0 (0)
Intermediate distributions			
$GARP_{50:50}$	100.0 (18)***	0.0 (0)	100.0 (18)***
Tied	0.0 (0)	0.0 (0)	0.0 (0)
$GARP_{Best\ subset}$	0.0 (0)	100.0 (18)***	0.0 (0)
Moderately widespread distributions			
$GARP_{50:50}$	31.3 (5)^{ns}	0.0 (0)	100.0 (16)***
Tied	0.0 (0)	0.0 (0)	0.0 (0)
$GARP_{Best\ subset}$	68.8 (11)	100.0 (16)***	0.0 (0)

Table 2 (continued).

Distribution extent and technique that was best predictor	Correctly predicted		
	Occurrences	Presences	Absences
Widespread distributions			
$GARP_{50:50}$	0.0 (0)	0.0 (0)	95.8 (23)***
Tied	0.0 (0)	0.0 (0)	4.2 (1)
$GARP_{Best\ subset}$	100.0 (24)***	100.0 (24)***	0.0 (0)

Table 3 – Paired comparison of techniques, indicating percent of sites (N=562) for which particular technique better predicted total occurrences, presences only, and absences only. Number in parentheses indicates number of sites for which technique was better predictor. Statistical significance (based on G-test) of deviation from random expectation (i.e. by chance alone two techniques would be equally likely to be best predictor; ***, $p < 0.001$). Ties not considered in statistical comparisons

Technique that was best predictor	Correctly predicted		
	Occurrences	Presences	Absences
Logistic regression vs. $GARP_{50:50}$			
Logistic regression	96.1 (540)***	17.4 (98)	96.3 (541)***
Tied	1.1 (6)	0.0 (0)	0.0 (0)
$GARP_{50:50}$	2.8 (16)	82.6 (464)***	3.7 (21)
Logistic regression vs. $GARP_{Best\ subset}$			
Logistic regression	98.4 (553)***	0.9 (5)	100.0 (562)***
Tied	0.2 (1)	1.1 (6)	0.0 (0)
$GARP_{Best\ subset}$	1.4 (8)	98.0 (551)***	0.0 (0)
$GARP_{50:50}$ vs. $GARP_{Best\ subset}$			
$GARP_{50:50}$	94.1 (529)***	0.5 (3)	100.0 (562)***
Tied	0.4 (2)	38.8 (218)	0.0 (0)
$GARP_{Best\ subset}$	5.5 (31)	60.7 (341)***	0.0 (0)

Table 4 – For subgroups of the 562 sites (based on number of species actually present at a site), paired comparisons indicating percent of time particular technique better predicted total occurrences, presences only, and absences only. Number in parentheses indicates number of sites for which technique was better predictor. Subgroups defined as including those sites with relatively low, moderate, and high species richnesses (24-49, 50-75, and 76-101 species per site, respectively). Statistical significance (based on *G*-test) of deviation from random expectation (i.e. by chance alone two techniques would be equally likely to be best predictor; ^{ns}, $p > 0.05$; ***, $p < 0.001$). Ties not considered in statistical comparisons

Subgroup and technique that was best predictor	Correctly predicted		
	Occurrences	Presences	Absences
	Logistic regression vs. $GARP_{50:50}$		
Low species richnesses			
Logistic regression	98.1 (155)***	34.2 (54)	91.1 (144)***
Tied	1.3 (2)	0.0 (0)	0.0 (0)
$GARP_{50:50}$	0.6 (1)	65.8 (104)***	8.9 (14)
Moderate species richnesses			
Logistic regression	97.8 (364)***	11.8 (44)	98.1 (365)***
Tied	1.3 (5)	0.0 (0)	0.0 (0)
$GARP_{50:50}$	0.8 (3)	88.2 (328)***	1.9 (7)
High species richnesses			
Logistic regression	59.4 (19)^{ns}	0.0 (0)	100.0 (32)***
Tied	3.1 (1)	0.0 (0)	0.0 (0)

Table 4 (continued).

Subgroup and technique that was best predictor	Correctly predicted		
	Occurrences	Presences	Absences
$GARP_{50:50}$	37.5 (12)	100.0 (32)***	0.0 (0)
Logistic regression vs. $GARP_{Best\ subset}$			
Low species richnesses			
Logistic regression	100.0 (158)***	2.5 (4)	100.0 (158)***
Tied	0.0 (0)	3.2 (5)	0.0 (0)
$GARP_{Best\ subset}$	0.0 (0)	94.3 (149)***	0.0 (0)
Moderate species richnesses			
Logistic regression	98.9 (368)***	0.5 (2)	100.0 (372)***
Tied	0.3 (1)	0.3 (1)	0.0 (0)
$GARP_{Best\ subset}$	0.8 (3)	99.2 (369)***	0.0 (0)
High species richnesses			
Logistic regression	84.4 (27)***	0.0 (0)	100.0 (32)***
Tied	0.0 (0)	0.0 (0)	0.0 (0)
$GARP_{Best\ subset}$	15.6 (5)	100.0 (32)***	0.0 (0)
$GARP_{50:50}$ vs. $GARP_{Best\ subset}$			
Low species richnesses			
$GARP_{50:50}$	97.5 (154)***	0.6 (1)	100.0 (158)***
Tied	0.0 (0)	36.1 (57)	0.0 (0)
$GARP_{Best\ subset}$	2.5 (4)	63.3 (100)***	0.0 (0)

Table 4 (continued).

Subgroup and technique that was best predictor	Correctly predicted		
	Occurrences	Presences	Absences
Moderate species richnesses			
$GARP_{50:50}$	92.7 (345)***	0.3 (1)	100.0 (372)***
Tied	0.5 (2)	40.3 (150)	0.0 (0)
$GARP_{Best\ subset}$	6.7 (25)	59.4 (221)***	0.0 (0)
High species richnesses			
$GARP_{50:50}$	93.8 (30)***	3.1 (1)	100.0 (32)***
Tied	0.0 (0)	34.4 (11)	0.0 (0)
$GARP_{Best\ subset}$	6.3 (2)	62.5 (20)***	0.0 (0)

Captions

Fig. 1 – Location of the 562 breeding-bird-atlas blocks in Oklahoma, indicating relative species richnesses of blocks. Squares are not to scale, being somewhat larger than actual survey blocks.

Fig. 2 – Percent of correctly predicted occurrences for each species for (a) logistic regression, (b) $GARP_{50:50}$, and (c) $GARP_{Best\ subset}$. From left to right, dotted lines separate sparse distributions (1-112 sites where species was present), moderately sparse distributions (113-225), intermediate distributions (226-338), moderately widespread distributions (339-451), and widespread distributions (452-562).

Fig. 3 – Percent of correctly predicted presences and absences, respecitively, for each species based on: (a, b) logistic regression; (c, d) $GARP_{50:50}$; and (e, f) $GARP_{Best\ subset}$. From left to right, dotted lines separate sparse distributions (1-112 sites), moderately sparse distributions (113-225), intermediate distributions (226-338), moderately widespread distributions (339-451), and widespread distributions (452-562).

Fig. 4 – Percent of correctly predicted occurrences of species for each block based on (a) logistic regression, (b) $GARP_{50:50}$, and (c) $GARP_{Best\ subset}$. Dotted lines partition localities into those with species richnesses that are relatively low, moderate, and high.

Fig. 5 – Percent of species at localities correctly predicted as being present or absent, respectively, using: (a, b) logistic regression; (c, d) $GARP_{50:50;}$ and (e, f) $GARP_{Best\ subset}$. Dotted lines partition localities into those with species richnesses that are relatively low, moderate, and high.

Fig. 6 – Scatter plot of actual species richness and predicted species richness for (a) logistic regression, (b) $GARP_{50:50}$, and (c) $GARP_{Best\ subset}$. Dotted line indicates where actual and predicted species richnesses are equal.

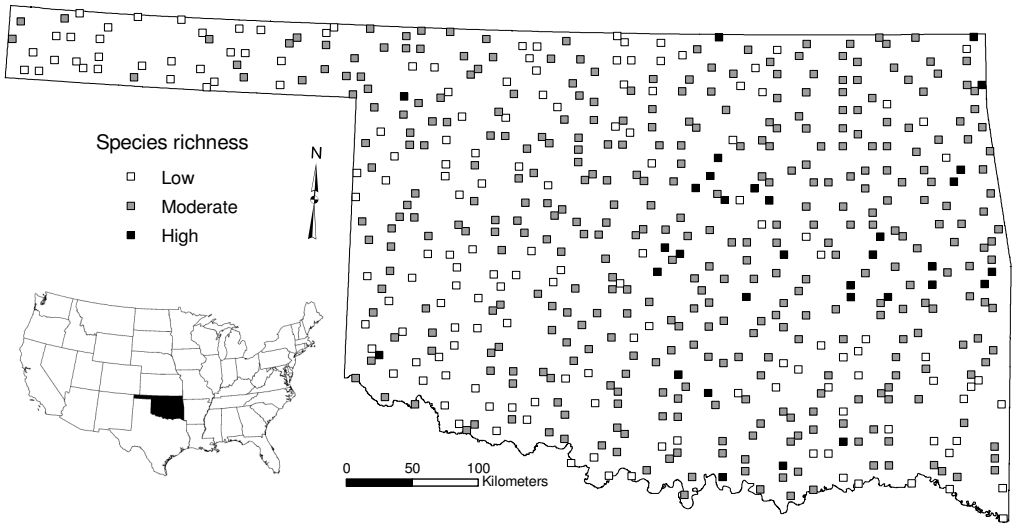

Figure 1

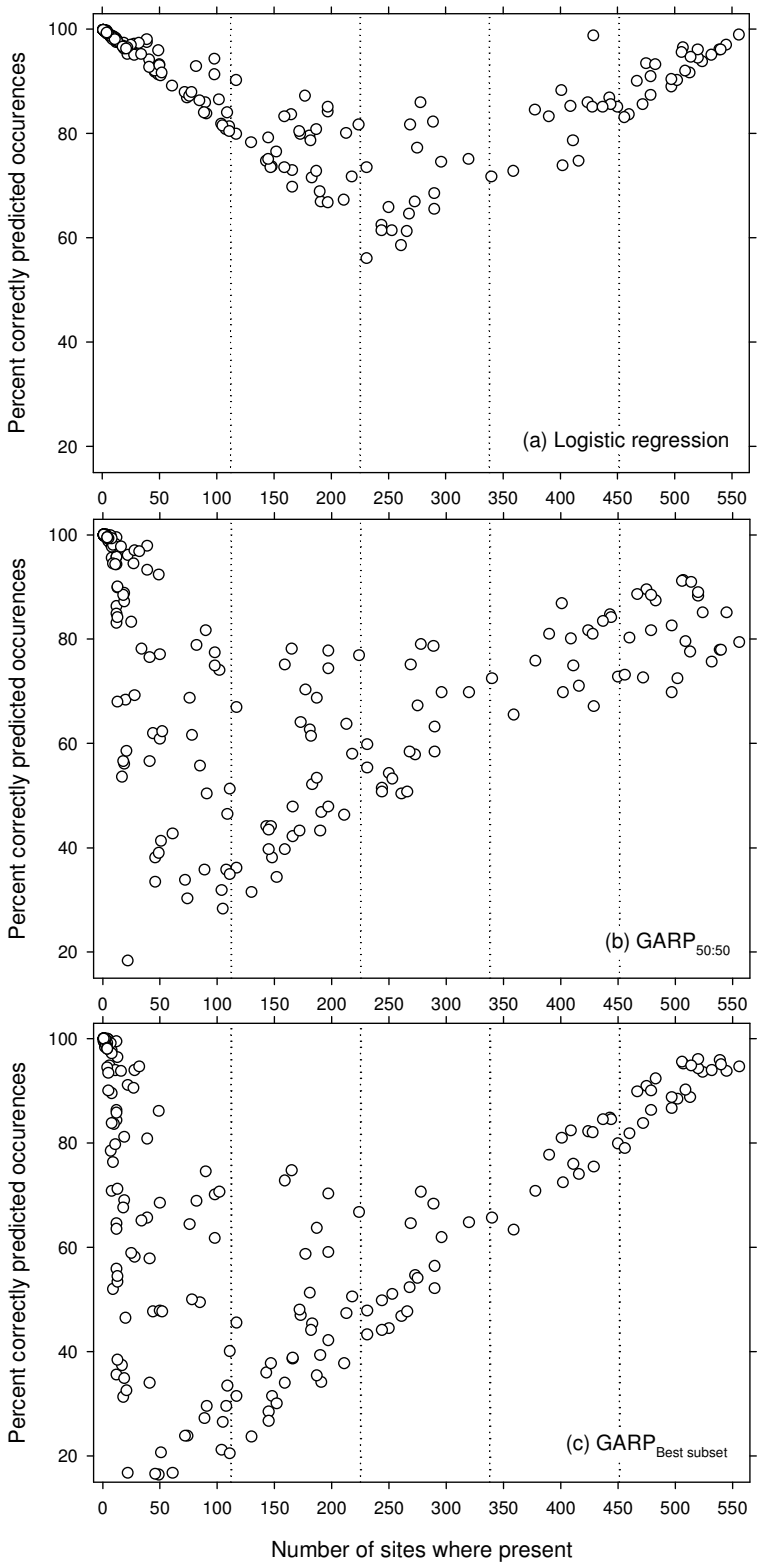
Percent correctly predicted occurences
100
80
60
40
20
0
50
100
150
200
250
300
350
400
450
500
550
(a) Logistic regression
(b) GARP50:50
(c) GARPBest subset
Number of sites where present

Figure 2

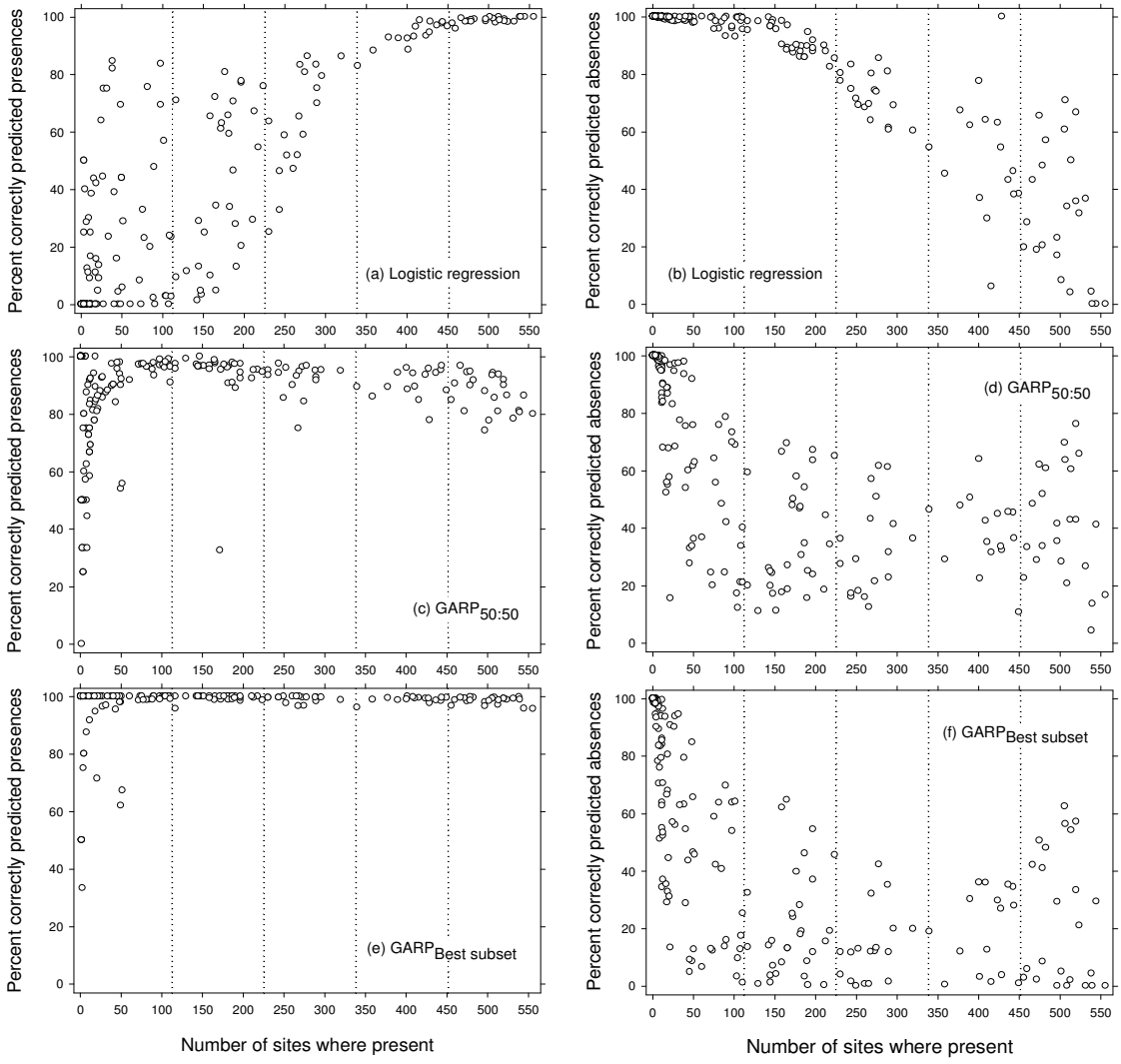
Percent correctly predicted presences
Percent correctly predicted absences
(a) Logistic regression
(b) Logistic regression
(c) GARP50:50
(d) GARP50:50
(e) GARPBest subset
(f) GARPBest subset
Number of sites where present
Number of sites where present

Figure 3

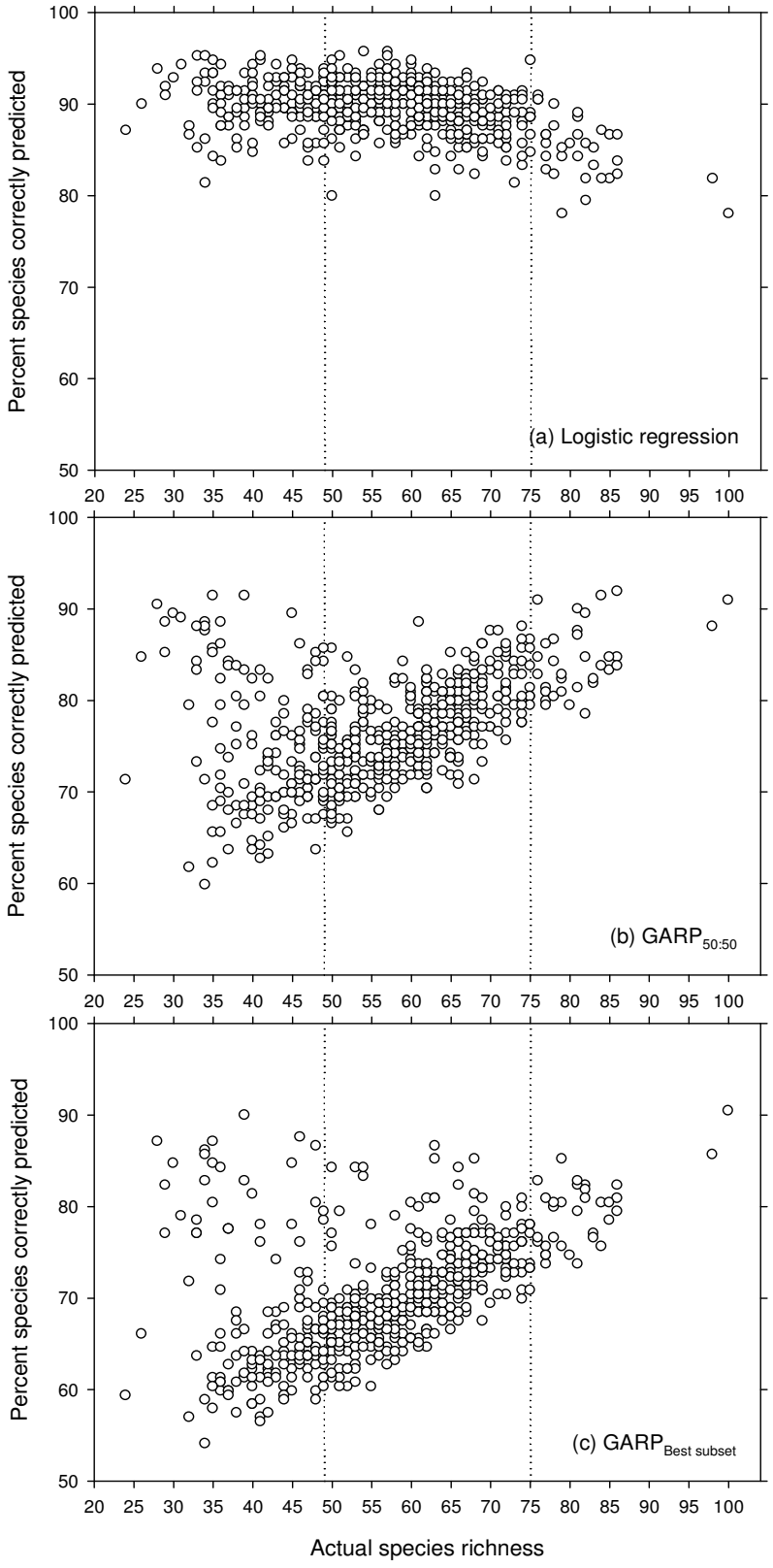
Percent species correctly predicted
(a) Logistic regression
(b) GARP50:50
(c) GARPBest subset
Actual species richness

Figure 4

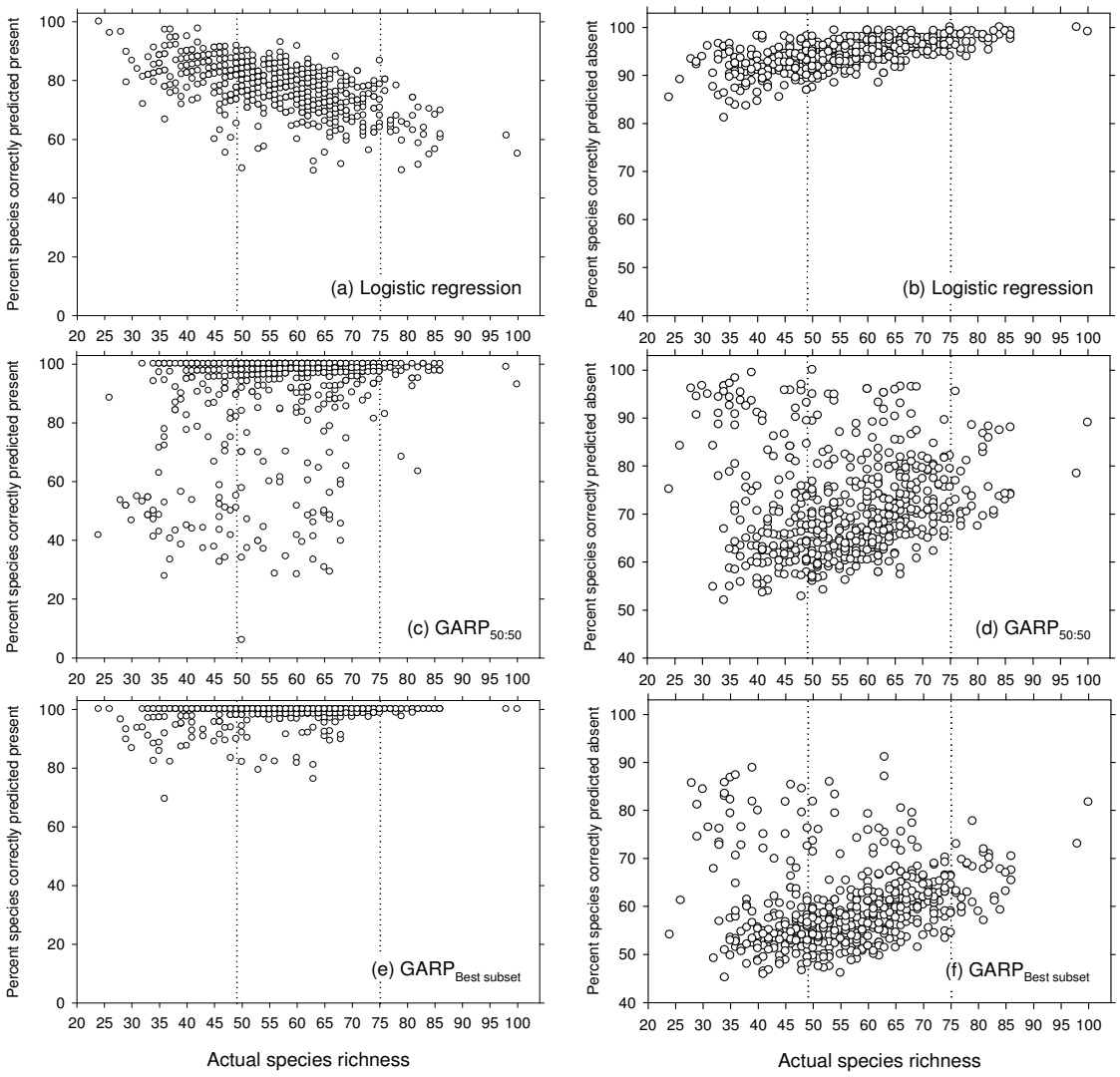
Percent species correctly predicted present
Percent species correctly predicted absent
(a) Logistic regression
(b) Logistic regression
(c) GARP50:50
(d) GARP50:50
(e) GARPBest subset
(f) GARPBest subset
Actual species richness
Actual species richness

Figure 5

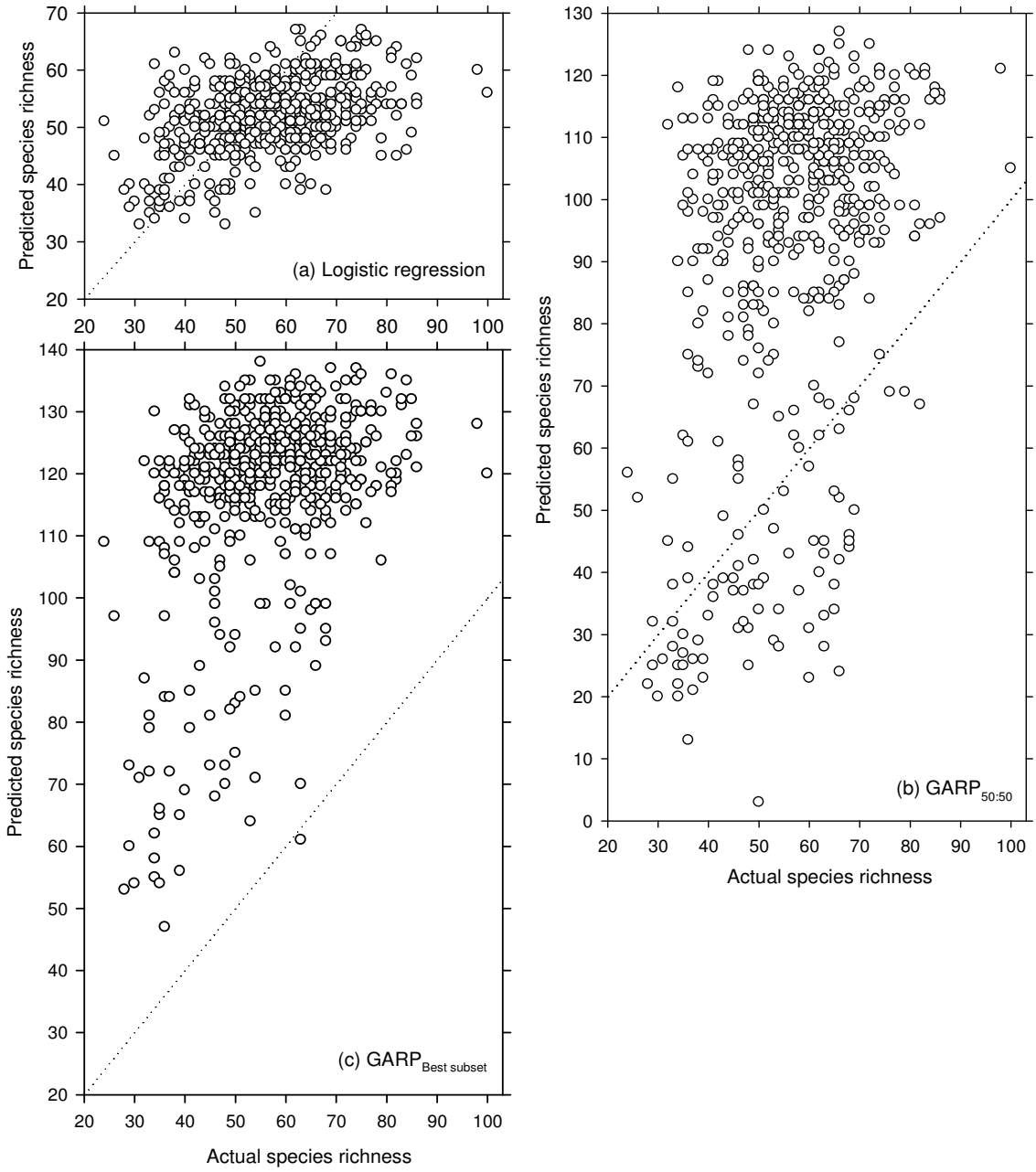
Predicted species richness
(a) Logistic regression
(b) GARP50:50
(c) GARPBest subset
Actual species richness

Figure 6