# Spatially explicit models to test climatic and land use influences on the distribution of Oklahoma grassland birds

Claire M. Curry

Jeremy Ross

Andrea Contina

Eli S. Bridge

# Introduction

Grasslands are one of the world’s most endangered ecosystems, with declines of 82.6-99.9% of tallgrass prairie, 30-99.9% of mixed-grass prairie, and 20-85.8% of short-grass prairie in the plains states and provinces of North America (Samson and Knopf 1994). Drivers of decline include land use conversion via agriculture and changes in fire and grazing regimes (Samson et al. 2004). Grassland report (Askin et al). The already tenuous status of grassland birds is further threatened by conversion to new crops resulting in permanent land use changes (Wright and Wimberly 2013), generational changes in land use (Higgins et al. 2002), changes in conservation programs for grassland habitats (Klute et al. 1997), and alterations to vegetation (Alward 1999) and ecosystem structure (Brown et al. 1997) from climate change (McCarty 2001). Grassland bird species are declining faster than other groups of birds as well (Sampson and Knopf 1994) and thus continued to be imperiled by these new threats to their habitat.

Species distribution modeling (SDM) uses predictors, usually climatic variables () but now expanding to biotic variables (), to predict what areas are most suitable for a given organism, and estimate what variables constrain a given species’ range. Range-wide predictions have been made for grassland birds (O’Connor et al. 1999) but some species with smaller ranges were not accurately modeled, perhaps because different drivers of distribution are important in different regions (Bakker et al. 2002). A new method called Spatiotemporal Exploratory Modeling (STEM) adds additional accuracy by using a stacked ensemble of smaller, regional models to account for local differences in variables that drive distributions (Fink et al. 2010). The original model was developed for data with a continent-wide scale and seasonal variation. Understanding impacts of changing land use and climate in a region with much ecosystem variety requires a dynamic approach to species distribution modeling, so we will use this approach of regional ensemble models at a smaller scale, within one state, within a single season (the breeding season). We will compare the spatially explicit ensemble’s effectiveness with a typical SDM at the statewide scale.

In the southern Great Plains, the U.S. state of Oklahoma contains a wide variety of grassland birds as its ecoregions range from tallgrass prairies in the east on the edge of eastern deciduous forests to several types of grasslands in the central part of the state and westward. Historically, each type of grassland comprised ?, ?, and ? ha, totaling ? ha (Sampson and Knopf paper). Modern estimates by comparable schemes are not available. Agriculture accounts for over $2.8 billion in the state’s gross domestic product in the study years (U.S. Bureau of Economic Analysis, 2017). Major crops in the state use. This combination of the state’s agricultural importance and forecast impact by climate change makes Oklahoma’s grassland birds vulnerable. The objectives of our study are to estimate the current distribution of Oklahoma grassland birds and understand what variables are important in their distribution. These data will allow managers to make decisions on what areas are important for preventing further decline in populations, what land use practices and trends may impact populations, and how climate change interacts with these. Considering management at the local scale is also important because many land use changes are made there (citations).

We will examine three aspects of Oklahoma grassland bird distribution. First, what is the current distribution of these species? We will use point count, transect, and citizen science data to create density estimates (from point count and transect data, comparing estimates because they are different habitats) and species distribution models (from all three data sources). We will create statewide distribution models for each species of interest and compare these with spatially explicit ensemble models at three scales. Second, what predictor variables determine the distribution of each species and as such what land use changes might make these species vulnerable? We look at land use, conservation easements, and climatic variables, and ask which variables were most important in the statewide species distribution model for each species. Finally, how will distributions change with climate change forecasts and potential land use changes? We use predicted climate change forecasts and estimates of potential land use change to estimate risks for Oklahoma’s grassland birds.

# Methods

## Study area

Brief discussion of ecoregions in Oklahoma, range of precipitation and temperature across the state, and what types of grasslands (and what areas exist) are here.

## Response data

We focused on grassland birds (Table 2).

* Survey methods
  + Point counts
  + Transects
* Citizen science data
  + eBird: All complete data (points and transects) for 2013 and 2014 for training data. 2011 and 2012 spatially sampled data for evaluation data.
  + Because some observers entered sightings from before and during our surveys into eBird.org, we eliminated ?? counts that were within two hours of the actual survey start time and within 15 km of the survey start location in 2013 and 2014.

## For presence/absence for use in species distribution models, we used survey sightings within 500 m of the point or transect. The maximum transect length was 4.3 km, so we restricted use of eBird transects to <4.3km to have comparable effort. We used the point count center or the transect midpoint as the count location to have comparable precision to eBird coordinates (Fink et al. 2010).

## Predictors

We used climate, land use (conservation easement status citation), and land cover variables to predict bird distribution. Table 1 shows the variables, sources, and their definitions. Neighborhood predictors were calculated by the values in rectangular areas around each point, at the scale of 5 x 5 pixels (150 x 150 m) and 15 x 15 pixels (450 x 450 m) (Fink et al. 2010). We looked at proportion of each land cover class and proportion of summed open space land covers (grasslands, hay/pasture, cropland, herbaceous wetlands, and barren land) since grassland bird occupancy can be influenced by the total non-structural cover (McDonald 2017). Neighborhoods were created in QGIS 2.14 with the GRASS r.neighbors processing tool (QGIS CITATION HERE).

## Analyses

### Density estimations

* Using distance sampling, possibly including detectability from repeated surveys
  + Comparison of point count vs transect effectiveness if sample size large enough for each and geographical overlap sufficient. However, point counts go along road and transects usually walking off-road. Alternative: comparison of estimations from road pcs vs “off road” transects?

### Species distribution models

To model species distributions based on our predictors, we created four models for each species at varying spatial scales: a single model statewide and three spatially explicit ensemble models. The statewide model will allow ranking of variable importance. This gives us one model for which we can interpret what predictor variables are influencing distribution. The remaining models are three spatiotemporal exploratory models (STEM) (Fink et al. 2010) at varying scales. In these models, smaller supporting models are created and then stacked to create a mosaic raster. The original STEM was used on continent-scale survey data and can be used with any base model (Fink et al. 2010, Fink et al. 2013). As our survey goals are to determine breeding distribution only, we did not specify temporal windows for this model (unlike Fink et al. 2010) and used data from April-June in all years. Hence, we will refer to our models as “spatially explicit” in this paper while adapting Fink et al’s averaging ensemble design.

Spatially explicit ensembles are likely to give more accurate predictions (Fink et al. 2010) but models with increasing local accuracy are harder to interpret and generalize (James et al. 2013). We see whether STEMs are useful at the state level (as opposed to continent-wide) by adapting the scale of our support sets to the state extent (i.e. our small, medium, and large scales are relatively smaller than those used in the original paper). With the diverse habitats and climatic variables found across Oklahoma, it should provide better predictions than the statewide model. Using both the statewide and spatially-explicit models give us complementary information on factors affecting species distribution in Oklahoma.

For all base models we used random forest classification trees (Breiman 2001). Random forest gives results competitive to other machine learning techniques such as boosted regression trees and bagged decision trees, with minimal tuning parameters required (Caruana and Niculescu-Mizil 2006, Cutler et al. 2007, Guo et al. 2010), including for species distribution models (Prasad et al. 2006, Lorena et al. 2011) and can use small sample sizes for presence records (Mi et al. 2017). The random forest algorithm bootstraps a subset of the data, fits some proportion of the predictor variables, and gives the error rate on training data using the “out of bag” sample (the portion of data not used in the bootstrap for each tree) (Hastie et al. 2001). The trees are then averaged for a final model. The use of a subset of variables per boostrapped tree also allows estimation of variable importance.

We used 7605 complete checklists (from survey and eBird data in 2013 and 2014) as the training set for models. For the eBird data, we used all data except casual counts. The statewide model was trained with the full training set (7065 checklists). We tested the models with eBird data from 2011 and 2012 (n checklists), sampled repeatedly for spatial uniformity (see below). Using data from different times results in a more accurate evaluation of whether the model generalizes well (Araújo and Guisan 2006, Cutler et al. 2007).

All models (four per species) were created using random forest classification trees in the R package randomForest (Liaw and Wiener 2002). A single random forest model was created for the statewide scale for each species using all training data, with a prediction raster created using the predict function in the R package ‘raster’ (Hijmans 2016). For the spatially explicit models, we created stratified random points in the study area (with the spsample function from the R package ‘sp’ (Pebesma and Bivand 2005, Bivand et al. 2013) and created a square of size small (200 points with 100 x 100 km), medium (100 points with 250 x 250 km), or large (25 points with 500 x 500 km) around these points. Each support set included all checklists from the training dataset located within its boundaries. The support set was discarded if it contained fewer than 25 checklists. As for the statewide model, we created a prediction raster for each support set model. All support set rasters for a given scale were stacked using the raster::mosaic function to get the mean value of each pixel, creating the spatially explicit ensemble (Hastie et al. 2001, Fink et al. 2010, Oppel et al. 2012) made of many smaller-scale random forests. This process was repeated at the three support set spatial scales, resulting in three spatially explicit, stacked ensembles per species.

To determine which predictors were important in species distributions, for each species we ranked variables using the statewide model. We used the mean decrease in accuracy and mean decrease in Gini index (define here) given by the randomForest R package to rank variable importance. We created partial dependence plots of the top variables for each statewide model to show how each variable increases or decreases probability of presence (Hastie et al. 2001, Cutler et al. 2007).

To evaluate model performance and choose between the four models for each species, we tested the models on the temporally independent evaluation dataset. To ensure spatially uniform testing (Fink et al. 2010), we created a statewide grid of ? x ? m cells. We randomly sampled no more than 5 observations from each grid cell in the evaluation dataset. We repeated the spatial sampling procedure randomly 50 times to get a distribution of performance for each model (Fink et al. 2010). We tested four models in this way for each species and chose the scale (small, medium, large, statewide) with best performance measures for each species. Performance measures were root mean square error (RMSE) and area under the receiver operating curve (AUC) more details on definitions here.

# Results

Map for STEM-type, basic model, and distance sampling estimates for each species.

# Discussion

We should be careful extrapolating climate change to extinction (Schwartz et al. 2006).

# Acknowledgements

This work was funded by USDA-NIFA grant #2013-67009-20369 to ESB. We thank technicians NAMES for help in gathering survey data. We thank eBird (look for ebird’s requested acknowledgement text).

# References

Alward, R. D. (1999). Grassland Vegetation Changes and Nocturnal Global Warming. Science 283:229–231. doi: 10.1126/science.283.5399.229

Araújo, M. B., and A. Guisan (2006). Five (or so) challenges for species distribution modelling. Journal of Biogeography 33:1677–1688. doi: 10.1111/j.1365-2699.2006.01584.x

Bakker, K. K., D. E. Naugle, and K. F. Higgins (2002). Incorporating Landscape Attributes into Models for Migratory Grassland Bird Conservation. Conservation Biology 16:1638–1646. doi: 10.1046/j.1523-1739.2002.01328.x

Bivand, R. S., E. Pebesma, and V. Gomez-Rubio (2013). Applied spatial data analysis with R. In. 2nd edition. Springer, New York.

Breiman, L. (2001). Random forests. Machine learning 45:5–32.

Brown, J. H., T. J. Valone, and C. G. Curtin (1997). Reorganization of an arid ecosystem in response to recent climate change. Proceedings of the National Academy of Sciences 94:9729–9733.

Caruana, R., and A. Niculescu-Mizil (2006). An empirical comparison of supervised learning algorithms. In Proceedings of the 23rd international conference on Machine learning. ACM, pp. 161–168.

Cutler, D. R., T. C. Edwards, K. H. Beard, A. Cutler, K. T. Hess, J. Gibson, and J. J. Lawler (2007). Random forests for classification in ecology. Ecology 88:2783–2792.

Fink, D., W. M. Hochachka, B. Zuckerberg, D. W. Winkler, B. Shaby, M. A. Munson, G. Hooker, M. Riedewald, D. Sheldon, and S. Kelling (2010). Spatiotemporal exploratory models for broad-scale survey data. Ecological Applications 20:2131–2147.

Guo, Y., A. Graber, R. N. McBurney, and R. Balasubramanian (2010). Sample size and statistical power considerations in high-dimensionality data settings: a comparative study of classification algorithms. BMC Bioinformatics 11:447. doi: 10.1186/1471-2105-11-447

Hastie, T., R. Tibshirani, and J. Friedman (2001). The elements of statistical learning. In. Springer series in statistics Springer, Berlin.

Higgins, K. F., D. E. Naugle, and K. J. Forman (2002). A Case Study of Changing Land Use Practices in the Northern Great Plains, U.S.A.: An Uncertain Future for Waterbird Conservation. Waterbirds: The International Journal of Waterbird Biology 25:42–50.

Hijmans, R. J. (2016). raster: Geographic Data Analysis and Modeling. In.

Hijmans, R. J., S. E. Cameron, J. L. Parra, P. G. Jones, and A. Jarvis (2005). Very high resolution interpolated climate surfaces for global land areas. International Journal of Climatology 25:1965–1978. doi: 10.1002/joc.1276

James, G., D. Witten, T. Hastie, and R. Tibshirani (2013). An Introduction to Statistical Learning. In. Springer New York, New York, NY.

Klute, D. S., R. J. Robel, and K. E. Kemp (1997). Will Conversion of Conservation Reserve Program (CRP) Lands to Pasture be Detrimental for Grassland Birds in Kansas? American Midland Naturalist 137:206. doi: 10.2307/2426840

Liaw, A., and M. Wiener (2002). Classification and regression by randomForest. R News 2:18–22.

Lorena, A. C., L. F. O. Jacintho, M. F. Siqueira, R. D. Giovanni, L. G. Lohmann, A. C. P. L. F. de Carvalho, and M. Yamamoto (2011). Comparing machine learning classifiers in potential distribution modelling. Expert Systems with Applications 38:5268–5275. doi: 10.1016/j.eswa.2010.10.031

McCarty (2001). 2001\_McCarty\_ecological\_consequences\_of\_climate\_change.pdf.

Mi, C., F. Huettmann, Y. Guo, X. Han, and L. Wen (2017). Why choose Random Forest to predict rare species distribution with few samples in large undersampled areas? Three Asian crane species models provide supporting evidence. PeerJ 5:e2849. doi: 10.7717/peerj.2849

Oppel, S., A. Meirinho, I. Ramírez, B. Gardner, A. F. O’Connell, P. I. Miller, and M. Louzao (2012). Comparison of five modelling techniques to predict the spatial distribution and abundance of seabirds. Biological Conservation 156:94–104. doi: 10.1016/j.biocon.2011.11.013

Pebesma, E., and R. S. Bivand (2005). Classes and methods for spatial data in R. R News 5.

Prasad, A. M., L. R. Iverson, and A. Liaw (2006). Newer Classification and Regression Tree Techniques: Bagging and Random Forests for Ecological Prediction. Ecosystems 9:181–199. doi: 10.1007/s10021-005-0054-1

Samson, F. B., F. L. Knopf, and W. Ostlie (2004). Great Plains ecosystems: past, present, and future. Wildlife Society Bulletin 32:6–15.

Samson, F., and F. L. Knopf (1994). Prairie conservation in North America. BioScience 44:418–421.

Wright, C. K., and M. C. Wimberly (2013). Recent land use change in the Western Corn Belt threatens grasslands and wetlands. Proceedings of the National Academy of Sciences 110:4134–4139. doi: 10.1073/pnas.1215404110

# Tables

Table . Predictors used in models.

|  |  |  |
| --- | --- | --- |
| Predictor | Definition | Source (needs changing to citation) |
| NRCS Conservation Easement Areas by State | Presence or absence of a conservation easement | <http://gws.ftw.nrcs.usda.gov/GWDL/3276698/easements_EASEAREA_ok_3276698_01.zip> |
| NRCS Conservation Easement Areas by State Calculated Area | Size of the conservation easement in which the pixel exists | <http://gws.ftw.nrcs.usda.gov/GWDL/3276698/easements_EASEAREA_ok_3276698_01.zip> |
| NLCD2011 Landcover | NLCD class | <http://gws.ftw.nrcs.usda.gov/GWDL/3276698/land_use_land_cover_NLCD_ok_3276698_02.zip> |
| NLCD 2.25 and 20.25 ha neighborhoods | 5 x5 and 15 x 15 pixel neighborhoods | All modified from NLCD landcover classes. |
|  | Open space (11, 31, 71, 81, 82, 95) |  |
|  | Open water 11 |  |
|  | Developed open space |  |
|  | Low intensity development (22) |  |
|  | Medium intensity development (23) |  |
|  | High intensity development (24) |  |
|  | Barren (31) |  |
|  | Forest (41, 42, 43) |  |
|  | Scrub/shrubland 52 |  |
|  | Grasslands 71 |  |
|  | Pasture and hay 81 |  |
|  | Croplands 82 |  |
|  | Woody wetlands 90 |  |
|  | Herbaceous wetlands 95 |  |
| Human population density | Number per square km | Census |
| BIO1 = Annual Mean Temperature |  | Bioclim variables from Worldclim (Hijmans et al. 2005) |
| BIO2 = Mean Diurnal Range (Mean of monthly (max temp - min temp)) |  |  |
| BIO3 = Isothermality (BIO2/BIO7) (\* 100) |  |  |
| BIO4 = Temperature Seasonality (standard deviation \*100) |  |  |
| BIO5 = Max Temperature of Warmest Month |  |  |
| BIO6 = Min Temperature of Coldest Month |  |  |
| BIO7 = Temperature Annual Range (BIO5-BIO6) |  |  |
| BIO8 = Mean Temperature of Wettest Quarter |  |  |
| BIO9 = Mean Temperature of Driest Quarter |  |  |
| BIO10 = Mean Temperature of Warmest Quarter |  |  |
| BIO11 = Mean Temperature of Coldest Quarter |  |  |
| BIO12 = Annual Precipitation |  |  |
| BIO13 = Precipitation of Wettest Month |  |  |
| BIO14 = Precipitation of Driest Month |  |  |
| BIO15 = Precipitation Seasonality (Coefficient of Variation) |  |  |
| BIO16 = Precipitation of Wettest Quarter |  |  |
| BIO17 = Precipitation of Driest Quarter |  |  |
| BIO18 = Precipitation of Warmest Quarter |  |  |
| BIO19 = Precipitation of Coldest Quarter |  |  |
|  |  |  |

Table 2. Study species with their conservation status.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Species | Breeding bird survey trend since YEAR | Oklahoma status | Federal status | IUCN status |
| Dickcissel |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |

# Figures