PREDICTIVE DISTRIBUTION MODELING OF SPECIES OF GREATEST CONSERVATION NEED IN TEXAS

Prepared by

Mark Andersen and Gary Beauvais

Wyoming Natural Diversity Database

University of Wyoming

Laramie, Wyoming

August 31, 2013

__Suggested citation for this report__: Andersen, M.D. and G.P. Beauvais. 2013. Predictive Distribution Modeling of Species of Greatest Conservation Need in Texas. Report prepared by the Wyoming Natural Diversity Database, Laramie Wyoming, for the Texas Natural Diversity Database, Texas Parks and Wildlife Department, Austin, TX. August 31, 2013.

CONTENTS

EXECUTIVE SUMMARY

Texas Parks and Wildlife Department (TPWD) identified over 1,300 Species of Greatest Conservation Need (SGCN) in their Texas Conservation Action Plan (TCAP)[1]. Understanding their geographic distributions is critical to efficient and effective management of these species. Occurrence data held in public databases (i.e., "dot maps") provide an incomplete picture of distribution, as these data tend to derive from opportunistic and often spatially-biased sampling. Predictive species distribution models quantify the relationships between spatially-referenced occurrence data for a species and the underlying environmental gradients, resulting in a map of predicted distribution, or potential occupancy, for a species. Such models provide a more comprehensive spatial representation of distribution that can be used to inform survey, assessment, and management of SGCNs.

Given the time and effort that would be required to generate predictive distribution models for the full set of SGCN, it was necessary prioritize species for initial distribution modeling work. To this end, the Texas Natural Diversity Database (TXNDD) and TPWD identified an initial set of 26 priority SGCN spanning a broad range of taxonomic groups on the basis of species characteristics and data availability. The Wyoming Natural Diversity Database (WYNDD), in fulfillment of an agreement with TPWD/TXNDD, generated predictive distribution models for this subset of Texas' SGCN. The resulting distribution models provide a replicable and scientifically rigorous estimate of each species' distribution within the State, and help set the stage for future modeling work to be carried out by TPWD/TXNDD.

In addition to this report, the project database, and the digital files containing model input and output datasets, WYNDD provided: (1) a modeling guidance document comprising a review of literature related to species distribution modeling; and (2) 3 days of direct consultation with and technical training of a TXPWD GIS specialist and species modeler. It is hoped that these initial models for SGCN, the supporting documentation, and the technical consultations and training will allow TXNDD and TPWD to continue the task of generating predictive distribution models for SGCN in the State of Texas.

INTRODUCTION

TEXAS' SPECIES OF GREATEST CONSERVATION NEED

Texas Parks and Wildlife Department (TPWD) is charged with developing the State's wildlife conservation plan[2]. In 2009, (TPWD) and its collaborators began reviewing the 2005 Comprehensive Wildlife Conservation Strategy[2], and based on this review, produced a revised plan, referred to as the Texas Conservation Action Plan (TCAP)[1]. In support of the TCAP, The working group produced a revised Species of Greatest Conservation Need (SGCN) list, based the 2005 Species of Greatest Conservation Need list[2], State and Federal protected species lists, NatureServe rankings [3], and other information related to species biology and life history[1]. This revised SGCN list identified over 1,300 species warranting attention by land and resource managers. With limited resources available for distribution modeling, TPWD/TXNDD developed a list of 26 target modeling taxa for this project based on conservation need and data availability (Table 1).

PREDICTIVE SPECIES DISTRIBUTION MODELING

One key to effective resource management is understanding the geographic distribution of the resource in question. The Texas Natural Diversity Database maintains occurrence data representing observations of individuals or populations of a species. These observation data, referred to variously as "observations," "occurrences," or "presence points," are often precise in their spatial location, but typically provide an incomplete picture of a species' distribution. A key problem is that negative data (i.e., locations where a species was surveyed for but not observed) are rarely recorded. Thus, it is often unclear whether the blank areas on these "dot maps" of species observations represent unoccupied areas, or are simply areas that were never sampled for the species. Likewise, it is unclear whether clusters of observations reflect areas of high suitability for a species, or are merely the product of uneven or biased sampling effort.

Distribution modeling has become a common method for filling in these blank areas with a prediction for occupancy by a species (i.e., present or absent). Deductive distribution models use expert knowledge to create a rule set that predicts suitability for occupancy based on important environmental characteristics of the landscape. For example, previous distribution modeling efforts at a statewide scale have generated deductive models for species by reclassifying land cover maps into suitable/unsuitable categories[4].

Inductive distribution models use statistical or machine learning methods to identify relationships between points of known presence or absence and the underlying environmental gradients, and model these relationships to allow the prediction of the species' distribution across the study area[5]. Whereas standard statistical methods for predicting the probability of a binary outcome (e.g., logistic regression) require training data for both classes (i.e., species presence/absence), only presence data are typically available in species observation databases. Specialized methods have thus been developed for producing species distribution models from presence-only data [6,7]. While these methods have proven effective in distribution modeling efforts [8,9], the resulting distribution models must be built, conveyed, and interpreted differently from those that might be generated using conventional methods such as logistic regression[10].

Although references in the literature typically refer to "*species* distribution modeling," we hereafter use the more general term 'taxon' rather than the more specific term 'species,' recognizing that several taxa on the target list are subspecies or specific populations. We used inductive modeling with a commonly applied algorithm and proven methods to generate models for these 27 taxa in Texas.

While all distribution models are subject to error, as with any model, they generally offer a more complete and useful representation of a taxon's distribution than do dot maps of observations. The resulting models can be used to guide surveys for new populations, or to assess potential overlap between modeled distributions and planned management activities or disturbances. In addition to model input and output data and documentation of our methods, we prepared a number of documents and scripts to facilitate future modeling work by TXNDD staff. Importantly, we also directly consulted with and trained TPWD modelers in the application of our methods over a 3-day period in August 2013.

METHODS

OVERVIEW

Occurrence data used for training models for the target taxa were derived from downloads of TXNDD's observation database. Occurrence data for non-target taxa obtained from the same database were used as background, or pseudo-absence data to allow a distinction between gradients at presence points for a target taxon and those gradients available on the landscape. These training presence and background data were related to GIS layers representing a suite of biologically-relevant environmental gradients using Maxent, a machine-learning algorithm for inductive modeling of distribution based on presence-only points[6]. The resulting models were "projected" onto the environmental gradient data to produce maps showing predicted distribution for the target taxa. We chose Maxent because it has repeatedly been shown to perform well with relatively small sample sizes, and does not require absence data for generating useful models. Model training, evaluation, and assessment were carried out using methods commonly employed in distribution modeling.

PRESENCE DATA COLLECTION AND PROCESSING

TXNDD was the sole source of occurrence data used in building models, and provided approximately 9,000 observation records of the target taxa from their Biotics database between May 28 and August 1 of 2013. The number of observation records available by taxon varied dramatically, from over 4,400 records for Black-tailed Prairie Dog (see Table 1 for Latin names of all taxa mentioned in the text), to just 21 total records for Chihuahua Balloon-vine (Table 1).

Since taxa may substantially shift their distributions over time in response to changes in climate and land use patterns, relating historical records to the environmental gradients might not produce a model that accurately predicts current distribution. Thus, biologists at the TXNDD provided record age cutoffs for select taxa (Table 1), corresponding to the timing of major distribution shifts, or, in some cases, shifts in data collection methods. Observation records for a given taxa collected prior to these cutoffs were excluded from modeling.

Table 1. Occurrence data by taxa. Total records indicates the number of records for the taxon provided by TXNDD. Modeling records indicates the number of records used for model training after filtering for record age, precision, season (where appropriate), and clustering.

Common Name	Scientific Name	Total Records	Modeling Records
Amphibians			
Sheep Frog	*Hypopachus variolosus*	46	45
Birds			
Lesser Prairie-chicken	*Tympanuchus pallidicinctus*	323	247
Piping Plover	*Charadrius melodus*	748	127
Black-capped Vireo	*Vireo atricapilla*	664	63
Bachman's Sparrow	*Aimophila aestivalis*	75	31
Mammals			
Black-tailed Prairie Dog	*Cynomys ludovicianus*	4,425	2,958
Texas Kangaroo Rat	*Dipodomys elator*	262	122
Swift Fox	*Vulpes velox*	42	41
Kit Fox	*Vulpes macrotis*	36	35
Black Bear – Western TX	*Ursus americanus*	115	108
Black Bear – Eastern TX	*Ursus americanus*	45	40
Ocelot	*Leopardus pardalis*	42	21
Reptiles			
Texas Tortoise	*Gopherus berlandieri*	99	61
Reticulate Collared Lizard	*Crotaphytus reticulatus*	40	39
Spot-tailed Earless Lizard	*Holbrookia lacerata*	142	133
Texas Horned Lizard	*Phrynosoma cornutum*	42	40
Texas Indigo Snake	*Drymarchon melanurus erebennus*	67	53
Louisiana Pine Snake	*Pituophis ruthveni*	37	29
Plants			
Texas Prairie Dawn	*Hymenoxys texana*	75	40
Threeflower Broomweed	*Thurovia triflora*	36	29
Zapata Bladderpod	*Physaria thamnophila*	96	12
Bracted Twistflower	*Streptanthus bracteatus*	282	25
Tobusch Fishhook Cactus	*Sclerocactus brevihamatus ssp. tobuschii*	322	83
Texabama Croton	*Croton alabamensis var. texensis*	37	16
Johnston's Frankenia	*Frankenia johnstonii*	145	92
Chihuahua Balloon-vine	*Cardiospermum dissectum*	21	16
Navasota Ladies'-tresses	*Spiranthes parksii*	908	74

Additionally, observation records may be very coarsely located geographically. For example, historical records from museum collections are often located only to county, or via crude spatial descriptions (e.g., "10 miles southwest of Austin"). Such records may provide information about a taxon's range at a very coarse scale, but generally do not provide useful information for building predictive distribution models. TXNDD, like most state heritage programs, maintains a field in their database representing the locational uncertainty, or mapping precision, of each record. Records with a locational uncertainty of greater than 8000 m were excluded from the modeling set. This value was chosen to balance spatial accuracy with the number of records available by taxa.

Two of the target species, Piping Plover and Black-capped Vireo, only occur seasonally in Texas. Piping Plover is a winter resident, so only observations recorded between July 15 and May 1 were included in modeling. Black-capped Vireo is a summer resident; records collected between March 15 and September 7 were used in modeling the species. These calendar date cutoffs were used to avoid modeling with migratory records, which may not be representative of winter and summer distributions for Piping Plover and Black-capped Vireo, respectively.

Lastly, occurrence points can exhibit clustering at multiple spatial scales due to sampling bias. For example, a single study for a taxon may result in many points within a Public Land Survey System (PLSS) section, and no points in the adjacent sections, even though the taxon may occur in these sections as well. At a broader scale, sampling bias may be related to accessibility, leading to more points near roads and cities and fewer points in remote areas with difficult access or on private lands. A distribution modeling algorithm effectively interprets higher densities of observations as higher suitability for a taxon, when in fact the higher densities are often due to more intensive sampling efforts. Thus, both of these types of sampling bias can lead to incorrect conclusions and models.

We used a spatial filtering process based on occurrence point quality described by Keinath et al[9] to address bias at the scale of an individual study. This approach eliminates occurrence points that are within a user-specified, minimum separation distance of other, higher quality points for the taxon, preserving the best available points while simultaneously reducing the amount of local-scale point clustering. Occurrence point quality can be measured in a variety of ways, but the key components of record accuracy are age, locational uncertainty, and reliability of taxonomic identification. TXNDD is confident in the identification of the taxa in each of their observation records, so we developed a Point Quality Index (PQI) based on a linear combination of scored record age (PQI-Date) and scored locational uncertainty (PQI-Precision), to indicate the overall quality of a record (Table 2).

To generate PQI-Date scores, we broke the range of acceptable dates for a taxon into four equal-sized bins. For taxa where biologists provided a specific record age threshold, the selected year was used as the oldest acceptable date, and the subsequent period was broken into four equal-sized bins for scoring purposes. If no specific record age cutoff was specified by biologists for a taxon, the oldest record available for the taxon was used as the oldest acceptable date, and the subsequent period was similarly divided into scoring bins. Records with an unknown year of observation were included as potential modeling records only if there was no cutoff provided by biologists for the taxon; these records were assigned the lowest PQI-Date score. To generate the PQI-Precision score, we used a locational uncertainty of 8000 m as the threshold for usable records. Records with a locational uncertainty better than this level were scored on a four-point system, similar to the PQI-Date system (Table 3). Occurrence data for Black-tailed Prairie Dog were based on a recent mapping of colonies, and were all scored equally for both PQI-Date and PQI-Precision.

Table 2. Classes for scoring occurrence points based on their date. Scores ranged from 1 to 4, and used taxon-specific date ranges.

Taxon	PQI-Date Score				
	Unusable	***1***	***2***	***3***	***4***
Bachman's Sparrow	NA	2005-2007	2008-2009	2010-2011	2012-2013
Black Bear	< 1995	1995-2000	2001-2004	2005-2009	2010-2013
Black-capped Vireo	< 2003	2003-2006	2007-2008	2009-2011	2012-2013
Black-Tailed Prairie Dog	NA	NA	NA	NA	All
Bracted Twistflower	< 1980	1993-1998	1999-2003	2004-2008	2009-2013
Chihuahua Balloon-vine	< 1970	1974-1984	1985-1994	1995-2003	2004-2013
Johnston's Frankenia	< 1980	1994-1999	2000-2004	2005-2008	2009-2013
Kit Fox	NA	1950-1966	1967-1982	1983-1997	1998-2013
Lesser Prairie-chicken	NA	2010	2011	2012	2013
Louisiana Pine Snake	NA	1956-1970	1971-1985	1986-1999	2000-2013
Navasota Ladies'-tresses	NA	1905-1932	1933-1959	1960-1986	1987-2013
Ocelot	< 1995	1997-2001	2002-2005	2006-2009	2010-2013
Piping Plover	NA	1988-1994	1995-2001	2002-2007	2008-2013
Reticulate Collared Lizard	NA	1933-1953	1954-1973	1974-1993	1994-2013
Sheep Frog	NA	1923-1946	1947-1968	1969-1991	1992-2013
Spot-tailed Earless Lizard	NA	1902-1930	1931-1958	1959-1985	1986-2013
Swift Fox	NA	1933-1953	1954-1973	1974-1993	1994-2013
Texabama Croton	< 1980	1990-1996	1997-2002	2003-2007	2008-2013
Texas Horned Lizard	NA	1991-1997	1998-2002	2003-2008	2009-2013
Texas Indigo Snake	NA	1967-1979	1980-1990	1991-2002	2003-2013
Texas Kangaroo Rat	< 1995	1996-2000	2001-2005	2006-2009	2010-2013
Texas Prairie Dawn	NA	1986-1993	1994-2000	2001-2006	2007-2013
Texas Tortoise	NA	1960-1973	1974-1987	1988-2000	2001-2013
Threeflower Broomweed	NA	1905-1932	1933-1959	1960-1986	1987-2013
Tobusch Fishhook Cactus	< 1980	1983-1991	1992-1998	1999-2006	2007-2013
Zapata Bladderpod	< 1990	1994-1999	2000-2004	2005-2008	2009-2013

Table 3. Classes for scoring occurrence points based on their location uncertainty, in meters. As with the PQI-Date score, scores for PQI-Precision ranged from 1 to 4. All taxon were scored using the same ranges of locational uncertainty values.

PQI-Precision Score	Definition
4	Location uncertainty ≤ 30 m
3	Location uncertainty > 30 m and ≤ 100 m
2	Location uncertainty > 100 m and ≤ 300 m
1	Location uncertainty > 300 m and ≤ 600 m
0	Location uncertainty > 600 m and ≤ 8000 m
NA	Record is unusable for modeling; uncertainty > 8000 m

An Overall PQI score for each occurrence point was generated by summing the scores for record age (PQI-Date) and for locational uncertainty (PQI-Precision). We developed a script in ArcGIS to remove any points for the taxon that were within 800 m of another point with a higher PQI score. We used 800 m as the minimum separation distance as a circle with this radius is similar in size to the PLSS section that is a common sampling frame for wildlife surveys.

Background Data Collection and Processing

With standard statistical modeling for a binary event such as presence/absence modeling for a taxon, both presence and absence data would be required to model its distribution. True absence data for a given taxon are seldom available. Thus, methods have been developed to use presence-only datasets to generate distribution models, using what are often termed "background" or "pseudo-absence" data[11]. Such methods use background data in order to distinguish between the environmental gradients present in areas that are "used by" versus those that are "available to" the target taxon. A default method for creating a background dataset is to select a large number (e.g., 10,000) of random points from the modeling area to represent the gradients available to a taxon[12].

Unfortunately, presence-only data contained in databases like TXNDD's Biotics database are seldom the product of random or exhaustive, systematic sampling efforts. Typically, these data derive from biased sampling that focuses more survey effort in areas that are easily accessible (e.g., near roads or populated places, on public lands, or in priority conservation areas), and less effort in areas that are more difficult to access due to rugged terrain, lack of roads, or private land ownership[13].

These types of broader-scale sampling biases were addressed by using a target background group approach[14], rather than the default method of selecting random background points. This approach attempts to mirror spatial sampling bias in the training occurrence data for a taxon by selecting background data – often occurrence locations for related taxa – that derive from surveys exhibiting similar spatial biases as those for the modeled taxon. Matching the biases in the training occurrences for a taxon with similar biases in the background data helps to factor out sampling bias in modeling, resulting in a model that more accurately reflects a taxon's distribution. If sampling bias is not accounted for, a presence-only modeling approach may produce a model that predicts sampling effort better than it predicts a taxon's true distribution[15]. When selecting target background data, the key consideration is ensuring that the points in the background dataset derive from surveys as similar as possible to those for the modeled taxon in terms of sampling effort, methods, and biases[14].

We evaluated data from a number of sources, including Herps of Texas[16] and eBird[17], to use as target background data for the modeled herptiles and birds, respectively. However, upon examining these data it was clear that different biases were present in these two datasets than were present in the training data from TXNDD's Biotics database. As citizen-science datasets, observations in Herps of Texas and eBird tend to be more strongly biased toward populated areas than observations in the TXNDD database.

Instead, we used a download of occurrences for non-modeled taxa from TXNDD's Biotics database, as our background dataset. This dataset comprised 8,665 occurrence locations represented as points, lines, and polygons, which we converted to points. This background dataset appeared to exhibit similar spatial biases as the training presences for modeled taxa. This is a reasonable expectation, given occurrence records are generally incorporated into the database from similar data sources across all taxa.

Environmental Data Collection and Processing

The predictor data used to build distribution models represent environmental characteristics or gradients assumed to be important in influencing species distributions, and are typically stored as raster datasets in Geographical Information Systems (GIS). Modelers commonly include predictor layers describing gradients related to climate, vegetation, elevation, and soils, but for selected taxa, more specific predictors representing various aspects of landscape pattern, hydrology, interspecific interactions, or disturbance may be important in limiting distribution[5].

The linkages between a distribution and these predictor layers may be direct, as in the case of a tree-nesting bird species that only occurs in forested land cover types. However, predictor layers used in building distribution models are often more indirectly related to distribution. For example, a plant species' distribution may be limited to areas with a particular soil moisture regime that is not directly represented with available GIS layers. Instead, indirect measures of site moisture such as slope or aspect might prove useful in modeling the species. Thus, a useful predictor set may contain attributes that are intuitively important to a taxon as well as attributes that are somewhat harder to interpret.

The factors that influence a taxon's distribution vary across differing spatial scales, from broad-scale gradients like climate to fine-scale parameters such as soil texture[18]. Accordingly, the spatial predictor layers used to build distribution models should represent a similar range of scales in order to produce the most reasonable models [19].

We reviewed available references [3,20-31] for the modeling taxa to generate a list of potentially useful predictor data layers for each taxon. Biologists from TXNDD and other cooperators provided comments on these lists and in some cases suggested additional predictor data layers that might be useful. We added standard climatic, elevation, and vegetation predictors to the list of potential predictors that were initially identified.

The full list of potential predictors included data layers related to climate, topography, land use/land cover, soils and substrate, and surface water. Climatic variables were downloaded from the WorldClim website (http://www.worldclim.org/current) and included the 30 arc-second Bioclim data, representing useful seasonal and monthly means, ranges, and extremes of temperature and precipitation[32]. Topographic variables were derived from the National Elevation Dataset[33] using a variety of transformations to provide representations of important topographic attributes, including elevation, slope, aspect, ruggedness, and site moisture. Hydrology predictors quantified Euclidean distance to water and prevalence of water on the landscape, based on hydrology layers prepared by the National GAP Program[34]. Land use variables included layers representing agricultural lands and the level of human impact across varying scales, and were also based on data prepared by the National GAP Program[35]. Land cover variables included vegetation height and percent cover for forest, shrubs, and herbaceous plants from the LANDFIRE dataset[36,37], and distance to ecotone boundaries from the National GAP Program[38]. Soils predictors described chemistry, texture, and moisture parameters derived from the Gridded SSURGO dataset[39]. Appendix 1 provides more detailed descriptions and references for each variable.

Model Generation, Validation, and Display

We used the Geospatial Modeling Environment (GME[40]) to attribute the shapefiles representing training presences and background points with values for all predictor variables, and exported the associated attribute tables as comma-delimited (CSV) files. Fields were added or deleted as needed to create the Samples with Data (SWD) format used by Maxent[41]. Providing the SWD format to Maxent, rather than providing only the spatial coordinates for each taxon's training points, decreases modeling time as the software does not need to sample predictor values to training and background data points during each iterative run.

Multicollinearity (i.e., strong correlations between predictor variables) can increase the standard errors of coefficients in regression[42], changing the interpretation of which predictors are most important in a model. Similar problems arise even when using machine learning methods such as Maxent[5]. Thus, we evaluated our set of predictor layers for pairwise correlations using a random sampling of approximately 10,000 points in using the statistical software, R[43]. We identified any predictor pairs where the pairwise Pearson's R was greater than 0.8, a relatively conservative value[44].

Bioclim variable pairs exhibited the highest degree of correlation. All but three of these predictors were highly correlated with at least one other Bioclim predictor ($R > 0.8$). Thus, we first generated a Maxent model for each taxon using only the Bioclim variables, to determine which subset of these predictors were the most powerful across all taxa. The Bioclim predictors were then ranked in descending order based on their mean percent contribution across taxa. Bioclim predictors that were highly correlated with another Bioclim predictor with a higher mean percent contribution were excluded from the set of potential predictors used in the initial models for each taxon.

Three to four successive Maxent models were generated iteratively for each taxon to identify a set of powerful predictors with low collinearity, starting with the selected Bioclim subset plus the taxa-specific predictors identified during the initial taxa review or in comments from TXNDD biologists. At each iteration, we eliminated variables that had a mean percent contribution of 0 or were highly correlated with another predictor that had a higher mean percent contribution. For this iterative process[5] in Maxent, we selected the default parameters for regularization, the "jackknife" option to determine the relative importance of each predictor, and the 10-fold cross-validation option to protect against overfitting during predictor selection. Final models were then generated using the reduced variable set identified through the iterative process, writing logistic model output as a BIL-format raster.

Models were validated using several approaches, depending upon the number of available occurrences. For all taxa, a model using 10-fold cross-validation was run using the same predictor variable set and parameters, and the mean area under curve (AUC) of the Receiver Operating Characteristic (ROC) plot was calculated based on the test set for each replicate to provide a measure of model accuracy[45]. Taxa with greater than 100 training points also had a randomly selected 20% of their training points excluded during model generation to provide an independent test dataset. Models were then run based on the remaining 80% of the training set, and model accuracy was assessed by calculating AUC for the 20% test set.

A Maxent model produced with the above methods is essentially a relative probability surface, with each map cell associated with a value representing the likelihood of occurrence by a taxon. Many applications of such models require that that they first be converted into binary maps of likely presence versus likely absence. Such conversion requires the selection of a cutpoint, or threshold, on the probability scale that distinguishes predicted presence from predicted absence. We applied a threshold to create a binary (i.e., "predicted absent/predicted present") expression of the final model for each taxa. The threshold was chosen based on the "Maximum Training Sensitivity Plus Specificity" metric provided by Maxent[41]; this threshold metric minimizes the total of omission and commission error. Additionally, Maxent was run using 10-fold cross-validation for each taxon to generate a surface representing the standard deviation of predictions across all replicates. The resulting raster output provides an indication of the level of uncertainty inherent in the final model.

RESULTS

SUMMARY OF MODEL INPUT DATA

The number of observation records removed from the training dataset due to record age, locational uncertainty, or spatial clustering varied substantially between taxa (Table 1). The proportion of points removed because they were older than the threshold specified by biologists was relatively high for Black-capped Vireo (346/664; 52%), Texas Kangaroo Rat (70/262; 27%), and Ocelot (13/42; 31%). Observation records for the plant taxa were the most spatially clustered of any group, with a substantial proportion of available points removed during the spatial filtering process for Navasota Ladies-tresses (834/908; 92%), Bracted Twistflower (254/282; 91%), Zapata Bladderpod (83/96; 87%), and Tobusch Fishook Cactus (224/322; 73%). For animal taxa, only Piping Plover (621/748; 83%) and Black-capped Vireo (252/315; 80%) had a similar proportion of their available point set reduced by spatial filtering. For taxa including Sheep Frog, Swift Fox, Kit Fox, Reticulate Collared Lizard, Spot-tailed Earless Lizard, and Texas Horned Lizard, nearly all available observation records were used as training points in modeling. This likely results from less intensive and targeted sampling efforts for these species compared to those where more points were eliminated due to spatial clustering.

Bird and plant taxa generally had the most recent records in their filtered modeling sets, while records for reptiles and the single amphibian were generally much older (Figure 1). Similarly, modeling points for plants and birds generally had the highest degree of spatial precision, while mapping precision for records of reptiles, amphibians, and mammals were generally worse (Figure 2). The lack of precision in amphibian and reptile records is likely due to their age; older observation records tend to have been less precisely located than modern records that are often mapped using GPS. The least precise records for mammals were generally for large, wary taxa for which locations are typically recorded from some distance away compared to less mobile or wary taxa like small mammals, herptiles, plants, or birds that can generally be located precisely by a nearby observer using GPS receivers. These patterns of record precision and age by taxonomic group are generally consistent with those observed in similar, multi-taxa distribution modeling projects for Wyoming[9] and the northwestern United States[8].

Figure 1. Beanplots [46] showing the distribution of record ages by taxon. Width of the gray areas corresponds to the density of observations around a particular year. The vertical black lines show the average year of observation for a taxon, while the dashed line shows the average year of observation across all taxa. Black-tailed Prairie Dog is not included in the chart as no year of observation was provided for the species' records.

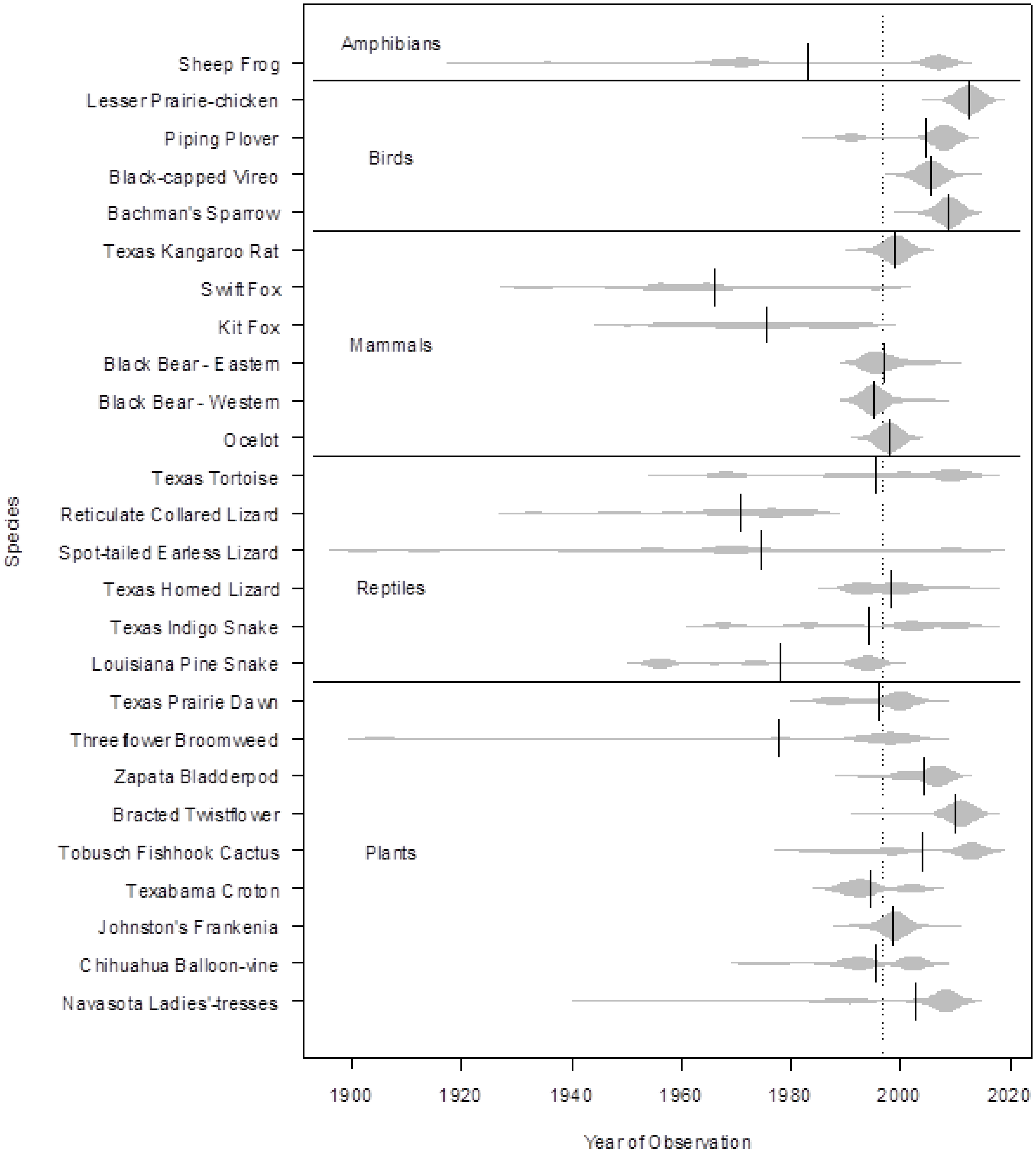
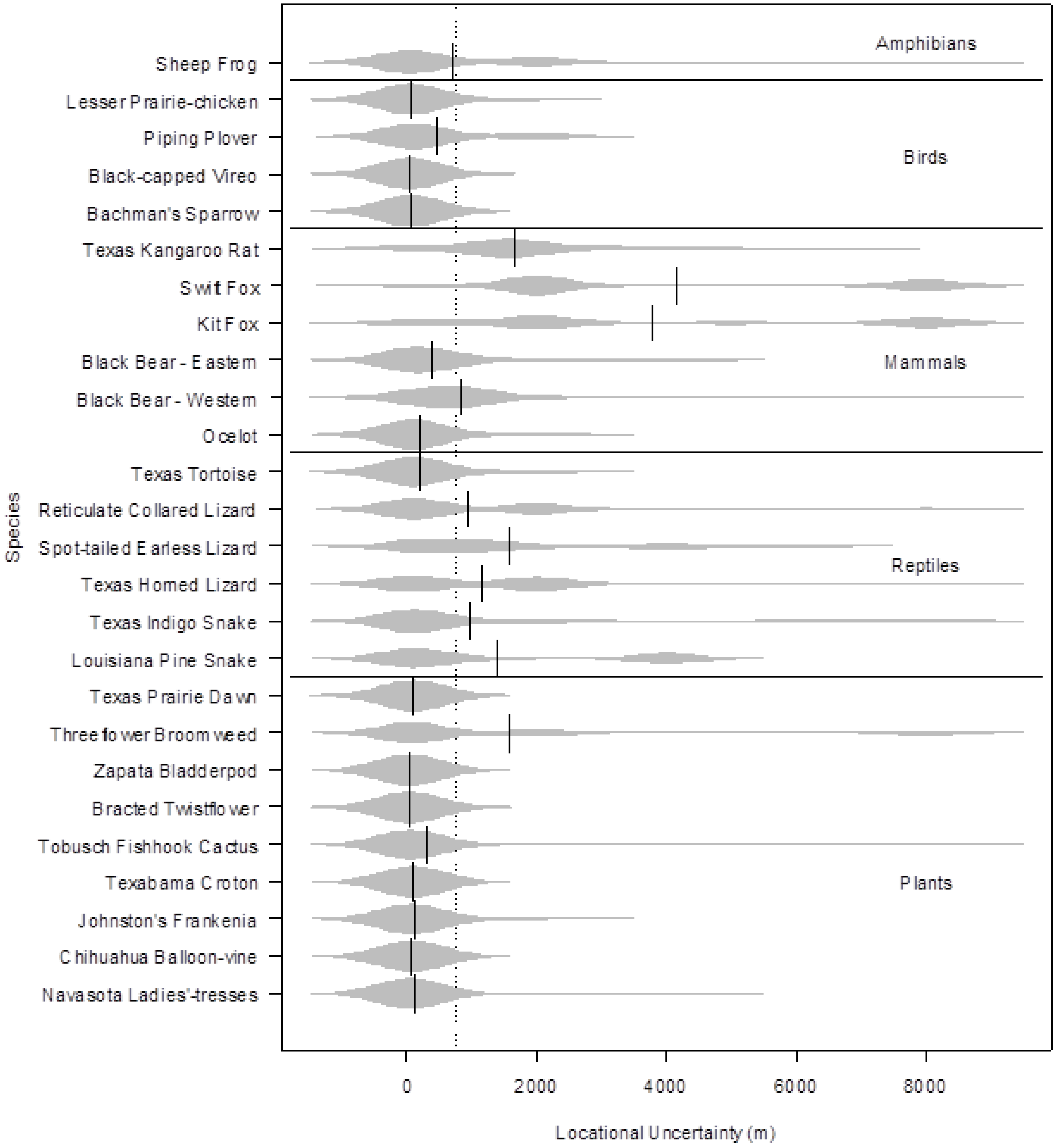


Figure 2. Beanplots showing the distribution of locational uncertainty by taxon. Width of the gray areas corresponds to the density of observations around a particular locational uncertainty value. The vertical black lines show the average locational uncertainty for records of a given taxon, while the dashed line shows the average locational uncertainty for records across all taxa. Black-tailed Prairie Dog is not included in the chart as no locational uncertainty was provided for the species' records.



A subset of six Bioclim layers was selected based on Maxent model runs for all taxa and evaluation of pairwise correlation coefficients for the Bioclim predictor set (Table 4). These six predictors appear to provide the most useful information about distributions for the modeling taxa in Texas, while minimizing collinearity. These predictors were included in the initial model runs for all taxa, along with the taxon-specific predictors we identified.

Table 4. Bioclim predictor evaluation. Bolded predictors were included in the core set of six Bioclim predictors used in the initial models for all taxa; italicized predictors were not included in initial models for any taxa, as they were highly correlated with other Bioclim predictors with a higher mean percent contribution in the variable evaluation model runs. Appendix 1 provides a full description of all predictor layers.

Bioclim Predictor [Predictor ID]	**Mean Percent Contribution**	**Strongest Correlation with More Powerful Predictor(s): Correlated Predictor (*Pearson's R)***
Min. Temperature of Coldest Month [Bio6]	12.6%	-
Precipitation Seasonality (Coefficient of Variation) [Bio15]	10.8%	Bio6 (-*0.47*)
Mean Diurnal Range (Mean of monthly (max temp - min temp) [Bio2]	6.2%	Bio6 (-*0.77*)
Precipitation of Warmest Quarter [Bio18]	4.8%	Bio15 (-*0.65*)
Mean Temperature of Warmest Quarter [Bio10]	4.4%	Bio6 (*0.77*)
Isothermality (BIO2/BIO7) (* 100) [Bio3]	2.7%	Bio15 (*0.67*)
Annual Mean Temperature [Bio1]	11.8%	Bio6 (0.96)
Precipitation of Coldest Quarter [Bio19]	8.0%	Bio15 (-0.91)
Temperature Annual Range (BIO5-BIO6) [Bio7]	6.4%	Bio6 (-0.95)
Mean Temperature of Coldest Quarter [Bio11]	6.4%	Bio6 (0.98)
*Temperature Seasonality (standard deviation *100) [Bio4]*	5.9%	Bio6 (-0.88)
Precipitation of Driest Month [Bio14]	5.7%	Bio15 (-0.93)
Precipitation of Wettest Month [Bio13]	3.8%	Bio2 (-0.88)
Precipitation of Driest Quarter [Bio17]	3.6%	Bio15 (-0.93)
Max. Temperature of Warmest Month [Bio5]	2.8%	Bio10 (0.82)
Annual Precipitation [Bio12]	2.7%	Bio15 (-0.93)
Precipitation of Wettest Quarter [Bio16]	1.5%	Bio15 (-0.87)
Mean Temperature of Wettest Quarter [Bio8]	-	-
Mean Temperature of Driest Quarter [Bio9]	-	-

As the selected Bioclim variables were included in the initial models for all taxa, and were removed during iterative modeling only if their percent contribution was equal to zero, these predictors were present in the greatest number of final models (Table 6). Bio6, the Minimum Temperature of the Coldest Month, appears in the final models for all 27 taxa, and has the greatest average percent contribution (27.4%) of any predictor across all taxa. Bio10 (Mean Temperature of the Warmest

Quarter) and Bio15 (Precipitation Seasonality) were the next most useful predictors across the full set of taxa, with percent contributions of 12.1% and 16.1%, respectively. Other predictors related to vegetation, soils, and topography were present in final models for fewer taxa, but in some cases were among the best predictors for those taxa.

Table 6. Predictor layers used in final models. Average percent contribution is the mean percent contribution for a predictor across final models for all taxa in which the predictor was included. Appendix 1 provides a full description of all predictor layers.

Predictor	**Number of Final Models Using Predictor**	**Average Percent Contribution**
bio6	27	27.4%
bio15	27	16.1%
bio18	25	6.0%
bio3	24	6.0%
bio2	23	6.2%
bio10	23	12.1%
nlcdcanopy	17	6.1%
percsand	13	3.3%
lfshrubcc	12	3.5%
lfevh	11	5.9%
dissect5	9	3.5%
percclay	8	3.9%
dissect10	7	0.9%
drainclass	7	9.7%
lfherbcc	7	2.2%
ned	7	22.0%
d2foredge	4	10.9%
soilph	4	12.6%
vrm10	4	5.6%
allwatdist	3	4.2%
lfforstcc	3	5.0%
slope	3	5.5%
water3200	3	2.9%
aprime45	2	1.8%
avoid12800	2	4.3%
ksat	2	8.1%
percsilt	2	2.7%
soilec	2	3.0%
vrm5	2	6.5%
aglands	1	0.6%
aprime90	1	2.5%
avoid1600	1	0.3%
avoid6400	1	0.3%
curve5	1	7.5%
d2wlsl	1	1.8%
hydgroup	1	7.4%
radld	1	6.1%
water1600	1	2.8%

Summary of Models

Full model reports showing the selected predictor layers, statistics on model fit and accuracy, and a map of the resulting probability surface for each taxon are provided in Appendix 2. A general summary of models across all taxa is provided here.

The initial model for Black Bears appeared to show fairly poor discrimination, likely due to differences in habitat usage between the western and eastern populations of the species. This is not surprising, given the dramatically different types of habitat available in these two portions of the state. We therefore split the data for the species into eastern and western populations and constructed separate models based on these separate point sets.

Across all taxa, models fit the training data (Figure 3a) and predicted test data (Figure 3b) well. As might be expected based on the input data quality (Figures 1 & 2), models for plants and birds were generally of higher quality than those for mammals and herptiles, though all models provided reasonably good accuracy on test data in cross-validation model runs. Black-tailed Prairie Dog appears as an outlier in the mammal group (AUC for training and cross-validation test data of 0.861 and 0.848, respectively). This is due to the fact that the training set for this species has a large number of points (2,355) covering a large geographic area. Since Maxent uses "fractional predicted area" (i.e., the percentage of the study area predicted present across varying thresholds) as a surrogate for the commission error typically used as the x-axis on a ROC plot, the model cannot achieve an AUC of 1, since an AUC of 1 would require classifying the entire study area as absent[47] while also correctly classifying all training presences. For this species, for example, the maximum AUC value possible would be 0.860, rather than 1, assuming the modeled prevalence matches the actual prevalence of the species on the landscape. This is true for AUC values reported for all taxa, though the effect is less pronounced for taxa with more narrow distributions. For the six taxa with a 20% test dataset, the AUC for test data closely matches the mean cross-validation AUC values, suggesting that the AUC values based on cross-validation are a reliable indicator of model quality in the absence of true test data.

Figure 3. Training AUC (a) and Mean Cross-Validation AUC across the modeling taxa, by taxonomic group. Training AUC is a measure of model fit, or specification, whereas the Mean Cross-Validation AUC provides an indication of model accuracy.

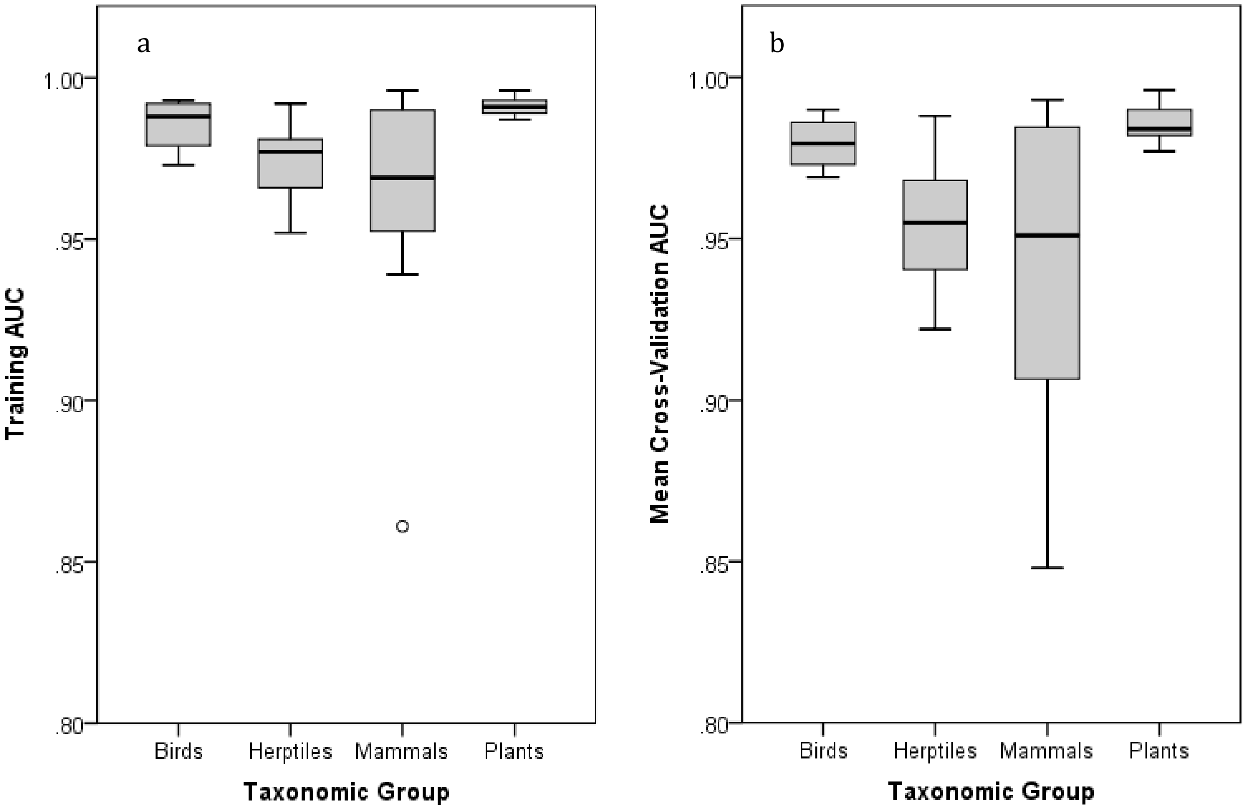
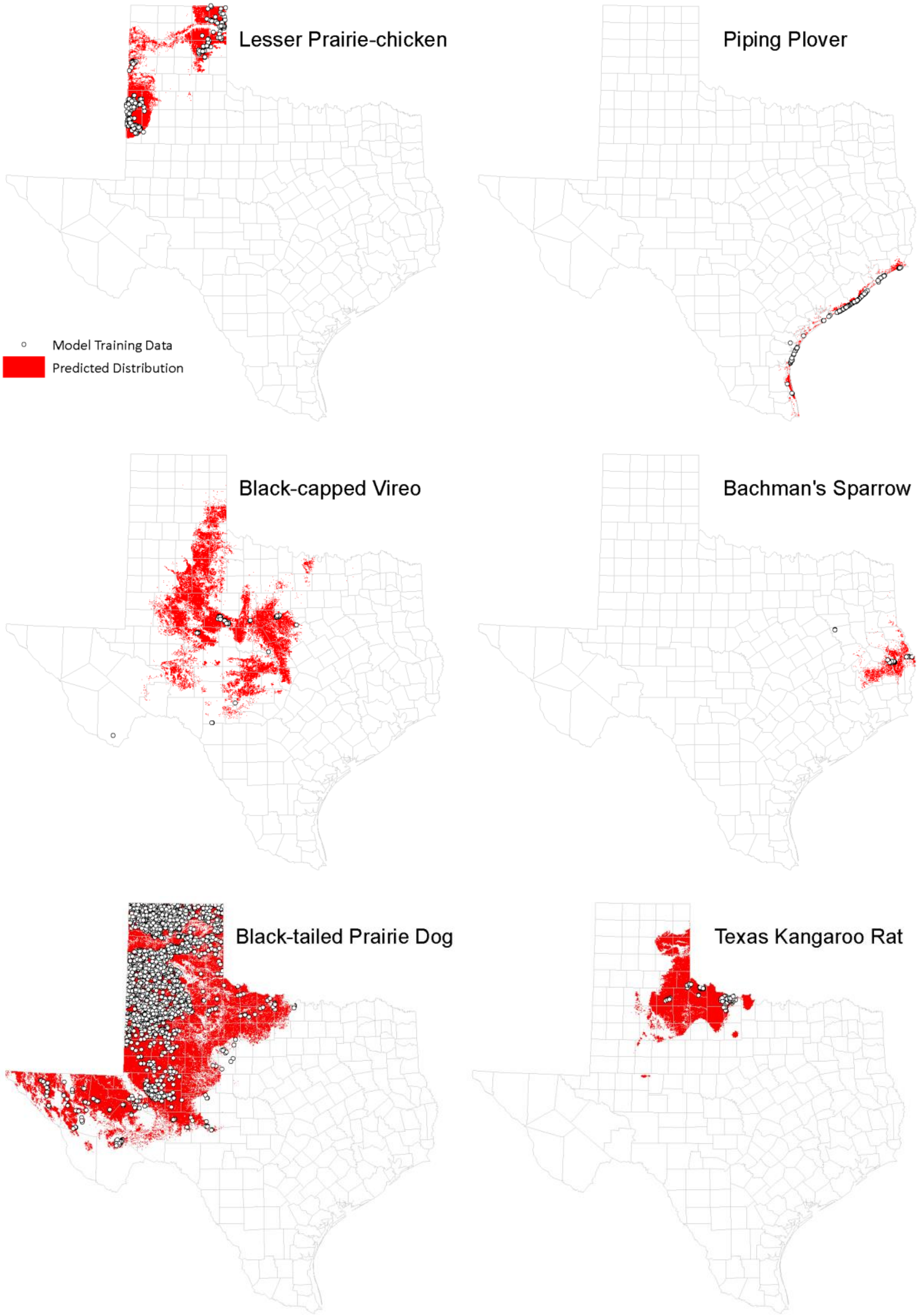
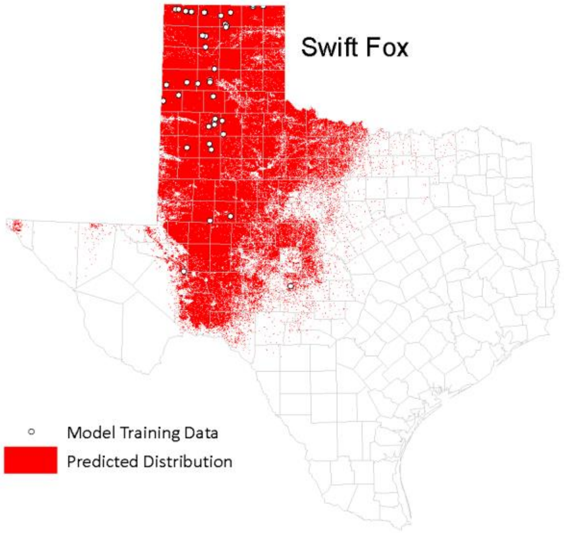


Table 5. Summary statistics for final distribution models, by taxon. For taxa with more than 100 points available for training, a random 20 percent of the points were held out for validation. Training AUC is the area under curve for a ROC plot based on the training data, and is a measure of how well the model fits the training data. Test AUC is based on a test dataset comprising the 20 percent of the data that was excluded from modeling, and is a measure of model accuracy. Mean Cross-Validation AUC gives the mean AUC value based on the test data across each of the 10 replicates in cross-validation, and was used as a measure of model accuracy when there were insufficient data to create a separate test dataset.

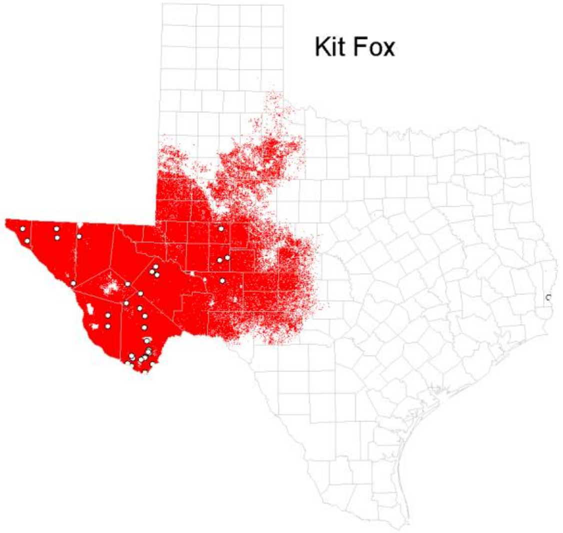
Taxon	**Number of Training Points**	**Number of Test Points**	**Training AUC**	**Test AUC**	**Mean Cross-Validation AUC**
Amphibians					
Sheep Frog	43	-	0.983	-	0.974
Birds					
Lesser Prairie-chicken	195	48	0.985	0.985	0.982
Piping Plover	84	20	0.993	0.990	0.990
Black-capped Vireo	63	-	0.991	-	0.977
Bachman's Sparrow	31	-	0.973	-	0.969
Mammals					
Black-tailed Prairie Dog	2,355	588	0.861	0.865	0.848
Texas Kangaroo Rat	98	24	0.994	0.995	0.993
Swift Fox	41	-	0.986	-	0.977
Kit Fox	35	-	0.939	-	0.887
Black Bear - Western	87	21	0.969	0.916	0.926
Black Bear - Eastern	40	-	0.966	-	0.951
Ocelot	21	-	0.996	-	0.992
Reptiles					
Texas Tortoise	60	-	0.952	-	0.929
Reticulate Collared Lizard	37	-	0.992	-	0.988
Spot-tailed Earless Lizard	104	26	0.974	0.954	0.955
Texas Horned Lizard	40	-	0.958	-	0.922
Texas Indigo Snake	50	-	0.977	-	0.952
Louisiana Pine Snake	29	-	0.979	-	0.962
Plants					
Texas Prairie Dawn	40	-	0.996	-	0.996
Threeflower Broomweed	28	-	0.989	-	0.982
Zapata Bladderpod	12	-	0.991	-	0.990
Bracted Twistflower	25	-	0.993	-	0.985
Tobusch Fishhook Cactus	82	-	0.990	-	0.984
Texabama Croton	16	-	0.995	-	0.982
Johnston's Frankenia	92	-	0.993	-	0.992
Chihuahua Balloon-vine	16	-	0.989	-	0.977
Navasota Ladies'-tresses	73	-	0.987	-	0.981

Figure 4. Thumbnails of binary models for each taxon.

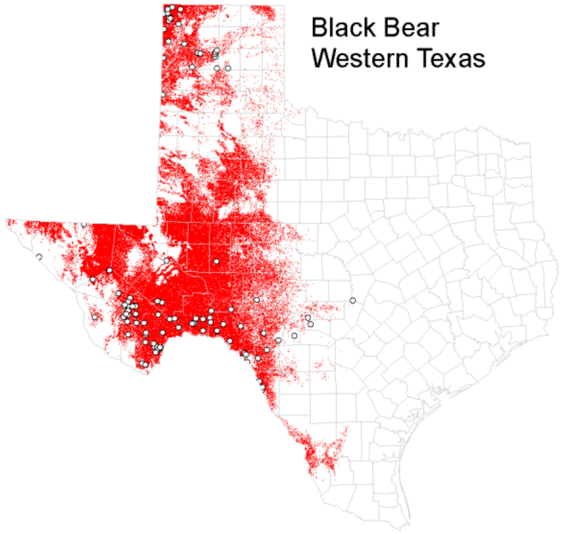




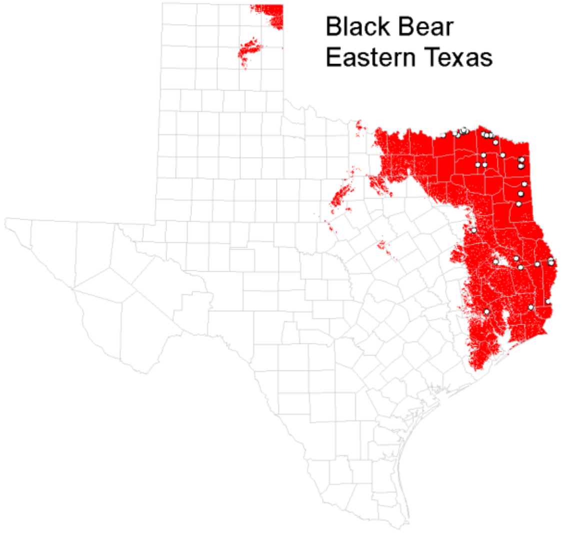
Swift Fox



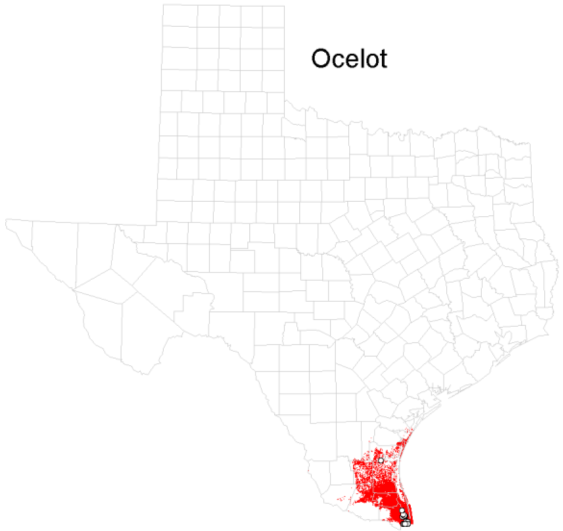
Kit Fox



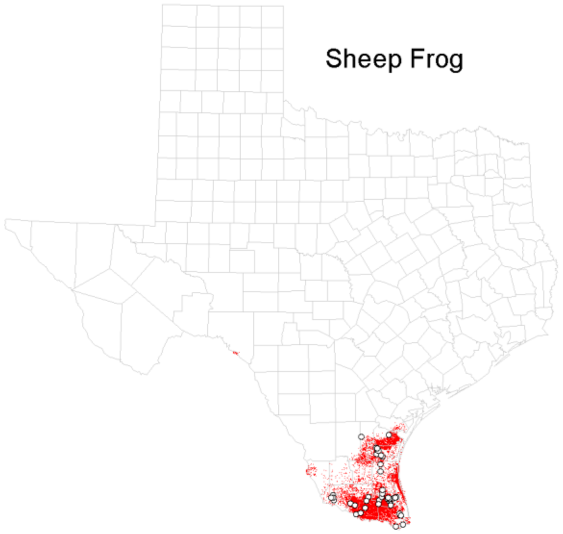
Black Bear
Western Texas



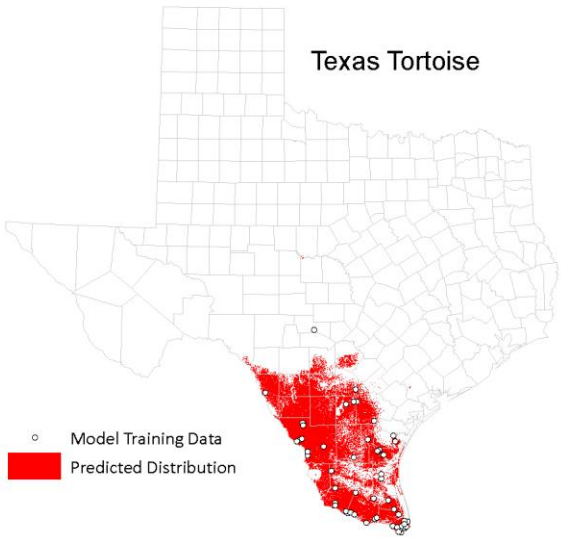
Black Bear
Eastern Texas

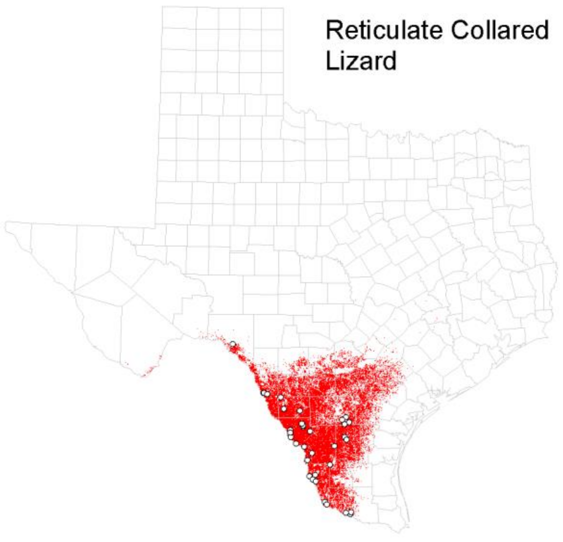


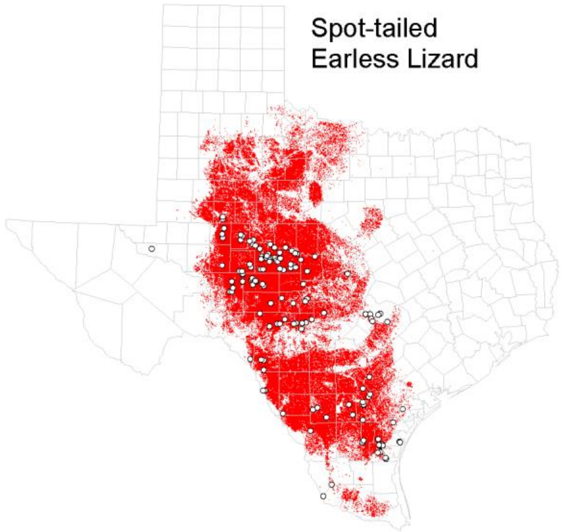
Ocelot

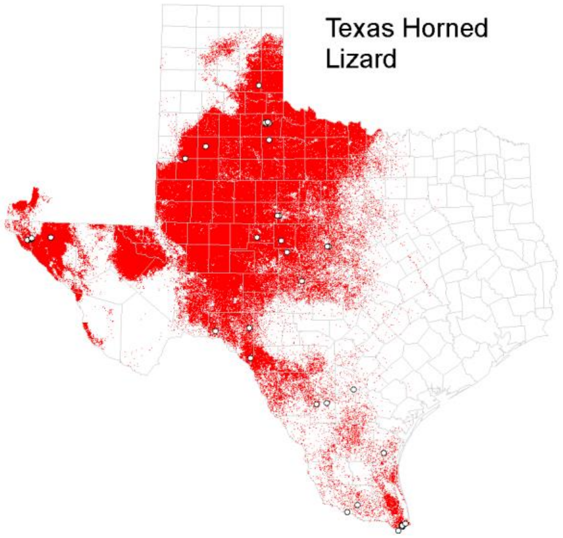


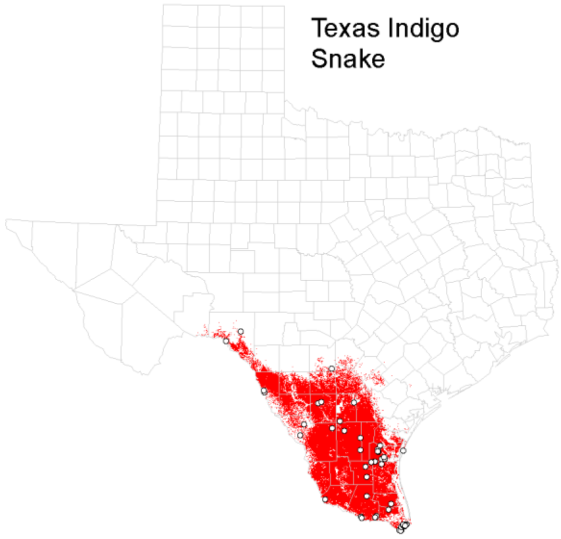
Sheep Frog

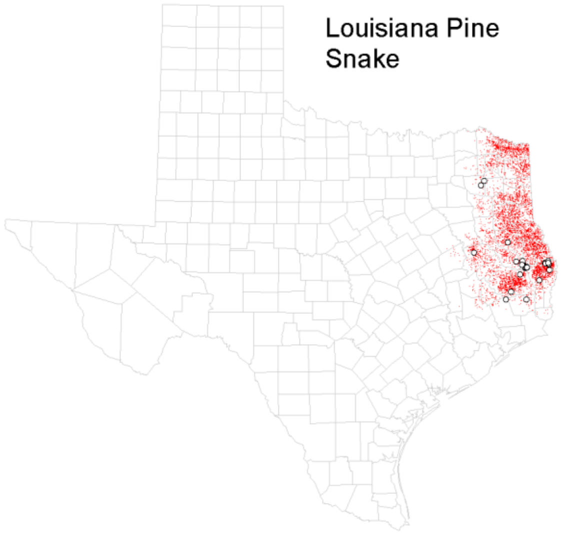

Texas Tortoise

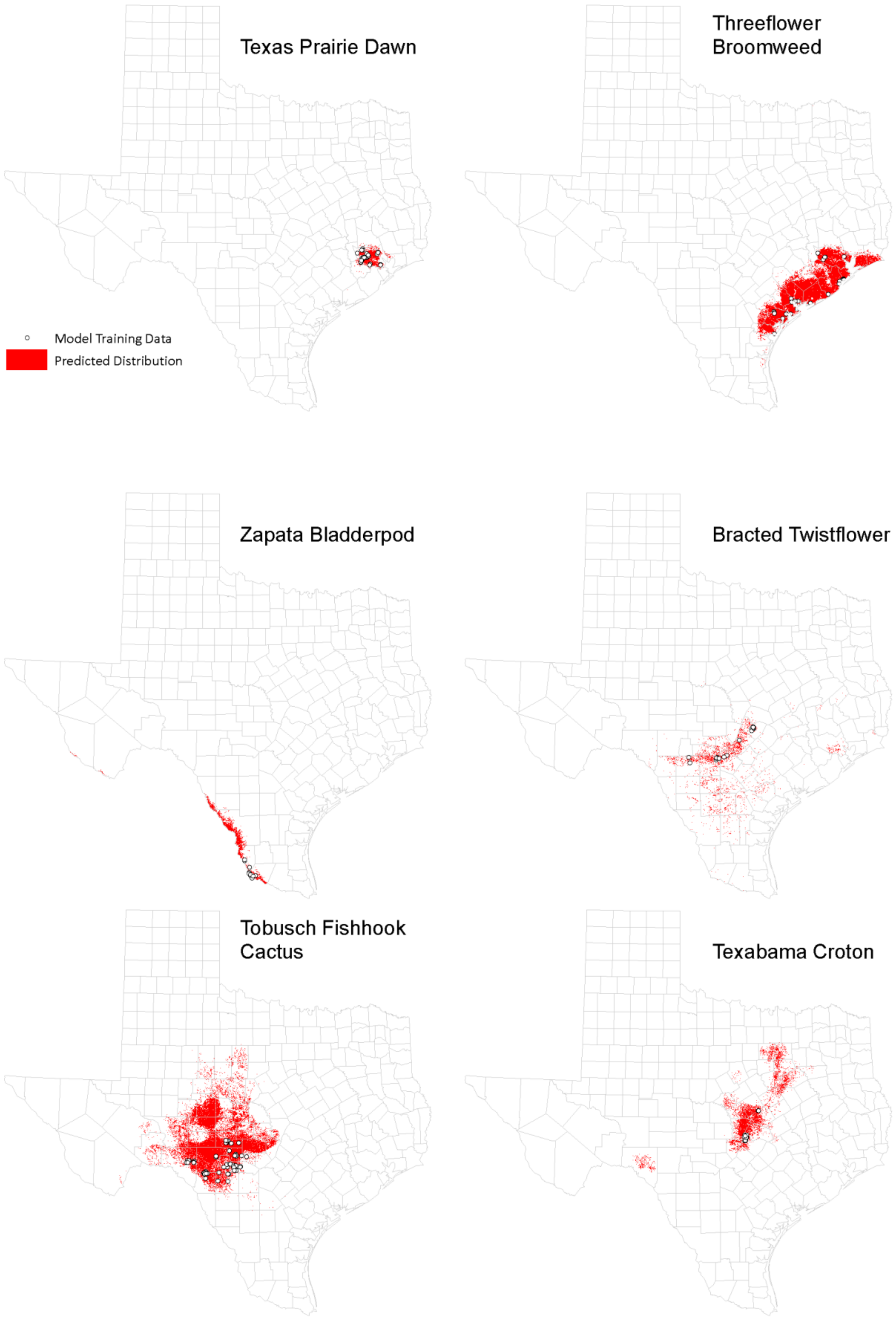

Reticulate Collared Lizard


Spot-tailed Earless Lizard


Texas Horned Lizard


Texas Indigo Snake


Louisiana Pine Snake



Texas Prairie Dawn

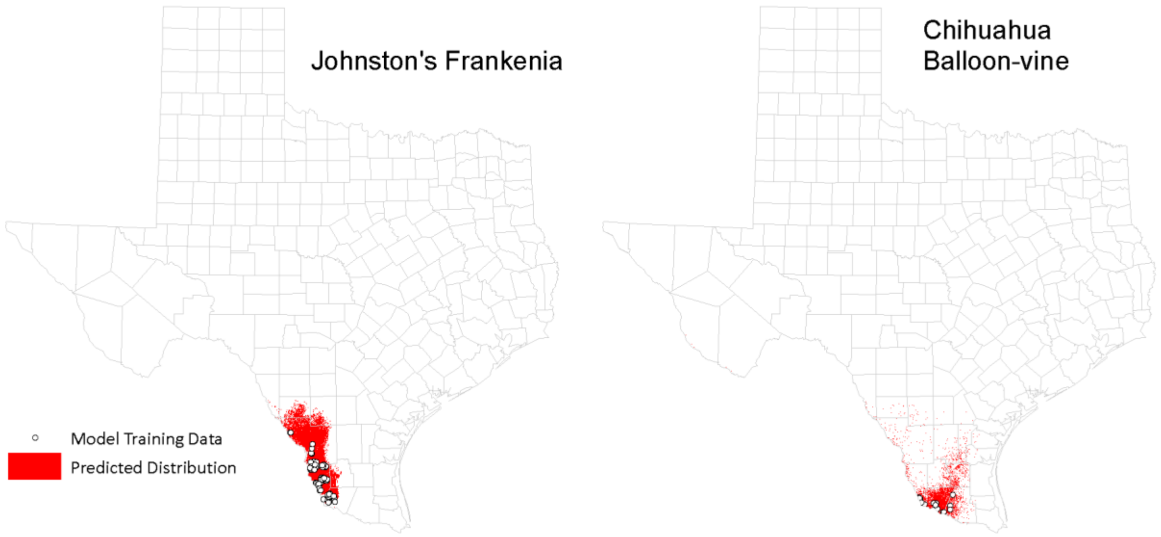
Threeflower Broomweed

Zapata Bladderpod

Bracted Twistflower

Tobusch Fishhook Cactus

Texabama Croton



DISCUSSION

USAGE AND LIMITATIONS OF DISTRIBUTION MODELING

Statistician George Box is perhaps best known for his famous axiom: "…essentially, all models are wrong, but some are useful."[48] His point was that models are always a simplification of some more complex system, and that in simplifying the system, important details are necessarily left out, leading to an imperfect model.

Box's adage certainly applies to all distribution models. In distribution modeling, there is uncertainty and error inherent in occurrence data, predictor layers, and the underlying mechanistic processes that shape actual distribution. Observation records can be poorly located, misidentified, or unrepresentative of a taxon's distribution, or they can derive from biased sampling efforts and as a consequence suggest an unrealistic picture of the taxon's distribution. Predictor layers can exhibit error in both position and value, leading to faulty conclusions about the relationship between a predictor and a taxon's distribution. Finally, the underlying mechanisms that influence a taxon's distribution can be inordinately complex or otherwise impossible to represent accurately with a simplified model. As Box's axiom suggests, though, distribution models can be both wrong *and* still very useful. The challenge to the modeler is twofold: 1) to interpret the model output correctly; and 2) to convey this interpretation to end users unfamiliar with modeling to assist them in using models in a way that acknowledges uncertainty and error.

Model Interpretation and Usage

Although the output GIS data from Maxent models are commonly thought of as logistic probability of occurrence by a taxon, the actual interpretation is more nuanced[12]. When only presence data are available and incomplete information exists on sampling effort, there is no way to determine the prevalence of a taxon on the landscape (i.e., the proportion of the modeling area occupied by a taxon)[10]. Thus, there is no means of determining what the true probability of occurrence is for a taxon in any given location based on a presence-only model. Rather, Maxent provides what might be considered a relative suitability[47], and individual values of probability at any given cell in a Maxent raster output are only meaningful in comparison to other cell values in the same model output. Higher output values do indicate a higher likelihood of occurrence, but output values from two different models cannot be directly compared (i.e., a value of 0.5 in a model for one taxon may not mean the same thing as a value of 0.5 in another taxon's model).

Caution must also be exercised when evaluating partial plots (graphs showing likelihood of occurrence as a function of each variable, holding all other variables constant). While indirect predictor layers such as elevation might contribute highly to the accuracy of a model, it would not be correct in most cases to state, for example, that elevation has a specific effect on distribution. Rather, elevation most likely influences temperature, precipitation, or other more direct gradients than more directly limit a taxon's distribution. Biological understanding is thus important in interpreting partial plots – particularly those for more indirect predictors.[49]

Distribution models such as those produced in this project can help identify areas of high biodiversity or important habitat or potential habitat for priority taxon, at a coarse scale. Planners can use such maps to identify suitable locations for ecological reserves, or to help them determine areas that are more suitable for development with minimal adverse impacts to biodiversity or to a particular taxon[5].

Distribution models can also be used to guide more efficient field surveys. By selecting the areas predicted by a model to be most suitable, researchers can hone in on the most likely locations to find a particular taxon to make the most of limited field project budgets. Moreover, by evaluating model output in the context of known occurrence locations, researchers can focus on areas a model deems suitable but that currently have no known records for the taxon, potentially expanding its known distribution.

However, models should not be used in place of site-level, clearance surveys for taxa with special management designations, as the predictor layers used to create distribution models are generally too coarse to make an accurate prediction at this scale. For project planning at a site-level, models can provide only an indication of whether the taxon is predicted to be "in the neighborhood," in which case on the ground surveys are likely warranted.

These latter two uses may require expressing a distribution model differently by applying different thresholding or symbology in mapping the model output. A biologist interested in trapping an organism to collect DNA samples, for example, would most benefit by limiting their sampling to only the areas predicted to be most suitable for their target taxon. Conversely, a biologist tasked with evaluating the potential impact of development for a taxon may want to err on the side of caution, by considering even areas of low predicted likelihood of occurrence to be potentially occupied and warranting field surveys. Lastly, researchers interested in the total area occupied by a taxon will be best served by selecting a threshold metric, like the "Maximum training sensitivity plus specificity" metric used here, that seeks to minimize total commission and omission error, resulting in a relatively unbiased estimate of area of occupancy.

Occurrence Data Limitations

While researchers have used a variety of modeling approaches to produce useful models with as few as ten training presences, model performance generally improves with increasing sample size[50,51], possibly leveling off at 50 to 100 training presences[11,52,53], particularly with Maxent [54]. While most of the modeled taxa had approximately 50 or more total points in the dataset provided by TXNDD, fifteen of the 27 modeling taxa had fewer than 50 points after the filtering process. Most of the points removed during filtering were removed due to high levels of clustering at small (<800 m) distances. Substantially better models may result if additional, independent observations were made in expanded portions of the taxa's distributions and added to the modeling sets for these taxa.

For some taxa, the high degree of clustering in their observation records may reflect areas of higher suitability rather than biased survey effort. In such cases, spatial filtering may not be required, allowing more of the observation records to be used for training data. This decision must be made based on an understanding of the spatial patterning and level of bias in the sampling effort that lead to the available training points.

No negative (i.e., absence) records were available for the modeled taxa. While modeling based on presence-only data is common, negative data can be used to provide models that discriminate more sharply between areas of predicted presence and absence[10]. Additionally, absence data allow the modeler a broader suite of potential modeling algorithms, including standard statistical methods such as logistic regression[42] (when sample sizes are large) and machine learning methods like Random Forest[55].

Absence data may be difficult to generate, as it requires relatively detailed knowledge survey effort and design, and the amount of survey effort required to confidently assign a location as an absence varies by taxa[56]. For example, with plants, one might only need to visit a site a single time during the flowering period for the taxon to conclude that it is absent. By contrast, for highly vagile taxa such as birds and large mammals that may occur in low numbers across large areas, many repeated surveys might be needed before a biologist would be willing to conclude that the taxon is absent. Nevertheless, given the benefits of absence data for distribution modeling, it warrants further consideration.

For rare plants, absence data may be reasonably generated by selecting observations recorded for other plant taxa during the flowering period for the target taxa, under the assumption that if the target taxon had been present and flowering, it would have been noted by the surveyor. Structured, repeated survey protocols such as the Breeding Bird Survey[57] and Christmas Bird Count[58] may provide enough sampling over the course of many years at specific sites that a taxon can be considered absent if it has never been recorded at a site. However, if an organism is particularly difficult to detect, as with many grassland birds outside of the breeding season, even multiple years of repeated surveys at a site might not constitute a reliable absence location.

Predictor Data Limitations

For many plants and most fossorial animals, soil characteristics are extremely important in shaping distribution. Unfortunately, though SSURGO[59] provides detailed soil survey data for Texas, it can be impractical to create derivative layers from these data given current limitations on file sizes and software capabilities. Many useful soil attributes, including measures of chemistry, texture, moisture, and depth, can be generated using the NRCS' Soil Data Viewer, but at present it is not possible to process the entire SSURGO dataset for Texas with this tool. By using the Gridded SSURGO (gSSURGO) dataset[39] with custom scripts provided by NRCS staff, we were able to derive many useful layers (e.g., Percent Clay/Silt/Sand, soil pH, soil drainage class), but were unable to create other potentially useful layers. For example, we have used a "depth to shallowest restrictive layer" in past modeling work to improve predictions for fossorial animals, as soil depth may limit distribution for some taxa, but we were unable to generate this layer using the available scripts in conjunction with the gSSURGO data. As with land cover data, where a sufficient conceptual understanding exists, soil series data can also be assigned taxa-specific suitability scores to generate binary or continuous indices that could improve predictions for selected taxa.

For at least three of the taxa modeled in this project (Black-capped Vireo[25], Bachman's Sparrow[60], and Louisiana Pine Snake[61]), disturbance history and seral stage an important factor influencing distribution. Fire history information has been successfully used to improve distribution models for birds that require early successional vegetation[62], but a substantial amount of effort may be required to compile and maintain the necessary disturbance layers. While some GIS data layers exist for disturbances like fire[63] and disease[64], these disturbances are highly variable in both space and time, so any models using disturbance layers as predictors would need to be updated frequently as new data become available. One practical alternative is to maintain a more general distribution model for such taxa, and to use ancillary data on disturbance to guide field work or assist with assessing and mapping habitat quality.

Suggestions for Future Work

As with any analysis or modeling project, collecting additional training data can greatly improve distribution models. Clearly, additional observation records for modeled taxa – particularly records some distance away from existing records – will provide additional information for modeling. Similarly, absence data for the modeled taxa could greatly improve models in two ways: 1) presence-absence models can draw a sharper distinction between occupied and unoccupied habitat; and 2) the availability of absence data in addition to presence data allows the use of many other modeling algorithms, including statistical methods like regression and machine learning methods like Random Forest. Inferences drawn from presence-absence models are generally more straightforward than those drawn from presence-only models. Absence data can be collected directly, when a taxon is surveyed for but not found, or it can be created retroactively based on prior survey work that found other taxa, but not the target taxon. We suggest that TXNDD consider collecting or creating absence data for all tracked taxa whenever possible.

We did not use data from online citizen science databases[16,17], as we felt these records were more useful as an independent dataset to be used in model validation than as additional training data. Currently, Herps of Texas does not have enough records for a thorough validation of most of the reptile and amphibian taxa we modeled, and obtaining and processing eBird data was not practical within the time constraints for this project. These datasets should be investigated as time permits for their potential value in validating the distribution models created here.

Likewise, collecting data for non-tracked taxa also can help improve model quality. When data are originally collected during field surveys, researchers will often record locations for a large number of taxa in addition to their target species. These additional records provide information about survey effort that can be used to gauge and factor out survey bias in records for the target taxa. Though many heritage programs do not collect data for taxa that they do not track, there typically is relatively little cost to integrating these records in the database, as they are often part of the datasets provided by researchers. We suggest that TXNDD consider storing observation data for any taxa contained in a dataset provided to TXNDD, to allow a more robust background dataset for future modeling work.

As with occurrence data, collection or generation of newer and better predictor datasets should continue to be a priority for modeling work. New data layers based on remotely sensed data are made available on a regular basis as satellite imagery becomes more ubiquitous. If time permits, developing taxa-specific layers by rescaling, scoring, or combining other datasets can greatly improve models. We were unable to generate numerous soil derivative layers (Appendix 1) we believed might be helpful in modeling due to time constraints and difficulties in compiling and processing statewide soil datasets. In comments regarding selected plant taxa, TXNDD's botanist suggested that it would be useful to have layers representing deciduous versus evergreen canopy cover or soil parent material, for example, and that some plants appear to have particular associations with soil series or geology formations of a particular age. These types of layers likely can be generated from existing data, but will require direct collaboration between biologists and modelers to ensure that the appropriate information is being represented.

Land cover layers were not directly used in constructing the models for these 27 taxa. Land cover layers no doubt contain useful information, but are problematic for inductive modeling, as variables with many of categories tend to be preferentially selected by modeling algorithms even when the relationship with the categories is spurious.[65] If a conceptual understanding of a taxon's distribution suggests that vegetative community strongly influences distribution, one could assign taxa-specific, numerical suitability ratings to each land cover type to produce a continuous index from these categorical data. While somewhat subjective in their definition, indices such as these have proven invaluable in previous modeling efforts.[9] Alternatively, land cover layers could be used to produce a standard deductive model that predicts distribution based on a binary suitability value (suitable/not suitable) for each land cover type. This deductive model could then be combined with an inductive model for the same taxon using a simple multiplicative raster overlay to eliminate areas that are not within suitable land cover types. This approach has also been successfully implemented in prior modeling work.[8]

Range maps that show occupancy of a taxon within predefined units of space (e.g., watersheds, counties) can be integrated in distribution modeling to improve predictions. Distribution models generally select more general, broad-scale predictors as study area size increases, and, conversely, tend to identify more direct, finer-scale predictors as the modeling area decreases. Thus, by limiting the modeling area to a taxon's range, predictions of distribution generally become sharper and more detailed. While watershed-based range maps were available for some of the modeling taxa, it would be helpful to have range maps across all taxa so that they could be integrated in a consistent modeling process across taxa.

A critical part of this project was direct consultation between experienced WYNDD modelers and TXNDD personnel anticipated to continue the modeling started here. We emphasize that such direct contact and training is a vital part of distribution modeling, and is at least as valuable as providing technical examples and written documentation. As discussed above, there are portions of the modeling process (e.g., occurrence data filtering, predictor variable selection) that involve some degree of 'art,' or subjective decision-making, as the variability in input data, important environmental gradients, study areas, and species biology have thus far precluded the development of hard and fast rules for all distribution modeling. In many cases, regionally and taxonomically-specific modeling experience provides the best available guidance related to modeling decisions.

REFERENCES

1 Texas Parks and Wildlife Department. Texas Conservation Action Plan 2012 - 2016: Overview. (2012).

2 Texas Parks and Wildlife Department. Texas Comprehensive Wildlife Conservation Strategy 2005 - 2010. . (2005).

3 NatureServe. *An online encyclopedia of life [web application], s.v. "Texas". Version 7.1.*, <http://www.natureserve.org/explorer/> (2009).

4 Merrill, E. H. *et al.* The Wyoming GAP Analysis Project: Final Report. 250pp.- (University of Wyoming, Laramie, 1996).

5 Franklin, J. & Miller, J. A. *Mapping species distributions: spatial inference and prediction*. Vol. 338 (Cambridge University Press Cambridge, 2009).

6 Maxent software for species distribution modeling (2005).

7 Hirzel, A., Hausser, J., Chessel, D. & Perrin, N. Ecological-niche factor analysis: how to compute habitat-suitability maps without absence data? *Ecology* **83**, 2027-2036 (2002).

8 Beauvais, G. P., Andersen, M. D. & Keinath, D. A. Range, distribution, and habitat of terrestrial vertebrates in the 5-state Northwest ReGAP region. Report prepared for the USDI Geological Survey - Gap Analysis Program (Moscow, Idaho). (University of Wyoming, Laramie, Wyoming, 2012).

9 Keinath, D., Andersen, M. & Beauvais, G. Range and modeled distribution of Wyoming's species of greatest conservation need. *Report prepared by the Wyoming Natural Diversity Database, Laramie Wyoming for the Wyoming Game and Fish Department, Cheyenne, Wyoming and the US Geological Survey, Fort Collins, Colorado* (2010).

10 Yackulic, C. B. *et al.* Presence-only modelling using MAXENT: when can we trust the inferences? *Methods in Ecology and Evolution* (2012).

11 Elith, J. *et al.* Novel methods improve prediction of species' distributions from occurrence data. *Ecography* **29**, 129-151 (2006).

12 Phillips, S. J. & Dudík, M. Modeling of species distributions with Maxent: new extensions and a comprehensive evaluation. *Ecography* **31**, 161-175 (2008).

13 Pearce, J. L. & Boyce, M. S. Modelling distribution and abundance with presence-only data. *Journal of Applied Ecology* **43**, 405-412 (2006).

14 Phillips, S. J. *et al.* Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. *Ecological Applications* **19**, 181-197 (2009).

15 Venier, L. A. & Pearce, J. L. Boreal forest landbirds in relation to forest composition, structure, and landscape: implications for forest management. *Canadian Journal of Forest Research-Revue Canadienne De Recherche Forestiere* **37**, doi:10.1139/x07-025 (2007).

16 *Herps of Texas*, <http://www.herpsoftexas.org/> (2013).

17 Sullivan, B. L. *et al.* eBird: A citizen-based bird observation network in the biological sciences. *Biological Conservation* **142**, 2282-2292 (2009).

18 Johnson, D. H. The comparison of usage and availability measurements for evaluating resource preference. *Ecology* **61**, 65-71 (1980).

19 Meyer, C. B. & Thuiller, W. Accuracy of resource selection functions across spatial scales. *Diversity and Distributions* **12**, 288-297 (2006).

20 Poole, J. M., Carr, W. R., Price, D. M. & Singhurst, J. R. *Rare plants of Texas: A field guide*. (Texas A & M University Press, 2007).

21 Banfield, A. (University of Toronto Press, Toronto, 1974).

22 Emmons, L. H. Neotropical rainforest mammals. A field guide. *Neotropical rainforest mammals. A field guide.*, i-xiv, 1-281 (1990).

23 Ernst, C. H., Barbour, R. W. & Lovich, J. E. Turtles of the United States and Canada. *Turtles of the United States and Canada.*, i-xxxviii, 1-578 (1994).

24 Giesen, K. M. Lesser prairie-chicken (Tympanuchus pallidicinctus). *Birds of North America* **364**, 1-19 (1998).

25 Grzybowski, J. A., Tazik, D. J. & Schnell, G. D. Regional analysis of black-capped vireo breeding habitats. *Condor*, 512-544 (1994).

26 Moss, S. P. & Mehlhop-Cifelli, P. Status of the Texas Kangaroo Rat, Dipodomys elator (Heteromyidae), in Oklahoma. *The Southwestern Naturalist* **35**, 356-358 (1990).

27 Pruss, S. D. Selection of natal dens by the swift fox (Vulpes velox) on the Canadian prairies. *Canadian Journal of Zoology* **77**, 646-652 (1999).

28 Stasey, W. C., Goetze, J. R., Sudman, P. D. & Nelson, A. D. Differential use of grazed and ungrazed plots by Dipodomys elator (Mammalia: heteromyidae) in north central texas. *Texas Journal of Science* **22** (2011).

29 Tennant, A. The snakes of Texas. *The snakes of Texas.*, 1-560 (1984).

30 Ratzlaff, A. Endangered and threatened wildlife and plants: determination of the black-capped vireo to be an endangered species. *Federal Register* **52**, 37420-37423 (1987).

31 Werler, J. E. & Dixon, J. R. Texas snakes: identification, distribution and natural history. *Texas snakes: identification, distribution and natural history.*, i-xv, 1-437 (2000).

32 Hijmans, R. J., Cameron, S. E., Parra, J. L., Jones, P. G. & Jarvis, A. Very high resolution interpolated climate surfaces for global land areas. *International journal of climatology* **25**, 1965-1978 (2005).

33 Gesch, D., Oimoen, M., Greenlee, S., Nelson, C., Steuck, M., and Tyler, D. The National Elevation Dataset: Photogrammetric Engineering and Remote Sensing. **68**, 5-11 (2002).

34 US Geological Survey (USGS) Gap Analysis Program (GAP). (USGS Gap Analysis Program Ancillary Data - Hydrography, http://gapanalysis.usgs.gov/data/species-data/, 2011).

35 US Geological Survey (USGS) Gap Analysis Program (GAP). (USGS Gap Analysis Program Ancillary Data - Human Impact Avoidance, http://gapanalysis.usgs.gov/data/species-data/, 2011).

36 U.S. Department of Interior - Geological Survey. (LANDFIRE: LANDFIRE Existing Vegetation Height layer, Available: http://landfire.cr.usgs.gov/viewer/, 2013).

37 U.S. Department of Interior - Geological Survey. (LANDFIRE: LANDFIRE Existing Vegetation Cover layer, Available: http://landfire.cr.usgs.gov/viewer/, 2013).

38 US Geological Survey (USGS) Gap Analysis Program (GAP). (USGS Gap Analysis Program Ancillary Data - Forest and Ecotone Habitats, http://gapanalysis.usgs.gov/data/species-data/, 2011).

39 Soil Survey Staff - United States Department of Agriculture - Natural Resources Conservation Service. (Gridded Soil Survey Geographic (gSSURGO) Database for Texas, Available online at http://datagateway.nrcs.usda.gov/, Accessed June 25, 2013).

40 Geospatial Modelling Environment (Version 0.7.2.1) (Hawthorne L. Beyer, URL: http://www.spatialecology.com/gme, 2012).

41 Phillips, S. (http://www.cs.princeton.edu/~schapire/maxent/tutorial/tutorial.doc, No Date.).

42 Hosmer, D. W. & Lemeshow, S. *Applied Logistic Regression*. (John Wiley & Sons, 1989).

43 R: A Language and Environment for Statistical Computing (R Foundation for Statistical Computing. Available at: http://www.R-project.org, Vienna, Austria, 2013).

44 Menard, S. *Applied logistic regression analysis, Second Edition*. (Sage, 2002).

45 Fielding, A. H. & Bell, J. F. A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environmental Conservation* **24**, 38-49 (1997).

46 Kampstra, P. Beanplot: A boxplot alternative for visual comparison of distributions. *Journal of Statistical Software* **28** (2008).

47 Phillips, S. J., Anderson, R. P. & Schapire, R. E. Maximum entropy modeling of species geographic distributions. *Ecological Modelling* **190**, 231-259 (2006).

48 Box, G. E. & Draper, N. R. *Empirical model-building and response surfaces*. (John Wiley & Sons, 1987).

49 Austin, M., Belbin, L., Meyers, J., Doherty, M. & Luoto, M. Evaluation of statistical models used for predicting plant species distributions: role of artificial data and theory. *Ecological Modelling* **199**, 197-216 (2006).

50 Hirzel, A. & Guisan, A. Which is the optimal sampling strategy for habitat suitability modelling. *Ecological Modelling* **157**, 331-341 (2002).

51 Cumming, G. S. Using between-model comparisons to fine-tune linear models of species ranges. *Journal of Biogeography* **27**, 441-455 (2000).

52 Stockwell, D. R. & Peterson, A. T. Effects of sample size on accuracy of species distribution models. *Ecological modelling* **148**, 1-13 (2002).

53 Kadmon, R., Farber, O. & Danin, A. A systematic analysis of factors affecting the performance of climatic envelope models. *Ecological Applications* **13**, 853-867 (2003).

54 Hernandez, P. A., Graham, C. H., Master, L. L. & and Albert, D. L. The effects of sample size and species characteristics on performance of different species distribution modeling methods. *Ecography* **29**, 773-785 (2006).

55 Breiman, L. Random forests. *Machine learning* **45**, 5-32 (2001).

56 Ward, G., Hastie, T., Barry, S., Elith, J. & Leathwick, J. R. Presence-Only Data and the EM Algorithm. *Biometrics* **65**, 554-563 (2009).

57 Sauer, J. R. *et al.* The North American Breeding Bird Survey, Results and Analysis 1966 - 2009. Version 3.23.2011. (USGS Patuxent Wildlife Research Center, Laurel, MD, 2011).

58 National Audubon Society. (National Audubon Society, www.christmasbirdcount.org, 2012).

59 Soil Survey Staff - United States Department of Agriculture - Natural Resources Conservation Service. (Soil Survey Geographic (SSURGO) &#x9;&#x9;Database for Texas, Available online at http://datagateway.nrcs.usda.gov/, 2013).

60 Tucker Jr, J. W., Robinson, W. D. & Grand, J. B. Influence of fire on Bachman's Sparrow, an endemic North American songbird. *Journal of Wildlife Management* **68**, 1114-1123 (2004).

61 Rudolph, D. C. & Burgdorf, S. J. Timber rattlesnakes and Louisiana pine snakes of the west Gulf Coastal Plain: hypotheses of decline. *Texas Journal of Science* **49**, l (1997).

62 Vallecillo, S., Brotons, L. & Thuiller, W. Dangers of predicting bird species distributions in response to land-cover changes. *Ecological Applications* **19**, 538-549 (2009).

63 Eidenshink, J. *et al.* A project for monitoring trends in burn severity. *The Journal of the Association for Fire Ecology* **3**, 3 (2007).

64 Bennett, D. D. & Tkacz, B. M. Forest health monitoring in the United States: a program overview. *Australian Forestry* **71**, 223-228 (2008).

65 Hastie, T., Tibshirani, R. & Friedman, J. Elements of Statistical Learning (Springer, 2009).