PREDICTIVE DISTRIBUTION MODELING OF BLM THREATENED, ENDANGERED, AND SENSITIVE PLANT SPECIES IN WYOMING

Prepared by

The Wyoming Natural Diversity Database

University of Wyoming

Laramie, Wyoming

November 18, 2016

Cooperative Agreement No. L12AC20036, Supplement 3

Suggested citation for this report:. Andersen, M.D., B. Heidel, and G.P. Beauvais. 2016. Predictive Distribution Modeling of BLM Threatened, Endangered, and Sensitive Plant Species in Wyoming. Report prepared by the Wyoming Natural Diversity Database, Laramie, Wyoming, for the Wyoming BLM.

CONTENTS

ABSTRACT

In Wyoming, the Bureau of Land Management (BLM) is responsible for managing Threatened, Endangered, and sensitive (TES) plant species. A critical aspect of managing TES plants is having a good understanding of their distribution on the landscape. The Wyoming Natural Diversity Database, Rocky Mountain Herbarium, and others have conducted botanical surveys and inventories for years, resulting in a large number of presence records for many rare plants in the state. However, some areas of the state have not been as thoroughly surveyed, leading to gaps in the available presence data for species. Predictive distribution modeling is becoming increasingly common as a method for filling in these types of survey gaps, providing a more complete picture of a species' potential distribution. We produced predictive distribution models for 47 TES plant species in Wyoming. The resulting models are not a substitute for field surveys, but they can be used to inform field surveys or to provide a "first-pass" filter for evaluating potential management actions.

INTRODUCTION

In Wyoming, the Bureau of Land Management (BLM) manages Special Status Species that include those plant species designated by BLM as sensitive[1], as well as Threatened and Endangered plant species recognized under the Endangered Species Act. One key to effective resource management is understanding the geographic distribution of the resource in question. The Wyoming Natural Diversity Database (WYNDD; University of Wyoming) maintains plant distribution data representing a synthesis of known distribution of Wyoming plant species of concern[2], including all TES species and other native species whose viability is in question in the state. The WYNDD database integrates available documentation including the collection data of the Rocky Mountain Herbarium (RM) and the robust plant survey data of WYNDD botanists, and other work by botanists statewide. These data have all been digitized to reflect any accompanying information available on location precision and geographic extent. They have been integrated to reflect spatial discreteness or overlap and accrual of information over time, such that records separated by some distance are inferred to represent separate populations. We refer to a spatially discrete record as an element occurrence records, i.e., a working approximations of a population. Throughout this report, the word "record" with no other qualifier refers to population-level data as processed and stored in the WYNDD database.

Nearly any set of records will provide an incomplete picture of a species' distribution, since they do not indicate whether a species may be present in unsurveyed areas. A key problem is that negative data (i.e., locations where a species was surveyed for but not found) are rarely archived in a database, unlike presence records. Thus, it is often unclear whether the blank areas on maps showing presence records represent unoccupied areas, or are simply areas that were never surveyed for the species. Likewise, it is sometimes unclear whether clusters of records reflect areas of high suitability for a species, or are merely the product of uneven or spatially-biased sampling effort resulting from constraints such as study areas limits or private land accessibility.

Distribution modeling has become a common method for assessing potential distribution in these blank areas with a prediction of suitability for occupancy by a species[3-7]. Deductive distribution models use expert knowledge to create a rule set that predicts suitability for occupancy based on important environmental characteristics of the landscape (e.g., land cover type). 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[8].

WYNDD previously modeled the potential distribution of TES plants species for BLM Wyoming, using a combination of deductive and inductive methods[3]. Inductive models were generated using Classification and Regression Trees (CART)[9] to model available presence and inferred absence locations based on underlying characteristics of topography, climate, substrate, and land cover. Deductive models were generated for species with insufficient presence data (i.e., fewer than eight presence locations available), and were created by the "range/intersection" method. The range/intersection method identified ranges or values for continuous or categorical environmental characteristics, respectively, based on observed values at presence locations, and then mapped the intersection of these "ranges" of suitable values as areas of predicted presence.

Although these models have proven useful tools for informing land management, planning, and field work in the state of Wyoming, the presence data were based on single centroids to represent

each record, and the negative data were based on RM thesis collection sites and a single centroid for the collection area of one or more sections. Since that time, all "centroid point" records have been converted either to polygons or to points buffered to indicate the degree of mapping uncertainty. In addition, many additional status reports entailing field surveys by WYNDD and additional floristic theses completed at RM have resulted in major expansions of distribution data for some species, as well as additions and deletions to the species of concern list.

Further, extensive research and application of species distribution modeling during the past 10-15 years have led to new insights, methods, and datasets that can be used to improve models. A number of new statistical and machine learning procedures have been developed for, or newly applied to, the problem of distribution modeling, including a number of methods that use randomization[10], regularization[11], cross-validation[5], and other techniques to improve models for species with limited presence data. Such methods were not computationally practical when the 2003 modeling work was underway, but are now possible due to faster computers and more efficient algorithms. Likewise, the library of available spatial layers representing environmental attributes for Wyoming continues to grow, allowing modelers to more closely match the biologically-relevant factors influencing species distributions. Finally, WYNDD and many other researchers have gained experience in distribution modeling that translates into new approaches that can help address previous shortcomings in modeling theory and data.

This project used inductive modeling with a commonly applied algorithm to generate predictive distribution models for 47 TES plant species in Wyoming. While distribution models, like all models, are subject to error, they offer a useful representation of a species' potential distribution that complements existing presence records. 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. All modeling input and output data, summary statistics, and methods are presented in this report, and are available as digital products from WYNDD. Suggestions are also given for how these models can be used, as well as for data collection and consolidation priorities for the future that could enhance the next generation of models for Wyoming's priority plant species.

METHODS

OVERVIEW

Presence data used to model the target species were derived from downloads of WYNDD's observation database (Biotics). Presence data for all plant species of concern, obtained from the same database, were used as background, or pseudo-absence data to allow for contrasting environmental conditions at sampled locations versus those where each species was recorded. These training presence and background data were evaluated against GIS layers representing a suite of biologically-relevant environmental gradients using Random Forest[12]. Random Forest is a machine-learning algorithm that builds upon CART models with a more computationally-intensive randomization algorithm that can boost performance dramatically over CART procedures. Random Forest was chosen over other approaches because it has been shown to perform well with relatively small sample sizes, and can automatically model interaction terms, non-linear responses, and categorical variables[12]. The resulting models were "projected" onto the environmental gradient

data to produce maps showing predicted distribution for the target species. Model training, evaluation, and assessment were carried out using methods commonly employed in distribution modeling.

Presence Data Collection and Processing

Threatened, Endangered and sensitive (TES) species represent 40 of the modeled species represented in this project, they are the only plant species in Wyoming with any federal status, and these TES species have all been the focus of one or more systematic WYNDD surveys. Each survey project sought to expand known distribution using some combination of remote sensing, GIS work with predictor layers, field reconnaissance, and extensive field work. As such, they have relatively robust presence data.

WYNDD's Biotics database was the source of all occurrence data used in building models, and provided approximately 3,000 observation records of the target species. The number of observation records available by species varied dramatically, from over 226 records for Beaver Rim phlox to just 2 total records for Winward's goldenweed (Table 1). Since species 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, records representing observations from before 1970 were excluded. Likewise, occurrence records with a mapping precision higher (i.e., worse) than 1200 meters were also excluded. This distance is commonly applied to the many plant records that have been located to section. Excluding less precise records reduces the possibility that a poorly-mapped point location will reduce model quality, by falsely indicating presence in an unsuitable setting.

Although species distribution modeling is typically based solely on points representing a documented observation for the target species, and values for the associated cells in GIS rasters representing the environmental predictors at that single, specific location, plant records from WYNDD's database comprise two basic and distinct types of representations, both referred to as "source features." First, WYNDD's database contains points of documented presence with an uncertainty buffer applied, resulting in a circular feature. All records that have collection data as their most detailed source of information are mapped as points.

Second, WYNDD maintains mapped polygons of occupied habitat, originating from boundaries drawn in GIS based on field notes and/or GPS data. Records that have survey data as their most detailed source of information are usually mapped as polygons, unless the entire source feature fits within an area of 20 m radius. These two different source feature types required different processing methods in order to relate the records to specific environmental gradient values (Figure 1). For buffered presence points, the centroid of each circular feature was used, as this minimized the potential spatial error in a point's location. For each mapped polygon of occupied habitat, a grid of points was generated at 30 m spacing within each polygon, aligned with the centroids of the 30 m raster cells used to represent environmental predictors. By creating multiple, gridded points to represent each habitat polygon rather than using a single polygon centroid, more information was available from each polygon. Although using all gridded points in a model would represent a form of pseudoreplication[13], iterative resampling (explained further in the "Model Generation" section, below) allowed the use of more of the information available in the gridded points while preventing pseudoreplication issues.

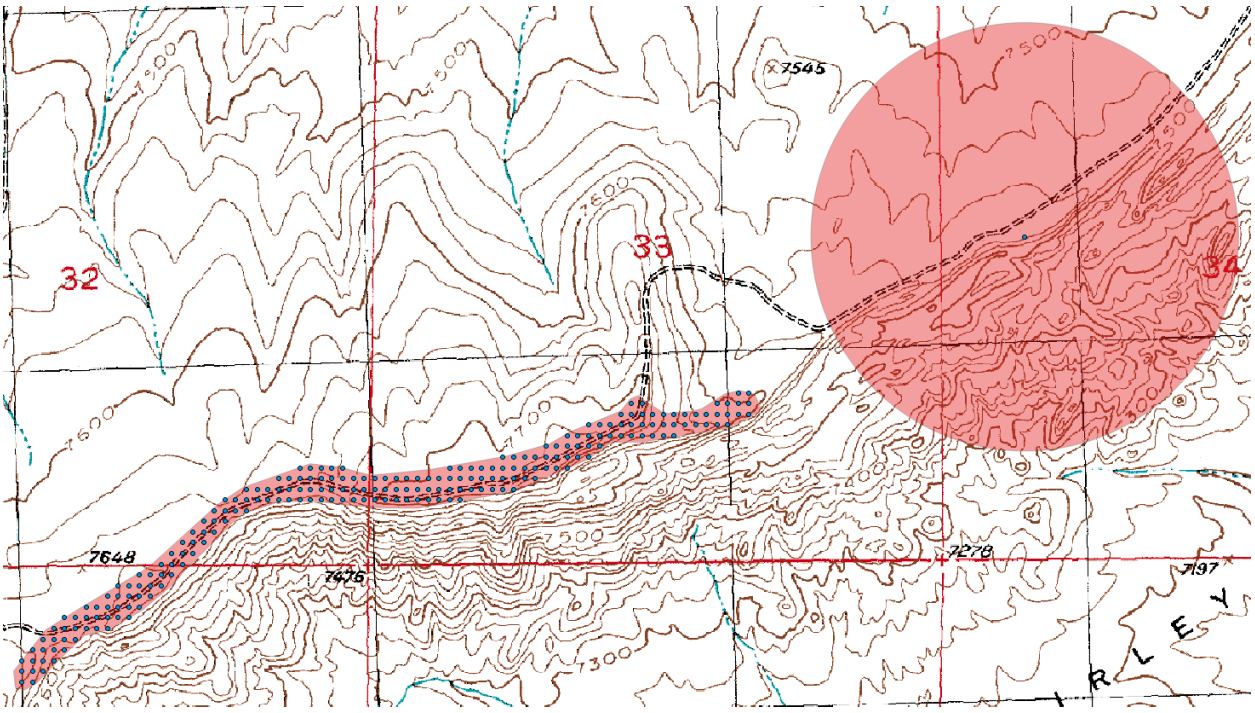

Figure 1. Polygon source features representing occupied habitat, as in the case of the elongate shape represented by a red polygon in the lower left, were sub-sampled using gridded points spaced at 30 m, shown here as blue dots. Source features representing buffered points were sub-sampled using a single centroid for the circular feature, as shown in the circular feature and corresponding centroid on the right. These two sets of points were combined to generate the presence points used in modeling.

Table 1. Presence location data by species. Totals indicate the number of presence locations for the species derived from WYNDD records. Modeling presence points indicate the number of presence locations used for model training after filtering for record age and precision.

Common Name	***Scientific Name***	***Federal Status***	***Total Presence Points***	***Modeling Presence Points***
Absaroka beardtongue	*Penstemon absarokensis*	Sensitive	136	134
Barneby's clover	*Trifolium barnebyi*	Sensitive	18	18
Beaver Rim phlox	*Phlox pungens*	Sensitive	226	225
Big Piney milkvetch	*Astragalus drabelliformis*	No status	117	107
Blowout penstemon	*Penstemon haydenii*	Endangered	28	28
Cary's beardtongue	*Penstemon caryi*	Formerly sensitive	109	100
Cedar Mountain Easter-daisy	*Townsendia microcephala*	Sensitive	12	12
Cedar Rim thistle	*Cirsium aridum*	Sensitive	40	29

Common Name	*Scientific Name*	*Federal Status*	*Total Presence Points*	*Modeling Presence Points*
Colorado butterfly plant	*Gaura neomexicana var. coloradensis*	Threatened	74	42
Desert yellowhead	*Yermo xanthocephalus*	Threatened	18	18
Devil's Gate Twinpod	*Physaria eburniflora*	No status	52	48
Dorn's Twinpod	*Physaria dornii*	Sensitive	56	54
Dubois milkvetch	*Astragalus gilviflorus var. purpureus*	Sensitive	86	42
Entire-leaved peppergrass	*Lepidium integrifolium*	Sensitive	23	23
Evert's waferparsnip	*Cymopterus evertii*	Sensitive	43	37
Fremont bladderpod	*Lesquerella fremontii*	Sensitive	99	94
Gibbens' beardtongue	*Penstemon gibbensii*	Sensitive	30	30
Green river greenthread	*Thelesperma caespitosum*	Sensitive	13	13
Hyattville milkvetch	*Astragalus jejunus var. articulatus*	Sensitive	13	13
Laramie columbine	*Aquilegia laramiensis*	Sensitive	81	81
Laramie false sagebrush	*Sphaeromeria simplex*	Sensitive	146	146
Large-fruited bladderpod	*Lesquerella macrocarpa*	Sensitive	34	34
Long-awned alkali wild-rye	*Elymus simplex var. luxurians*	Sensitive	221	220
Many-stemmed spider-flower	*Cleome multicaulis*	Sensitive	9	9
Meadow milkvetch	*Astragalus diversifolius*	Sensitive	16	16
Meadow pussytoes	*Antennaria arcuata*	Sensitive	92	92
Nelson's milkvetch	*Astragalus nelsonianus*	Formerly sensitive	87	75
Opal phlox	*Phlox opalensis*	No status	108	100
Owl Creek miner's candle	*Cryptantha subcapitata*	Sensitive	18	16
Ownbey's thistle	*Cirsium ownbeyi*	Sensitive	59	58
Pale blue-eye-grass	*Sisyrinchium pallidum*	Formerly sensitive	41	40
Payson beardtongue	*Penstemon paysoniorum*	No status	59	47
Persistent sepal yellowcress	*Rorippa calycina*	Sensitive	111	109
Porter's sagebrush	*Artemisia porteri*	Sensitive	164	164

Common Name	*Scientific Name*	*Federal Status*	*Total Presence Points*	*Modeling Presence Points*
Precocious milkvetch	*Astragalus proimanthus*	Sensitive	26	26
Prostrate bladderpod	*Lesquerella prostrata*	Sensitive	27	27
Rocky Mountain twinpod	*Physaria saximontana var. saximontana*	Sensitive	94	88
Shoshonea	*Shoshonea pulvinata*	Sensitive	52	51
Sidesaddle bladderpod	*Lesquerella arenosa var. argillosa*	Sensitive	23	21
Small rockcress	*Boechera pusilla*	Candidate (was Sensitive)	8	8
Stemless beardtongue	*Penstemon acaulis*	Sensitive	54	54
Trelease's racemose milkvetch	*Astragalus racemosus var. treleasei*	Sensitive	35	32
Tufted Twinpod	*Physaria condensata*	Sensitive	95	90
Uinta greenthread	*Thelesperma pubescens*	Sensitive	51	51
Ute ladies' tresses	*Spiranthes diluvialis*	Threatened	28	28
Ward's goldenweed	*Oonopsis wardii*	No status	55	47
Williams' waferparsnip	*Cymopterus williamsii*	Sensitive	64	64
Woolly Twinpod	*Physaria lanata*	No status	52	46
Wyoming tansymustard	*Descurainia torulosa*	Sensitive	37	30

NEGATIVE DATA COLLECTION AND PROCESSING

True absence data for a given species are seldom available, since researchers typically do not explicitly record locations where they surveyed for a species but failed to find it. Even when negative results from surveys are recorded, they seldom are databased in any way that makes them readily accessible in numbers sufficient for distribution modeling. Survey routes have routinely been recorded in field notes, but such negative data have not routinely been digitized and databased.

Moreover, it is often unclear whether a survey that did not record a species truly represents an unoccupied location; it is possible, instead, that the surveyor simply failed to detect the species, for a variety of reasons[14]. For example, many plant species cannot be found or identified with certainty outside of a discrete phenological window (e.g., Ute ladies-tresses), and some species are cryptic or at least small and easily-overlooked (e.g., Cedar Mountain Easter-daisy). A few species have the

capacity to remain alive belowground through the growing season so that individual plants, if not their populations, are dormant for all survey purposes (e.g., Ute ladies-tresses).

Given the typical lack of reliable absence data, methods have been developed to generate "pseudo-absence" or "background" data points[15, 16]. Such methods use background data in order to distinguish between the environmental gradients present in areas that are *used by* a species versus those that are *available to* the species. One method for creating a background dataset that is commonly used with the Maxent algorithm for distribution modeling, for example, is to select a large number (e.g., 10,000) of random points from the modeling area to represent the gradients available to a species[17]. However, this approach assumes that the presence dataset that will be contrasted to the random points is itself a product of random or exhaustive sampling.

Botanical initiatives of recent decades, including the pioneering studies of Robert Dorn and the floristic inventories of RM, have aimed for systematic approaches to botanical survey coverage across the state. WYNDD surveys have addressed the need for detailed information on the rarest species. However, these survey initiatives have had discrete scopes and logistical constraints, and nearly all were focused on public lands, so gaps remain in coverage of rare plant presence records.

If sampling bias is not accounted for, a presence-only modeling approach which uses randomized background points may produce a model that predicts sampling effort better than it predicts a species' true distribution[18]. These types of broader-scale sampling biases were addressed in this project by using a target background group approach[16], rather than the default method of selecting random background points or some other means of generating background data. This approach attempts to mirror spatial sampling bias in the presence data for a species by selecting background data – often records for related species – that derive from surveys exhibiting similar spatial biases. Matching the biases in the presence data for a target species with similar biases in the background data helps to factor out systemic sampling bias in modeling, resulting in a model that more accurately reflects a species' distribution. Further, since most of the rare plant presence points used in this modeling project derive from WYNDD botanical surveys, it is reasonable to assume that, had a species been present at any surveyed site, it would have been recorded. Thus, to generate background data for modeling a given species, locations from WYNDD's database for *all other species* of concern or potential concern were used. Rather than use the gridded points generated for all species, a centroid was generated for the source features representing all plant species of concern or potential concern in Wyoming, to avoid skewing the background dataset to the largest source features. Highly imbalanced sets can result in models that emphasize correct classification of the majority class – absence, in the case of the plant models – over correct classification of the minority class (presences). Thus, a down-sampling approach was also used to balance the number of presence and absence points used in modeling (see the "Model Generation" section below, for details).

For two BLM sensitive species, Meadow Milkvetch and Meadow Pussytoes, preliminary models generated as part of this project were used to select additional sampling targets for a separate BLM project that involved testing early rounds of the potential distribution models in 2014 field work, as well as revisiting imprecisely mapped populations, and identifying potential habitat using photointerpretation[19]. Field work resulted in approximately 10 new presence points based on photointerpretation, better mapping of several previously documented locations, and approximately 82 locations where surveys were conducted but the species were not found. The source features resulting from these new presence locations, and the surveyed, negative locations

were added to the presence and pseudo-absence sets, respectively, for subsequent modeling. New records of a third BLM sensitive species, Multi-stemmed spider-flower, were also documented, and presence points were generated from results that were added for revising its model as well. The field work also pointed to the need for adding wetland layers for modeling the potential distribution of these wetland species.

Environmental Data Organization and Processing

The predictor data used to build distribution models represent environmental characteristics or gradients identified as important in influencing species distributions, and are typically stored as raster datasets in a Geographical Information Systems (GIS) platform. Modelers commonly include predictor layers describing gradients related to climate, vegetation, elevation, and soils, but for selected species, more specific predictors representing other characteristics of landscape pattern, hydrology, interspecific interactions, or disturbance may be important in limiting distribution[8].

The linkages between a species' distribution and these predictor layers may be direct, as in the case of a grassland plant species that occurs only in locations with no forest canopy cover. 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 topographic position or climate might prove useful in modeling the species. Thus, a useful predictor set may contain attributes that are intuitively important to a species as well as attributes that are somewhat harder to interpret.

The factors that influence a species' distribution vary across differing spatial scales, from broad-scale gradients like climate to fine-scale parameters such as soil texture[20]. 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[21]. A list of potentially useful predictor data layers was generated after reviewing available information for the modeled species. Standard climatic, elevation, and vegetation predictors were added 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[22]. Topographic variables were derived from the National Elevation Dataset[23] 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 or wetland habitats, and prevalence of water on the landscape, based on hydrology layers prepared by the National GAP Program[24]. Land cover variables included percent cover for forest, shrubs, and herbaceous plants, as well as bare ground, from the LANDFIRE dataset[25, 26]. Soils predictors described chemistry, texture, and moisture parameters derived from the STATSGO dataset[27]. Many layers that were categorical in their native format (e.g., soils, vegetation) were transformed into continuous gradients to avoid issues that can result from inclusion of categorical predictors in models. More detailed descriptions and references for each variable are provided in Appendix 1.

MODEL GENERATION, VALIDATION, AND DISPLAY

The Geospatial Modeling Environment (GME[28]) was used to attribute the shapefiles representing training presences and background points with values for all potential predictor variables, and the associated attribute tables were exported as comma-delimited (CSV) files. Predictor variable values were evaluated for multicollinearity. Multicollinearity (i.e., strong correlations between predictor variables) can increase the standard errors of coefficients in regression[29], changing the interpretation of which predictors are most important in a model. Although algorithms such as Random Forest are more robust to the effects of multicollinearity than is regression, these methods may still overfit when the number of potential predictors is high relative to the number of training data points[8]. Thus, the set of predictor layers was evaluated for pairwise correlations using the statistical software, R[30]. Sets of potential predictors with high pairwise correlations were identified, and in each case the subset of predictors that had the lowest multicollinearity was retained. For example, the Bioclim predictors describing Mean Diurnal Temperature Range (bioclim2), Temperature Seasonality (bioclim4), and Annual Temperature Range (bioclim7) have moderate to high pairwise correlations (Figure 2). Bioclim2 and bioclim4 are the most *different* pair among this set, and can predict bioclim7 with an adjusted r-squared of 0.988. In this example set, bioclim7 was therefore excluded from the potential predictors because it contains the least unique information in the set.

After eliminating highly collinear predictors, each species was evaluated and the pertinent literature was reviewed to flag relevant predictors for inclusion in an initial round of modeling on a species-specific basis. An initial round of modeling was done to identify the most informative subset of predictors, and a subsequent round of modeling used the resulting predictor subset for each species.

Since using all gridded points for presence source features would amount to pseudoreplication, and using all background points would result in class imbalance[31, 32], an iterative approach was taken that used resampling to build many models from many subsets of the presence and absence data. For each iteration, a subsample of presence and absence data were selected, and a model was constructed based on those subsamples. Presence point subsamples for the iteration were created by randomly selecting a point from each source feature, to avoid pseudoreplication that would be caused by taking multiple points from a single source feature. Absence points were randomly "down-sampled"[32] so that there were three times the number of absence points compared to presence points.

Prior to modeling, a covariance matrix was generated from the full set of absence data, to provide a measurement of covariance for the predictor values at all absence points. Next, the initial round of modeling for each species used all potential predictors identified for the species, and generated summary statistics to allow for objective variable reduction in the next round. After each training set subsample (presence and absence) was drawn, a single Random Forest model was generated by growing 500 trees with the *mtry* parameter -- the number of predictors to try at each node -- set by default to the square root of the number of total predictors[12]. The subsample of absence data from that iteration was then appended to the absence subsamples from previous iterations, and a new covariance matrix was calculated from this cumulative set of absences. The resulting covariance matrix was then checked for equivalence with the covariance matrix for the full absence dataset (at

alpha level of 0.005), to ensure that the full range of variability contained in the full absence dataset, with respect to the predictor variable values, was captured in the cumulative subsets of

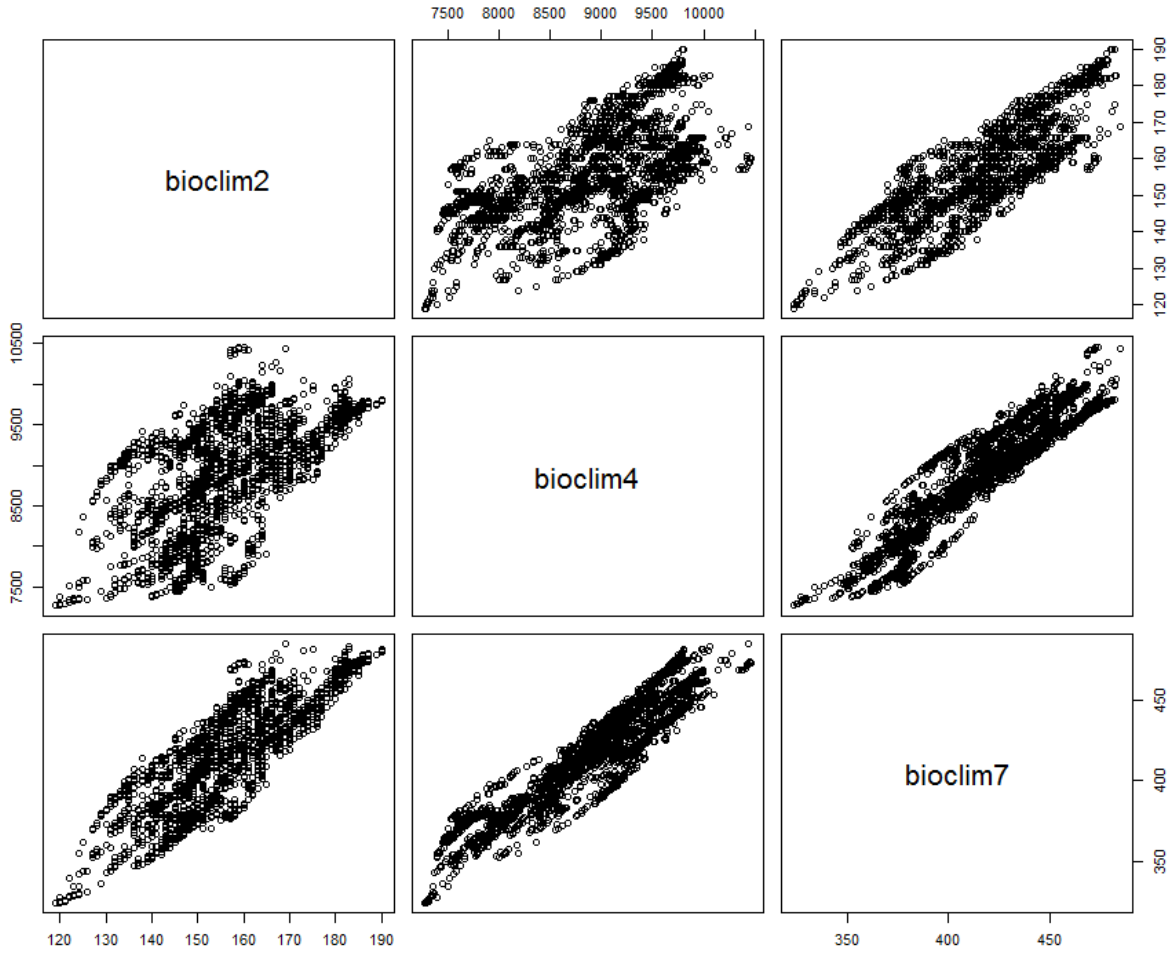

Figure 2. Mean Diurnal Temperature Range (bioclim2), Temperature Seasonality (bioclim4), and Annual Temperature Range (bioclim7) were correlated with one another at moderate to high levels in pairwise comparisons, with bioclim2 and bioclim4 being the most different. Bioclim7 was excluded from the set of potential predictors as it can be predicted by a combination of bioclim2 and bioclim4 with an adjusted R-squared of 0.988.

training absences. Iterative subsampling and modeling continued until covariance of background data converged and the number of iterations was at least 100. The resulting Random Forest models from each iteration were then combined into a single Random Forest model comprising at least 50,000 trees. A plot of the out-of-bag (OOB) error rate[12] for the combined model resulting from each iteration was generated to help determine whether models had stabilized within the number of iterations run. No spatial output was produced from this round of modeling, as the intent was simply to provide the summary statistics necessary to perform objective variable reduction in a subsequent round of modeling.

To determine which predictors should be removed for each species based on output from the first round of modeling, the Mean Decrease in Accuracy (MDA)[33] importance parameter was stored for each model iteration. Then, a p-value was generated as the proportion of iterations where the Mean Decrease in Accuracy for a predictor was *not* greater than 0. This is the probability, in other words, that permuting (randomizing) the values for a given predictor and re-running the model with the permuted values lead to a *better* model, which would only be expected if the variable were not truly informative. As the number of tests performed is equal to the number of predictors evaluated, the number of predictors included in the model was multiplied by the calculated p-value to get a Bonferroni-adjusted p-value. This Bonferroni-adjusted p-value was then used as a criterion for identifying predictors to exclude from the next modeling round. Any predictors with a Bonferroni-adjusted p-value > 0.05 were excluded from the next round of modeling. Permutation-based approaches such as this one have been shown to be less prone to overfitting when compared to a backward stepwise elimination[34-36].

Following the first round of modeling, statistical and spatial output was reviewed. Input datasets were modified in order to improve models, where possible. It provided a database quality control in which questionable presence points generated from the WYNDD database were identified as such for edits in the WYNDD database and removal from the modeling dataset. New (2015) locations for the two sensitive species of alkaline meadows were added to the presence dataset, while data representing negative surveys were added to the absence dataset.

Three predictor layers were added to the set of potential predictors, based on model output review and an evaluation of key predictor layer shortcomings in the first round of modeling. A "distance to wetland habitat" layer was generated by combining data layers on streams, wetlands, and other wetland habitat types, since wetland species, as a group, appeared to have the least refined models in the first round. A "bedrock calcium" layer was created by rescoring a bedrock geology layer into ordinal categories of calcium carbonate concentration, to help refine models for calciphiles – another group for which the first round of models were largely unsatisfactory. Finally, a "biome" layer was generated by combining ecoregions into a more generalized map, to help narrow model predictions to the appropriate biome for species that were extremely limited in this regard. Appendix 1 contains more details related to the creation of each of these new predictors.

After modification of the presence/absence and predictor datasets, two rounds of modeling were repeated: 1) an initial round with all potential predictors and no spatial output, to help identify the most important predictors; and 2) a second round using the reduced predictor set and writing full spatial output and summary statistics. The resulting set of models was compared to the original set generated in the first two rounds, to determine whether the revised input data improved the model for each species. Statistical output and spatial output were evaluated to determine which model was most useful for each species. As each of the 100 iterations of a model for a species generated 500 trees, the summary statistics were calculated as the mean of each statistic, based on OOB samples, across all 50,000 trees, using a predicted probability threshold of 50%. Statistics used to evaluate models included the OOB error, True Skill Statistic (TSS)[37], and sensitivity and specificity[38].

RESULTS

OVERVIEW

Of the fifty-two plant species identified as potential candidates for modeling, three were not able to be modeled due to data limitations. A preliminary model could not be generated for Winward's goldenweed, as only two presence locations exist for the species[39], so preliminary distribution models were created for 51 species. Following review of preliminary models, two other species – Wyoming locoweed and Desert glandular phacelia – were eliminated from the modeling set. Wyoming locoweed was eliminated because only a small fraction of known records for the species have currently been incorporated into WYNDD's database. Desert glandular phacelia was excluded from subsequent modeling due to taxonomic uncertainties.

Potential distribution models therefore were produced for 49 plant species, including TES species as well as others that are globally rare and that are found on BLM lands in Wyoming. Models for two of the Threatened and Endangered species – Desert yellowhead and Blowout penstemon – were rejected following review of all model output. The model generated for Desert yellowhead as part of this project was rejected due to the existence of a better model, generated previously as a separate project, with more intensive and species-specific methods[6]. The model for Blowout penstemon was rejected because the active sand dune habitats with which it is associated are dynamic in nature[40] and therefore cannot be represented by a static predictor layer. In total, predictive distribution models for 47 species were selected and delivered as final products of this project. A detailed summary of the final model for each species, including presence and predictor data overviews, model performance measures, and maps showing model output for each species can be found in Appendix 2 of this report.

PRESENCE DATA

The number of presence points used in modeling ranged from 8 (Small rockcress) to 225 (Beaver Rim phlox), with a median of 46. Useful species distribution models have been generated by other researchers with as few as 5-50 presence points[8, 15, 41-46], though there appears to be some consensus that having greater than 50 presence points results in more robust models[8]. Twenty-one species had greater than 50 presence points, while 13 species had fewer than 30 presence points, a lower limit suggested by several authors (see Franklin and Miller[8] for a thorough discussion of sample size considerations). Interestingly, the model with the poorest performance statistics – Payson beardtongue – had the median number of locations, 47, used in modeling, while the model for Colorado butterfly plant had a similar number of locations (42), and was among the best-performing models. Presence data for Payson beardtongue was almost exclusively based on collection records (i.e., single points recorded near where a species was collected), whereas Colorado butterfly plant had presence data based primarily on surveys that generated detailed polygons of occupied habitat, rather than a single collection point. This suggests that the polygon level data, and the iterative subsampling that was done to maximize the information gleaned from these polygon data, may enable the production of better models when compared to single-point data.

An alternate possible explanation is that Payson beardtongue occurs across a broader range of settings compared to other modeled species (foothills in southwest Wyoming to dry, semi-desert

basins in the central part of the state), whereas Colorado butterfly plant occupies a fairly narrow and specific niche (riparian zones in extreme southeast Wyoming). More presence locations are typically needed when a species is more broadly distributed and general in its habitat requirements[8]. Further, Payson beardtongue is not listed as a BLM Sensitive species, and has not been the focus of targeted surveys, meaning that the current collection of presence locations might offer an incomplete picture of its distribution.

The WYNDD database reflects current species taxonomic understanding, and presence data as used for this project followed the same framework. For example, it has been hypothesized that *Phlox pungens*, a state endemic, comprises two separate varieties in two separate geographic areas of its distribution. The taxonomic research is ongoing, and *Phlox pungens* presence points were modeled in a single set rather than as two sets.

Predictor Data

A total of 63 potential predictors (Table 2) were identified for inclusion in the variable selection run for each species, after eliminating predictors to reduce collinearity. The Bioclim predictor set[22] was trimmed from 19 to 10 potential predictors, to reduce collinearity in the predictor set tested for each species. Three soil parameters, Percent Clay, Percent Sand, and Percent Silt, were perfectly collinear as a set, so for each species where soil texture was identified as important, the most biologically-relevant subset of two were included as potential predictor, and the third was excluded. The Topographic Position Index (TPI[47, 48]) and Vector Ruggedness Measure (VRM[49]) predictors, each created using neighborhood sizes of 3, 5, 11, 21, and 31 cells, were highly collinear. Eliminating the versions of these predictors created with 5 and 21 cell neighborhoods greatly reduced collinearity among the set, so TPI and VRM predictors based on neighborhood sizes of 3, 11, and 31 cells were included as potential predictors, where identified as potentially informative. Topographic Position Index, in particular, may work best when layers calculated at multiple neighborhood sizes are used, as this helps to identify landforms (e.g., ridgetops, valley bottoms, midslopes[48]) that may be important to plants.

Bioclimatic predictors appeared in final models for the greatest number of species, and also had high average importance, across species. Bioclim13 (precipitation of wettest month) and bioclim6 (minimum temperature of coldest month) were included in nearly all models and had the highest average importance values of any predictors, across all species. The high relative importance of these two bioclimatic predictors that describe extremes in temperature and precipitation is in line with findings of other researchers who concluded that seasonal highs and lows are more important in defining species' distributions than are annual means of these climate parameters (see [8]).

Bare ground cover (bare), ruggedness (vrm11 and vrm31), topographic position (tpi_11 and tpi_31), elevation (elev), and biome all appeared in the final models for at least half of the modeled species, with elevation and bare ground cover having the highest average importance across all models. Many of the soil parameters (e.g., percent sand, percent silt, soil pH, and organic matter) were eliminated from the models for most species due to low importance values, but had high average importance values to the models in which they remained. Predictors that measured proximity to water or presence of saturated soil conditions (e.g., distance to wetland habitat, distance to water, available soil water content, ksat_surf) had high importance values for wetland and riparian species.

Three predictors developed specifically for this project – biome, calcium carbonate concentration and distance to wetland habitat – proved important in refining models for select species. Distance to wetland habitat refined models for wetland species by better representing various wetland types, compared to layers previously used in modeling. Biome helped to constrain predictions for species that primarily occurred in a limited set of biomes (e.g., reducing the amount of distribution predicted in montane areas for species that are only known from basin settings).

Table 2. Predictor layers included in modeling, with the number of final models that included the predictor and the average importance (Mean Decrease Accuracy) of that predictor across models including that predictor.

Name	***Predictor***	***Number of Models***	***Average Importance***
A[1] (Transformed Aspect -- Southeast/Northwest Gradient)	aprime135	2	0.002
A[1] (Transformed Aspect -- North/South Gradient)	aprime180	3	0.003
A[1] (Transformed Aspect -- Southwest/Northeast Gradient)	aprime45	3	0.004
A[1] (Transformed Aspect -- West/East Gradient)	aprime90	4	0.002
Available water capacity, top 200 cm	awc	4	0.009
Available water capacity, surface soil layer	awc_surf	2	0.003
Bare ground cover	bare	43	0.016
Mean Temperature of Warmest Quarter	bioclim10	47	0.021
Annual Precipitation	bioclim12	47	0.024
Precipitation of Wettest Month	bioclim13	45	0.036
Precipitation Seasonality (Coefficient of Variation)	bioclim15	46	0.026
Precipitation of Driest Quarter	bioclim17	47	0.018
Precipitation of Warmest Quarter	bioclim18	47	0.026
Mean Diurnal Range (Mean of monthly (max temp - min temp))	bioclim2	47	0.022
Isothermality (BIO2/BIO7) (* 100)	bioclim3	46	0.022
Temperature Seasonality (standard deviation *100)	bioclim4	46	0.027
Min Temperature of Coldest Month	bioclim6	46	0.035
Biome	biome	26	0.013
Calcium Carbonate Percentage top 200 cm	caco3	11	0.006
Calcium Carbonate Percentage surface layer of Soil	caco3surf	3	0.010
Soil cation-exchange capacity, top 200 cm	cec	0	NA
Soil cation-exchange capacity, surface soil layer	cec_surf	3	0.008
Conifer Index	confr	0	NA
Landscape Contagion Index	contag	1	0.015
Compound Topographic Index	cti	15	0.004
Distance to Cliffs	d2cliffs40	3	0.021
Distance to Rock Outcrop	d2outcrop	5	0.004
Distance to Permanent Standing Water	d2psw	2	0.014
Depth to shallowest restrictive layer	d2srl	2	0.014
Distance to Any Water	d2w	3	0.028

Name	*Predictor*	*Number of Models*	*Average Importance*
Distance to Wetland Habitat	d2wethab	8	0.025
Depth to water table	dep2watr	2	0.002
Soil Electrical Conductivity top 200 cm	ec	1	0.003
Elevation	elev	31	0.019
Flooding Frequency Class	flood_freq	2	0.001
Mean Forest Cover	forest	2	0.003
Calcium rating of bedrock geology formation	geol_calc	6	0.019
Growing Degree Days	growdd	1	0.029
Herbaceous Cover Index	herb	1	0.002
Heat Load Index	HLI	9	0.003
Erosion Factor, K, Whole Soil, surface layer of Soil	kfact_surf	1	0.017
Saturated Hydraulic Conductivity (KSAT), top 200 cm	ksat	2	0.003
Saturated Hydraulic Conductivity (KSAT), surface soil layer	ksat_surf	2	0.008
Soil organic matter, top 200 cm	orgmat	9	0.024
Soil organic matter, surface soil layer	orgmatsurf	4	0.032
Percent clay, top 200 cm	pclay	6	0.009
Percent clay, surface soil layer	pclaysurf	2	0.033
Percent sand, top 200 cm	psand	18	0.021
Percent sand, surface soil layer	psandsurf	6	0.028
Percent silt, top 200 cm	psilt	19	0.016
Percent silt, surface soil layer	psiltsurf	6	0.009
Sagebrush Index	sage	2	0.007
Soil sodium adsorption ratio top 200 cm	sar	1	0.007
Shrub Index	shrub	3	0.018
Degree Slope	slope	22	0.008
Soil pH, top 200 cm	soilph	14	0.022
Soil pH, surface soil layer	soilphsurf	6	0.013
Topographic Position Index, 11-cell focal window	tpi_11	28	0.004
Topographic Position Index, 3-cell focal window	tpi_3	15	0.002
Topographic Position Index, 31-cell focal window	tpi_31	27	0.007
Vector Ruggedness Measure, 11-cell focal window	vrm11	29	0.010
Vector Ruggedness Measure, 3-cell focal window	vrm3	18	0.005
Vector Ruggedness Measure, 31-cell focal window	vrm31	36	0.014

DISTRIBUTION MODEL OUTPUT

Initial model runs indicated that covariance matrices derived from cumulative, subsampled presence and pseudo-absence locations converged within 100 subsampling iterations for nearly all species (Figure 3a). This meant that by the time 100 subsampling iterations had been performed, the majority of variability in predictor layer values across all presence/absence locations had been

captured, so performing additional iterations would not have a significant impact on the resulting models. Thus, 100 subsampling iterations were used in all subsequent modeling. Likewise, preliminary modeling showed that OOB error converged with 500 or fewer trees grown per subsampling iteration (Figure 3b), so this number of trees was used for all subsequent models.

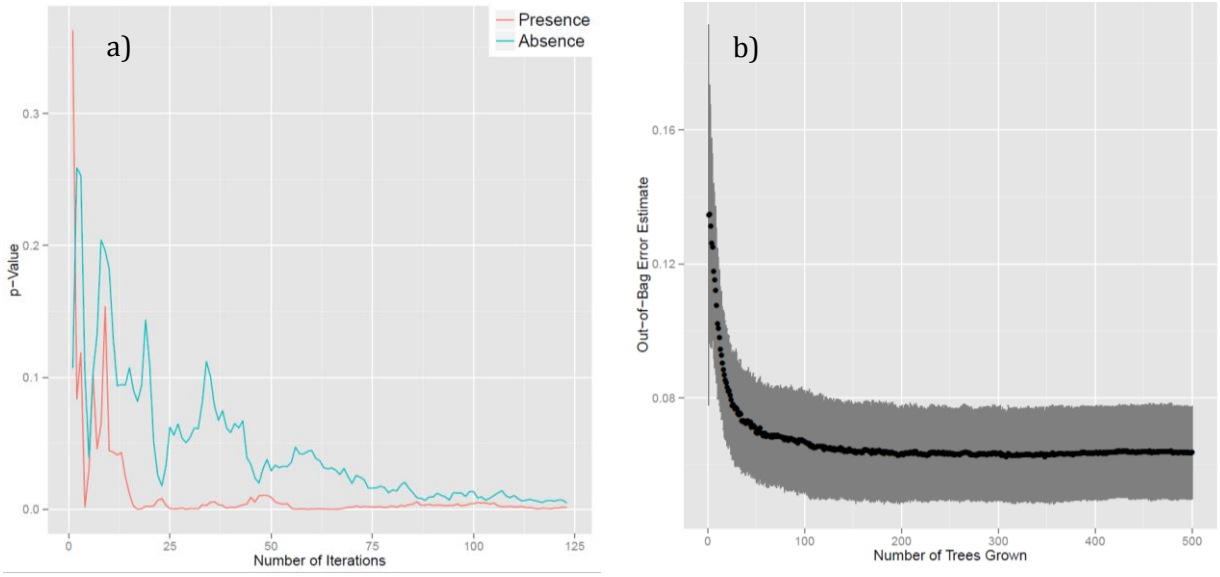

Figure 3. P-values for covariance matrix equivalence tests for cumulative subsamples of presence and absence data versus the full set of presence and absence data, by subsampling iteration, for Williams' waferparsnip (a). Most species showed a pattern similar to this one, indicating that 100 subsampling iterations were sufficient for capturing the variability in predictor values across all presence and absence points. OOB error by number of trees grown in each subsampling iteration, for Williams' waferparsnip (b). Black dots indicate the mean OOB error associated with the number of trees grown in a given Random Forest grown for a subsample of training data; gray bars show one standard deviation around the mean. Models for all species exhibited a similar pattern, suggesting that growing 500 trees per subsample iteration ensured error convergence.

The median OOB error across final models for all species was 3.1%, indicating that final models for most species had low error rates and high accuracy for OOB data points. Similarly, median values were high for TSS (91.3%), Sensitivity (93.6%), Specificity (98.0%), and Kappa (91.8%) across all species. The model for Long-awned alkali wild-rye had the highest model performance statistics, overall, with an OOB error of 0.3%. This species is restricted to active sand dunes and adjacent plains with sandy soils, in a contiguous area in southwest Wyoming[50]. Unlike the dune fields on which Blowout penstemon occurs, the features upon which Long-awned alkali wild-rye depends appear to be well defined by available soils data and other predictor layers. The most important predictors included in the model for Long-awned alkali wild-rye, in descending order of importance, were soil organic matter, shrub cover, precipitation of warmest quarter, and percent sand, and the relationships with these predictor layers were relatively strong and were consistent with habitat descriptions for the species[50].

It is worth noting that the summary statistics presented here pertain to binary versions of models created with a logistic probability threshold set at 50%. Though a commonly used, default

threshold, 50% is often not the optimal threshold for a distribution model based on presence and pseudo-absence data deriving from spatially biased or otherwise unrepresentative sampling[51]. Thus, binary versions of each model based on a refined probability threshold likely would have higher model performance scores.

Model deliverables for each species are highlighted in Appendix 2 and comprise two expressions of the final model selected for the species: 1) a version representing the raw, predicted probability values; and 2) a four-category version created by applying thresholds to produce a simplified and more interpretable layer. The three threshold values used to generate the four-category expression of the model were:

1) The minimum predicted probability assigned to any known presence locations
2) The predicted probability associated with the 25th percentile of probability values assigned to known presence locations (i.e., a threshold selected such that 25% of known presence locations had *lower* predicted probabilities than this threshold)
3) The predicted probability associated with the 75th percentile of probability values assigned to known presences

Binning the model output into categories based on predicted probability at known presences resulted in models that display varying levels of likelihood, while at the same time conveying a sense of the uncertainty that is inherent with any model. The categories resulting from application of the above thresholds can be interpreted as:

1) Predicted Absent
2) Low predicted probability of presence
3) Medium predicted probability of presence
4) Highest predicted probability of presence

Further interpretations and guidance regarding these categories can be found in the Discussion section. Additionally, another threshold, "MaxTSS" -- the threshold that maximizes the TSS – was identified using the GENetic Optimization Using Derivatives routine in the rgenoud[52] package for R. This is an optimization routine that was used to identify an optimal threshold that maximizes a function summing the specificity and sensitivity for a binary version of the model with a given threshold. Applying this threshold to the probability raster will yield a binary version of each model that balances the tradeoff between correctly predicting presence, and minimizing incorrect prediction of absences as presence (i.e., commission error).

Table 3. Model performance for final models, by species. Statistics shown here are based on out-of-bag (OOB) samples, using a classification threshold of 50% predicted probability.

Common	OOB Error	TSS	Sensitivity	Specificity	Kappa	Max S+S Threshhold
Absaroka beardtongue	6.8%	83.9%	89.5%	94.5%	82.3%	0.5911
Barneby's clover	4.4%	87.7%	90.3%	97.4%	88.2%	0.5285
Beaver Rim phlox	5.1%	88.8%	93.4%	95.4%	86.8%	0.4783
Big Piney milkvetch	4.6%	89.5%	93.4%	96.1%	88.0%	0.6666
Cary's beardtongue	5.5%	86.3%	90.5%	95.8%	85.4%	0.6680
Cedar Mountain Easter-daisy	2.8%	92.9%	94.9%	98.0%	92.7%	0.7624

Common	OOB Error	TSS	Sensitivity	Specificity	Kappa	Max S+S Threshhold
Cedar Rim thistle	7.7%	77.8%	82.0%	95.8%	79.2%	0.6443
Colorado butterfly plant	0.9%	98.0%	98.7%	99.3%	97.7%	0.4500
Devil's Gate twinpod	4.8%	87.6%	90.9%	96.7%	87.4%	0.5112
Dorn's twinpod	2.0%	95.2%	96.7%	98.5%	94.8%	0.7393
Dubois milkvetch	0.9%	97.1%	97.4%	99.7%	97.6%	0.6404
Entire-leaved Peppergrass	1.8%	94.0%	94.7%	99.3%	95.0%	0.6938
Evert's waferparsnip	6.0%	83.3%	86.8%	96.4%	83.9%	0.4225
Fremont bladderpod	1.2%	95.6%	95.7%	99.8%	96.8%	0.5964
Gibbens' beardtongue	2.7%	90.3%	90.8%	99.5%	92.7%	0.4357
Green river greenthread	1.3%	98.2%	99.8%	98.3%	96.6%	0.8875
Hyattville milkvetch	0.6%	99.2%	99.9%	99.3%	98.5%	0.5854
Laramie columbine	1.4%	94.5%	94.7%	99.8%	96.1%	0.6786
Laramie false sagebrush	0.9%	98.1%	98.8%	99.2%	97.7%	0.5575
Large-fruited bladderpod	6.0%	86.0%	91.1%	94.9%	84.3%	0.4027
Long-awned alkali wild-rye	0.3%	99.3%	99.6%	99.8%	99.2%	0.7032
Many-stemmed spider-flower	1.8%	97.4%	99.7%	97.7%	95.3%	0.7180
Meadow milkvetch	5.8%	83.7%	87.3%	96.5%	84.3%	0.4568
Meadow pussytoes	1.6%	95.6%	96.6%	99.0%	95.7%	0.4764
Nelson's milkvetch	6.7%	79.8%	83.3%	96.6%	81.6%	0.6331
Opal phlox	6.5%	82.6%	86.9%	95.8%	82.7%	0.5173
Owl Creek miner's candle	3.6%	90.7%	93.3%	97.4%	90.4%	0.1996
Ownbey's thistle	3.1%	91.5%	93.6%	98.0%	91.6%	0.4359
Pale blue-eye-grass	4.9%	85.7%	88.5%	97.3%	86.7%	0.6107
Payson Beardtongue	14.6%	57.8%	65.8%	92.0%	59.8%	0.5921
Persistent sepal yellowcress	3.1%	90.9%	92.4%	98.5%	91.8%	0.7264
Porter's sagebrush	0.9%	97.7%	98.3%	99.4%	97.7%	0.5339
Precocious milkvetch	1.0%	98.6%	99.9%	98.7%	97.4%	0.8373
Prostrate bladderpod	2.7%	92.9%	94.9%	98.0%	92.7%	0.7109
Rocky Mountain twinpod	6.1%	82.3%	85.7%	96.6%	83.5%	0.4242
Shoshonea	4.9%	87.2%	90.5%	96.7%	87.0%	0.4207
Sidesaddle bladderpod	0.5%	99.3%	100.0%	99.4%	98.7%	0.6917
Small rockcress	0.5%	99.3%	100.0%	99.3%	98.6%	0.8574
Stemless beardtongue	1.9%	95.6%	97.1%	98.5%	95.1%	0.5816
Trelease's racemose milkvetch	4.9%	88.9%	93.2%	95.8%	87.2%	0.7572
Tufted twinpod	3.5%	91.3%	94.0%	97.3%	90.6%	0.6463
Uinta greenthread	1.4%	97.5%	99.0%	98.5%	96.4%	0.5918
Ute ladies' tresses	1.3%	97.9%	99.4%	98.5%	96.6%	0.8618
Ward's goldenweed	3.5%	89.2%	90.9%	98.4%	90.5%	0.5857
Williams' waferparsnip	2.5%	93.0%	94.6%	98.4%	93.2%	0.5132
Woolly twinpod	7.7%	77.3%	81.2%	96.1%	79.1%	0.5026
Wyoming tansymustard	6.5%	83.3%	88.0%	95.3%	82.7%	0.4944

DISCUSSION

USAGE AND LIMITATIONS OF DISTRIBUTION MODELING

In species distribution modeling, there is uncertainty and error inherent in presence and pseudo-absence points, predictor layers, and the underlying mechanistic processes that shape actual distribution. Presence points can be mismapped in processing data or in any facet of reporting field results, including transcription error or over- or under-representation of occupied habitat. Presence points can also be based on misidentification, a shortcoming that WYNDD addresses in quality control steps. Finally, presence points may not be representative of a species' distribution, or they can derive from biased sampling efforts and as a consequence suggest an unrealistic picture of the species' distribution. Predictor layers can exhibit error in both position and value, leading to spurious conclusions about the relationship between a predictor and a species' distribution. Finally, the underlying mechanisms that influence a species' distribution can be inordinately complex, nuanced, or otherwise challenging to represent accurately with a simplified model. For many rare plant species in Wyoming, distribution appears to be a function of both available habitat and processes associated with geographic isolation. Thus, models like those prepared in this project represent "potential distribution." Despite the presence of error and uncertainty, such models remain useful hypotheses about a species' geographic distribution, as long as users understand the inherent limitations of each model.

MODEL INTERPRETATION AND USAGE

The output values from distribution models are commonly thought of as a logistic probability of a species' presence, but the actual interpretation is typically more nuanced. Without substantial and representative absence data, it is impossible to determine the species' extent across large landscapes[51], since it is unknown whether empty spots on the "dot maps" of species observations are truly unoccupied. Thus, there is no direct way to estimate the true probability of a species' presence at any given location. Instead, output values from models such as the ones developed for this project should be viewed as relative indices of suitability for a species. Output from two different models cannot be directly compared (i.e., a value of 0.5 in a model for one species may not mean the same thing as a value of 0.5 in another species' model, and the true probability of a species occurring in such a location may not be 50% in either case). Higher output values should generally correspond with a higher probability of presence, and vice-versa, so models can be used to identify the areas that have the highest potential for species' presence.

Distribution models such as those produced in this project can help identify the species of interest that are potentially present in a proposed project area. They might also identify areas of potentially high concentrations of target species, or potential habitat for a priority species, at a coarse scale. Planners can use such maps to identify possible locations for Areas of Critical Environmental Concern, or to help them determine areas that may be more suitable for development with minimal adverse impacts to biodiversity or to a particular species[8].

Distribution models can also be used to guide 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 species to make the most of limited field project budgets. Moreover, by evaluating model output in the context of known presence points, researchers can focus on areas a model deems suitable but

that currently have no known records for the species, potentially expanding its known distribution. However, models should not be used in place of site-level, clearance surveys for TES species, 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 species is predicted to be "in the neighborhood," in which case field surveys are likely warranted.

Final model products from this project were delivered both as continuous, predicted suitability values and as simplified output showing four ordinal categories of suitability. Any use of a distribution model may require expressing the model differently by applying different thresholding or symbology in mapping the model output. A biologist interested in locating a particular species, for example, would most benefit by limiting their sampling to only the areas predicted to be most highly suitable for their target species (i.e., focusing on only the top-most category – "Highest predicted probability of presence"). Conversely, a manager tasked with evaluating the potential impact of development for a species or group of species may want to err on the side of caution, by considering even areas of lower predicted likelihood of presence to be potentially occupied and warranting field surveys (i.e., ruling out only the "Predicted Absent" category").

Caution must also be exercised when evaluating partial plots: graphs showing likelihood of presence as a function of each variable, holding all other variables constant. While indirect predictor layers such as elevation might contribute substantially 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, vegetation, soils, or other gradients that more directly limit a species' distribution. Biological understanding is thus important in interpreting partial plots – particularly those for more indirect predictors.[53]

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[54, 55], possibly leveling off somewhat at 50 to 100 training presence points[15, 42, 43]. Approximately half of the modeled species had 50 or more usable presence points, and all but two had 10 or more points, so it is likely that the resulting models will have utility in management. Some of the species represented by these models may have distributions that extend beyond the currently known distributions; in these cases, substantially better models may result if additional, independent observations are made in expanded portions of the species' distributions and added to the modeling sets for these species.

True negative (i.e., absence) records were not available in a readily usable format for the set of modeled species. While modeling based on presence-background data is common, true negative data can be used to provide models that discriminate more sharply between areas of predicted presence and absence[56]. Additionally, true absence data allow the modeler a broader suite of potential modeling algorithms, including standard statistical methods such as generalized linear models[29], or methods like occupancy modeling[57] that directly account for imperfect detection, when sample sizes are sufficiently large. Absence data may be difficult to generate, as it requires relatively detailed knowledge of survey effort and design, and the amount of survey effort required to confidently assign a location as an absence varies by species[58]. Nevertheless, given the benefits of absence data for distribution modeling, it warrants further consideration. WYNDD is currently in

the process of creating a new observations database that will provide the ability to store and compile survey efforts and negative datasets for species. This will allow for the production of more refined models as time goes by and negative data accumulate in the database.

Predictor Data Limitations

For most plants, soil characteristics are extremely important in limiting distribution, and this has been recognized for native plants in Wyoming[59]. Unfortunately, detailed digital soils data layers (SSURGO[27]) are not currently available as a statewide coverage for Wyoming, and likely will not become available for a number of years (J. Bauchert, pers. comm.). Although portions of the state have coverage of the detailed soils layers, making use of the data in a subset of the study area introduces large areas where no predictions can be made, as they lack the necessary predictor values. Statewide soils layers provide some information that can help in model building, but completion of the more detailed, SSURGO soil data layer would allow much more precise predictions to be made. Users of the models can use SSURGO data layers, in combination with model output if SSURGO coverage is available for their project area.

For at least one of the species modeled in this project -- Blowout penstemon – habitat suitability is a dynamic characteristic on the landscape[40]. This species is restricted to very active sand dune features[60, 61], and both the dunes and features within them are discontinuous, meaning that the spatial extent and configuration of habitat may vary over time. While some existing GIS data layers map dune extents over time, any models using these layers as predictors would need to be updated frequently to provide a current prediction of species distribution. One practical alternative is to maintain a more general distribution model for such species, and to use ancillary data such as digital aerial imagery, to guide precise field work or assist with assessing and mapping habitat quality.

For other species, such as Ute ladies-tresses, there were excellent presence/pseudo-absence data for eastern Wyoming but not for the rest of the state. In theory, results of WYNDD wetland survey projects could be used to generate meaningful absence data. But at this point, the species' distribution from adjoining states cannot readily be put to use because the environmental data layers do not span state boundaries.

Suggestions for Future Work

As with any analysis or modeling project, collecting additional training data can improve distribution models. Clearly, additional observation records for modeled species – particularly records some distance away from existing records – will provide additional information for modeling. Similarly, absence data for the modeled species 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 established statistical methods like regression. 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 species is surveyed for but not found, or it can be created retroactively based on prior survey work that found other species, but not the target species. While creating pseudo-absence data from locations where other species were recorded seems reasonable in this case, explicitly building negative datasets for

each species could greatly improve models, particularly at fine scales. WYNDD has recently created a new database that allows for the structured storage and retrieval of negative data. Consolidating the negative records that currently exist as assorted GIS files and printed records into the new database would help lay the groundwork for a future modeling effort using true presence/absence methods.

As with presence point data, collection or generation of newer and better predictor datasets should continue to be a priority for modeling work. This includes refinement of existing data layers, and development of new data layers based on remotely sensed data that are made available on a regular basis as satellite imagery becomes more ubiquitous. Ideally, there would be replacement of surrogate layers with ones that are directly derived. The rare species addressed in this project are habitat specialists, and while we tried to address those that are calciphiles, there are others that have substrate requirements that are not shared with other species. If time permits, developing species-specific layers by rescaling, scoring, or combining other datasets may improve models.

Land cover layers were not directly used in constructing the models for these species. 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.[62] If a conceptual understanding of a species' distribution suggests that vegetative community strongly influences distribution, one could assign species-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.[5] 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 species 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.[4]

ACKNOWLEDGEMENTS

We thank the BLM for providing funding for this project, under agreement L12AC20036. Joy Handley of WYNDD lead the multi-year work of digitizing plant species of concern distribution data before the start of this project, assisted in entering some of the new 2015 records, in correcting errors in existing records, and in reviewing model output. Melanie Arnett provided data exports at key times. Jeffrey Evans of The Nature Conservancy provided helpful comments and R code related to the subsampling approach and covariance convergence testing. Timothy Howard of the New York Natural Heritage Program suggested methods for decreasing the time required for each model run in R. Victoria Ramos assisted with compilation of graphics and tables for Appendix 2.

REFERENCES

[1] U.S. Department of the Interior Bureau of Land Management (BLM) Wyoming. (2010) BLM Wyoming Sensitive Species Policy and List.

[2] Heidel, B. (2012) Wyoming plant species of concern, Wyoming Natural Diversity Database, University of Wyoming, Laramie, WY.

[3] Fertig, W., and Thurston, R. (2003) Modeling the potential distribution of BLM Sensitive and USFWS Threatened and Endangered plant species in Wyoming. Report prepared for the Bureau of Land Management Wyoming State Office by the Wyoming Natural Diversity Database., Laramie, WY.

[4] Beauvais, G. P., Andersen, M. D., and Keinath, D. A. (2012) 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.

[5] Keinath, D., Andersen, M., and Beauvais, G. (2010) 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*.

[6] Heidel, B., Handley, J., and Andersen, M. (2011) Distribution and habitat requirements of Desert Yellowhead (Yermo xanthocephalus), Fremont County, Wyoming.

[7] Andersen, M., and Beauvais, G. (2013) Predictive Distribution Modeling of Species of Greatest Conservation Need in Texas. Report prepared by the Wyoming Natural Diversity Database., University of Wyoming, Laramie, WY.

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

[9] Breiman, L., Friedman, J., Olshen, R., and Stone, C. (1984) *Classification and Regression Trees*, Wadsworth, Inc., Belmont, California, United States of America.

[10] Cutler, D. R., Edwards, Thomas C., Jr., Beard, Karen H., Cutler, Adele, Hess, Kyle T., Gibson, Jacob, Lawler, Joshua J. (2007) Random forests for classification in ecology, *Ecology 88*, 2783-2792.

[11] Phillips, S. J., Anderson, R. P., and Schapire, R. E. (2006) Maximum entropy modeling of species geographic distributions, *Ecological Modeling 190*, 231-259.

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

[13] Hurlbert, S. H. (1984) Pseudoreplication and the design of ecological field experiments, *Ecological monographs 54*, 187-211.

[14] MacKenzie, D., Nichols, J., Royle, J., Pollock, K., Bailey, L., and Hines, J. (2005) *Occupancy estimation and modeling: inferring patterns and dynamics of species occurrence*, Academic Press.

[15] Elith, J., Graham, C. H., Anderson, R. P., Dudík, M., Ferrier, S., Guisan, A., Hijmans, R., Huettmann, F., Leathwick, J. R., Lehman, A., Li, J., Lohmann, L. G., Loiselle, B. A., Manion, G., and Moritz, C. (2006) Novel methods improve prediction of species' distributions from occurrence data, *Ecography 29*, 129-151.

[16] Phillips, S. J., Dudík, M., Elith, J., Graham, C. H., Lehmann, A., Leathwick, J., and Ferrier, S. (2009) Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data, *Ecological Applications 19*, 181-197.

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

[18] Venier, L. A., and Pearce, J. L. (2007) 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*.
[19] Heidel, B. (2014) Inventory of Alkaline Meadows for BLM Sensitive Plant Species: Antennaria arcuata (Meadow pussytoes), Astragalus diversifolius (Meadow milkvetch) and Cleome multicaulis (Many-stemmed Spiderflower) with Field-Testing of Potential Distribution Models; Fremont and Sweetwater Counties, Wyoming.
[20] Johnson, D. H. (1980) The comparison of usage and availability measurements for evaluating resource preference, *Ecology 61*, 65-71.
[21] Meyer, C. B., and Thuiller, W. (2006) Accuracy of resource selection functions across spatial scales, *Diversity and Distributions 12*, 288-297.
[22] Hijmans, R. J., Cameron, S. E., Parra, J. L., Jones, P. G., and Jarvis, A. (2005) Very high resolution interpolated climate surfaces for global land areas, *International journal of climatology 25*, 1965-1978.
[23] Gesch, D., Oimoen, M., Greenlee, S., Nelson, C., Steuck, M., and Tyler, D. (2002) The National Elevation Dataset: Photogrammetric Engineering and Remote Sensing, *68*, 5-11.
[24] US Geological Survey (USGS) Gap Analysis Program (GAP). (2011) USGS Gap Analysis Program Ancillary Data - Hydrography, February 22, 2011 ed., USGS Gap Analysis Program Ancillary Data - Hydrography, http://gapanalysis.usgs.gov/data/species-data/.
[25] U.S. Department of Interior - Geological Survey. (2013) LANDFIRE: LANDFIRE Existing Vegetation Height layer, LANDFIRE: LANDFIRE Existing Vegetation Height layer, Available: http://landfire.cr.usgs.gov/viewer/.
[26] U.S. Department of Interior - Geological Survey. (2013) LANDFIRE: LANDFIRE Existing Vegetation Cover layer, LANDFIRE: LANDFIRE Existing Vegetation Cover layer, Available: http://landfire.cr.usgs.gov/viewer/.
[27] Soil Survey Staff. General Soil Map (STATSGO2). Available online at http://sdmdataaccess.nrcs.usda.gov/, Natural Resources Conservation Service, United States Department of Agriculture.
[28] Beyer, H. L. (2012) Geospatial Modelling Environment (Version 0.7.2.1), Hawthorne L. Beyer, URL: http://www.spatialecology.com/gme.
[29] Hosmer, D. W., and Lemeshow, S. (1989) *Applied Logistic Regression*, John Wiley & Sons, New York, New York, USA.
[30] R Core Team. (2013) R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing. Available at: http://www.R-project.org, Vienna, Austria.
[31] Chawla, N. V., Japkowicz, N., and Kotcz, A. (2004) Editorial: special issue on learning from imbalanced data sets, *ACM Sigkdd Explorations Newsletter 6*, 1-6.
[32] Chen, C., Liaw, A., and Breiman, L. (2004) Using random forest to learn imbalanced data, *University of California, Berkeley*.
[33] Nicodemus, K. K., and Malley, J. D. (2009) Predictor correlation impacts machine learning algorithms: implications for genomic studies, *Bioinformatics 25*, 1884-1890.
[34] Svetnik, V., Liaw, A., Tong, C., and Wang, T. (2004) Application of Breiman's random forest to modeling structure-activity relationships of pharmaceutical molecules, In *Multiple Classifier Systems*, pp 334-343, Springer.
[35] Strobl, C., Boulesteix, A.-L., Zeileis, A., and Hothorn, T. (2007) Bias in random forest variable importance measures: Illustrations, sources and a solution, *BMC bioinformatics 8*, 25.
[36] Hapfelmeier, A., and Ulm, K. (2013) A new variable selection approach using Random Forests, *Computational Statistics & Data Analysis 60*, 50-69.
[37] Allouche, O., Tsoar, A., and and Kadmon, R. (2006) Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS), *Journal of Applied Ecology 43*, 1223-1232.

[38] Fielding, A. H., and Bell, J. F. (1997) A review of methods for the assessment of prediction errors in conservation presence/absence models, *Environmental Conservation 24*, 38-49.
[39] Fertig, W. (2012) Status of Winward's goldenweed (Ericameria discoidea var. winwardii) in Wyoming. Unpublished report prepared for the Bureau of Land Management Wyoming State Office and Wyoming Natural Diversity Database by Moenave Botanical Consulting, Kanab, UT.
[40] Heidel, B., Cox, S., and Blomquist, F. (2014) Dune habitat trends of an Endangered species, Penstemon haydenii (blowout penstemon) in Wyoming, Wyoming Natural Diversity Database, University of Wyoming, Laramie, WY.
[41] Hernandez, P. A., Graham, C. H., Master, L. L., and and Albert, D. L. (2006) The effects of sample size and species characteristics on performance of different species distribution modeling methods, *Ecography 29*, 773-785.
[42] Stockwell, D. R., and Peterson, A. T. (2002) Effects of sample size on accuracy of species distribution models, *Ecological modelling 148*, 1-13.
[43] Kadmon, R., Farber, O., and Danin, A. (2003) A systematic analysis of factors affecting the performance of climatic envelope models, *Ecological Applications 13*, 853-867.
[44] Loiselle, B. A., Jørgensen, P. M., Consiglio, T., Jiménez, I., Blake, J. G., Lohmann, L. G., and Montiel, O. M. (2008) Predicting species distributions from herbarium collections: does climate bias in collection sampling influence model outcomes?, *Journal of Biogeography 35*, 105-116.
[45] Kremen, C., Cameron, A., Moilanen, A., Phillips, S., Thomas, C., Beentje, H., Dransfield, J., Fisher, B., Glaw, F., and Good, T. (2008) Aligning conservation priorities across taxa in Madagascar with high-resolution planning tools, *Science 320*, 222-226.
[46] Wisz, M. S., Hijmans, R., Li, J., Peterson, A. T., Graham, C., and Guisan, A. (2008) Effects of sample size on the performance of species distribution models, *Diversity and Distributions 14*, 763-773.
[47] Jenness, J. (2006) Topographic Position Index (tpi_jen.avx) extension for ArcView 3.x, v. 1.3a, Jenness Enterprises.
[48] Weiss, A. D. (2001) Topographic Position and Landforms Analysis, Poster Presentation, (Conservancy, T. N., Ed.), San Diego, CA.
[49] Sappington, M. J., Longshore, K. M., and Thompson, D. B. (2007) Quantifying Landscape Ruggedness for Animal Habitat Analysis: A Case Study Using Bighorn Sheep in the Mojave Desert, *Journal of Wildlife Management 71*, 1419-1426.
[50] Heidel, B. (2012) Status of Elymus simplex var. luxurians (Long-awned alkali wild-rye), southwestern Wyoming.
[51] Hastie, T., and Fithian, W. (2013) Inference from presence-only data; the ongoing controversy, *Ecography 36*, 864-867.
[52] Mebane Jr, W. R., and Sekhon, J. S. (2011) Genetic optimization using derivatives: the rgenoud package for R, *Journal of Statistical Software 42*, 1-26.
[53] Austin, M., Belbin, L., Meyers, J., Doherty, M., and Luoto, M. (2006) Evaluation of statistical models used for predicting plant species distributions: role of artificial data and theory, *Ecological Modelling 199*, 197-216.
[54] Hirzel, A., and Guisan, A. (2002) Which is the optimal sampling strategy for habitat suitability modelling, *Ecological Modelling 157*, 331-341.
[55] Cumming, G. S. (2000) Using between-model comparisons to fine-tune linear models of species ranges, *Journal of Biogeography 27*, 441-455.
[56] Yackulic, C. B., Chandler, R., Zipkin, E. F., Royle, J. A., Nichols, J. D., Campbell Grant, E. H., and Veran, S. (2012) Presence-only modelling using MAXENT: when can we trust the inferences?, *Methods in Ecology and Evolution*.
[57] Royle, J. A., Nichols, J. D., and Kéry, M. (2005) Modelling occurrence and abundance of species when detection is imperfect, *Oikos 110*, 353-359.

[58] Ward, G., Hastie, T., Barry, S., Elith, J., and Leathwick, J. R. (2009) Presence-Only Data and the EM Algorithm, *Biometrics 65*, 554-563.
[59] Fertig, W. (2011) Strategies for plant conservation in Wyoming: Distributional modeling, gap analysis, and identifying species at risk, In *Department of Botany*, University of Wyoming, Laramie, Wyoming.
[60] Fertig, W. (2000) Status of blowout penstemon (Penstemon haydenii) in Wyoming, *Report prepared for the Wyoming Cooperative Fish and Wildlife Research Unit, US Fish and Wildlife Service, and Wyoming Game and Fish Department by the Wyoming Natural Diversity Database, Laramie, Wyoming*.
[61] Heidel, B. (2012) Status of Penstemon haydenii (blowout penstemon) in Wyoming. Unpublished report prepared for the Bureau of Land Management Rawlins and Rock Springs Field Offices by the Wyoming Natural Diversity Database, University of Wyoming, Laramie, WY.
[62] Hastie, T., Tibshirani, R., and Friedman, J. (2009) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Second ed., Springer, New York, New York, United States of America.