RESTORATION ECOLOGY
The Journal of the Society for Ecological Restoration

SER SOCIETY FOR ECOLOGICAL RESTORATION

RESEARCH ARTICLE

Cannot see the random forest for the decision trees: selecting predictive models for restoration ecology

David M. Barnard[1], Matthew J. Germino[1,2], David S. Pilliod[1], Robert S. Arkle[1], Cara Applestein[1], Bill E. Davidson[1], Matthew R. Fisk[1]

Improving predictions of restoration outcomes is increasingly important to resource managers for accountability and adaptive management, yet there is limited guidance for selecting a predictive model from the multitude available. The goal of this article was to identify an optimal predictive framework for restoration ecology using 11 modeling frameworks (including machine learning, inferential, and ensemble approaches) and three data groups (field data, geographic data [GIS], and a combination thereof). We test this approach with a dataset from a large postfire sagebrush reestablishment project in the Great Basin, U.S.A. Predictive power varied among models and data groups, ranging from 58% to 79% accuracy. Finer-scale field data generally had the greatest predictive power, although GIS data were present in the best models overall. An ensemble prediction computed from the 10 models parameterized to field data was well above average for accuracy but was outperformed by others that prioritized model parsimony by selecting predictor variables based on rankings of their importance among all candidate models. The variation in predictive power among a suite of modeling frameworks underscores the importance of a model comparison and refinement approach that evaluates multiple models and data groups, and selects variables based on their contribution to predictive power. The enhanced understanding of factors influencing restoration outcomes accomplished by this framework has the potential to aid the adaptive management process for improving future restoration outcomes.

Key words: ecological prediction, ensemble modeling, machine learning, model comparison, postfire restoration, predictive framework

Implications for Practice

- The framework presented herein will assist practitioners in selecting and comparing among the numerous models available for predicting restoration treatment outcomes.
- The variability we found among models, and the ability to refine accuracy through repeated testing, implies that multimodel comparisons are the best approach to optimizing predictive accuracy for restoration ecology projects.
- Optimizing predictive power for restoration ecology models will aid resource managers to locate treatments in areas with the greatest chance of success, thereby maximizing efficiency and resources.

Introduction

Prediction, as a means of demonstrating scientific understanding, has long been heralded as the ultimate goal of ecology (Peters 1991; Brudvig 2011; Houlahan et al. 2017). However, accurate predictions have proven troublesome due to the inherent complexity of ecological systems and the volume of data needed to adequately represent pattern–process relationships (Brudvig et al. 2017; Houlahan et al. 2017; Applestein et al. 2018). Computationally rigorous models (e.g. machine learning algorithms [MLAs]) excel at extracting patterns from complex datasets and are becoming more common in ecological studies (Peters et al. 2014), but the number of MLAs germane to any specific research problem outpace the guidance available for selecting from, testing, and refining candidate models. There is hence a need to clarify the model selection process, especially in relation to restoration ecology projects where developing restoration strategies, evaluating restoration outcomes, and applying adaptive management principles can be confounded by intricacies in the data such as idiosyncratic outcomes and environmental variability.

Developing predictive models typically involves an unconstrained search of the dataset for patterns that explain variation in the dependent variable. The traditional inferential approach to hypothesis testing has focused on using central tendencies and linear relationships in the data to assess meaningful patterns. However, ecological data are often complex and characterized

Author contributions: DB, MG conceived the project; MG obtained funding, and contributed to all work phases; RA, MF, CA, BD designed the sampling plan; MF, CA, BD led the sampling and data organization; DB performed modeling and analyses; DP contributed to project development; DB led and all coauthors contributed to writing.

[1]U.S. Geological Survey, Forest and Rangeland Ecosystem Science Center, 970 S. Lusk Street, Boise, ID 83706, U.S.A.
[2]Address correspondence to M.J. Germino, email mgermino@usgs.gov

doi: 10.1111/rec.12938
Supporting information at:
http://onlinelibrary.wiley.com/doi/10.1111/rec.12938/suppinfo

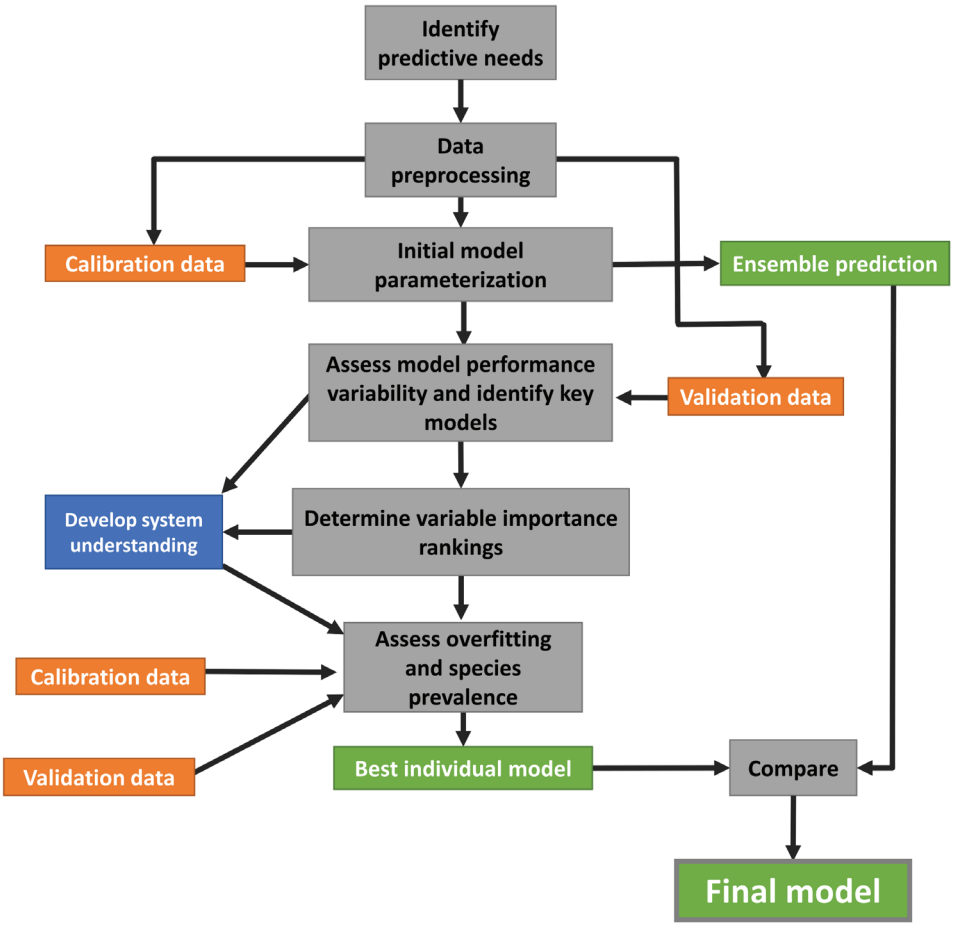

Figure 1. Conceptual diagram of the model comparison framework used in this study. Gray boxes indicate modeling processes, orange boxes indicate data inputs, blue boxes indicate knowledge outputs, and green boxes indicate developed models or predictions. Note that class imbalance assessments are only required on classification models, not regression models.

by latent patterns, nonlinear relationships, and higher-order interactions that can be more difficult to identify using inferential approaches compared to MLAs (Cutler et al. 2007; Olden et al. 2008). Notwithstanding, differences in mathematical formulation among MLAs can lead to inconsistent model performance given a specific dataset (Olden et al. 2008; Kampichler et al. 2010; Witten et al. 2016), suggesting that comparisons are needed to select the model with the greatest predictive accuracy. Alternatively, model ensembles can be developed to "smooth out" predictions from multiple models (Buisson et al. 2010; Elith et al. 2011), but this approach is rarely compared to individual models that have been refined and thus ensemble model benefits are typically assumed rather than tested.

The most objective assessment of predictive power is to test a model on validation data independent of those used for model parameterization, but such tests are rarely reported (Houlahan et al. 2017). Instead, it is more common to estimate predictive accuracy by cross-validating a model on random subsamples of the parameterization dataset (Arlot & Celisse 2010). This manner of out-of-sample testing lacks independence given the potential for inadvertent information transfer between training and testing data subsets (so-called "data-leakage"; Witten et al. 2016), and thus previously reported accuracies may be overly optimistic.

Predictive model development tends to be less limited by the often-philosophical constraints placed on inferential models built for hypothesis testing (Houlahan et al. 2017). Thus, more accurate models may be developed by using all potential predictor variables and an iterative approach of repeated model testing, reduction, and refinement based on relative rankings of variable importance. Hence, the goal of this study was to find the most predictive model of restoration outcomes using an iterative model comparison approach including MLAs, inferential tests, and model-ensemble predictions (Fig. 1, Table 1, Appendix S1, Supporting Information). We used three groups of predictor data (Table 2): (1) high sampling intensity plot-level data collected by technicians (Field-Data), (2) spatially representative and remotely sensed or interpolated geographic and landscape data that are readily available (GIS-Data), and (3) a combination thereof (All-Data) to establish a maximum predictive accuracy. We then tested this framework using a case study of postfire sagebrush occupancy following aerial seeding

Table 1. Modeling frameworks including R packages training parameters used.

Algorithm Abbreviation	*Algorithm Description*	*R Package and Reference*	*Training Parameters*	*Training Parameter Description*
CART	Classification and regression tree	"rpart" (Therneau et al. 2010)	cp	Complexity parameter
Boosted CART	Boosted classification and regression tree	"ada" (Culp et al. 2006)	Iter	Number of trees to grow
			Maxdepth	Maximum tree depth
			nu	Learning rate
Random forest	Random forest	"randomForest" (Breiman et al. 2011)	mtry	Number of predictors to randomly select for each tree
SVM	Support vector machine	"ada" (Culp et al. 2006)	degree	Polynomial degree
			Scale	Scale
			c	Cost
ANN	Artificial neural network	"nnet" (Ripley 2013)	Size	Number of hidden units
			Decay	Weight decay
			bag	Bagging iterations
Boosted logistic	Boosted logistic regression	"caTools" (Tuszynski 2012)	nIter	Number of boosting iterations
GLM	Generalized linear logistic regression with least absolute shrinkage and selection operator (LASSO)	"glmnet" (Friedman et al. 2016)	alpha	Mixing percentage
			lambda	Regularization parameter
Bayesian GLM	Bayesian generalized linear model	"arm" (Gelman et al. 2009)	–	
Naïve Bayes	Naïve Bayes classifier	"e1071" (Dimitriadou et al. 2009)	fL	Laplace correction
			Usekernal	Distribution type
			Adjust	Bandwidth adjustment
k-NN	*k*-nearest neighbor	"kknn" (Schliep et al. 2010)	Kmax	Maximum number of neighbors
			Distance	Distance
			Kernal	Kernal

treatments in the Great Basin, U.S.A. We selected sagebrush occupancy, a coarse metric for applied ecological outcomes, as our response variable due to the functional importance of sagebrush in the study ecosystem and on the premise that occupancy may be more predictable than other recovery metrics such as canopy cover or relative abundance (Brudvig et al. 2017; Laughlin et al. 2017). In the process of building our predictive models, we also explore variable importance rankings, model predictive ability on data independent from that used for model development, and agreement in model predictions along gradients in key landscape characteristics to improve the ecological understanding of our study system.

Methods

Site Description and Data Collection

Observations of sagebrush occupancy, vegetation, and soil surface characteristics (i.e. Field-Data) were collected at 2,171 plots in the year following the 2015 Soda fire, which burned over 113,000 ha in southwestern Idaho and southeastern Oregon (Fig. S1). During the winter and spring of 2016, >75,000 ha of the burn area were treated with three different applications of aerial sagebrush seed ranging from 0.89 to 1.97 lb per ha of pure live seed (PLS). Sagebrush seed mixes included Wyoming big sagebrush (*Artemesia tridentata* ssp. *wyomingensis*), basin big sagebrush (*Artemesia tridentata* ssp. *tridentata*), and low sagebrush (*Artemesia arbuscula*). In addition, 11,122 ha were also treated with a preemergent herbicide (Imazapic; applied to control annual grasses and some broad leaf weeds) and another 7,241 ha were drill-seeded with perennial bunchgrass seed mixes which included primarily blue bunch wheatgrass (*Pseudoroegneria spicata*), crested wheatgrass (*Agropyron cristatum*), and Sandberg's bluegrass (*Poa secunda*). The polygons for each treatment overlapped in some regions, and thus individual sampling points may have had any combination of the three treatments applied.

The 2,171 plots used in this study are from two independent datasets: one collected under a Bureau of Land Management Emergency Stabilization and Rehabilitation (BLM-ESR) project and the second collected using similar methods for a Joint Fire Sciences Program (JFSP) project. Plot locations for the BLM-ESR project were selected using a stratified-random approach but were discarded if slope angle was >40°, they were within 400 m of a water source, or if cobbles, rocks, or trails made up >20% of an area within an 18 m radius of the plot center. Plots in the JFSP study were selected at defined distances from a road (100, 200, and 300 m). To improve detection

Table 2. Potential predictor variables used for modeling. In the treatment predictors (i.e. top three rows), a zero indicates the treatment was not applied and a one indicates the treatment was applied. A Y (yes) or N (no) in the data group columns identifies whether the data represent field measurements (Field-Data) or if they were collected from spatial geographic data (GIS), readily available online. The All-Data subset used in this study included all variables.

				Data Group	
Parameter	*Data Class*	*Data Range*	*Source*	*Field*	*GIS*
Sagebrush seed applied	Categorical	0–1	BLM polygons	Y	Y
Herbicide applied	Categorical	0–1	BLM polygons	Y	Y
Grass seed applied	Categorical	0–1	BLM polygons	Y	Y
Exotic annual grass cover (%)	Numerical	0–100	Measured in field	Y	N
Perennial bunch grass cover (%)	Numerical	0–100	Measured in field	Y	N
Fertile island cover (%)	Numerical	0–100	Measured in field	Y	N
Pedoderm class	Categorical	–	Measured in field	Y	N
Mean annual precipitation (mm)	Numerical	234–530	PRISM	N	Y
Plot slope aspect (degrees)	Numerical	0.1–360	Digital elevation model	N	Y
Plot slope angle (degrees)	Numerical	0.2–37.8	Digital elevation model	N	Y
Heatload (unitless)	Numerical	0.35–0.99	Digital elevation model	N	Y
Topographic wetness index	Numerical	5.13–17.31	Digital elevation model	N	Y
Site potential	Categorical	–	LANDFIRE	N	Y
LANDFIRE vegetation cover	Categorical	–	LANDFIRE	N	Y
Burn history (no. fires)	Numerical	0–2	LANDFIRE	N	Y
Soil taxonomic class	Categorical	–	SSURGO	N	Y
Soil erodibility (unitless)	Numerical	0–0.44	SSURGO	N	Y
Depth to bedrock (cm)	Numerical	0–99	SSURGO	N	Y

probability, sagebrush occupancy was determined by surveying circular areas of increasing radius (5.5, 9, 13, and 18 m) from the plot center-point. If no sagebrush were observed within an 18 m radius, the plot was recorded as unoccupied.

All data used for modeling are summarized in Table 1, and data summaries are presented in Table S1 and Figure S2. For Field-Data, we conducted an unguided ocular assessment of exotic annual grass, perennial bunch grass, and soil fertile island cover, and recorded the soil "pedoderm" classification of each plot. Fertile islands of soil are "legacy" imprints of prefire shrub crowns that have enhanced biogeochemical and hydrologic soil properties which persist after a fire (Hoover & Germino 2012). They are visually identified by substantially darker surface color. Soil pedoderm classification is based on the biological and physical features at the soil surface (Burkett et al. 2011).

For GIS-Data we used a digital elevation map to determine slope angle and aspect at each plot location. These parameters were then used to calculate heatload (McCune et al. 2002) and topographic wetness index (TWI; Beven & Kirkby 1979). A 30-year mean of annual precipitation was determined from gridded PRISM data (800 m pixels; PRISM Climate, PRISM-Climate-Group 2004). We collected prefire vegetation classification and ecological site potential from the LANDFIRE database (Rollins 2009) and we collected spatial soil mapping data from the soil geographic database (SSURGO; Soil Survey Staff 2017).

Supplementing field observations with spatial or remotely sensed data (i.e. GIS-Data in this study) is a standard practice in ecological studies, but there is potential for error and misclassification to be introduced when GIS data are disseminated at the plot level (Dietze 2017), and little is known about the trade-offs between the increased information of including GIS at the cost of reduced resolution, especially in the context of multimodel comparisons. While practitioners rarely rely solely on GIS-Data, in lieu of field sampling, the spatial intensity are often limited by budgetary constraints, and it is not always clear if landscape gradients are characterized with adequate resolution (Applestein et al. 2018).

Model Comparison Framework

Data Preparation and Preprocessing

Data were preprocessed by filtering out missing values and predictors with zero or near-zero variance. To determine near-zero variance predictors, we used two metrics from Kuhn (2008): (1) percent of unique values, defined as the number of unique values divided by the total number of samples and (2) the frequency ratio, defined as the frequency of the most prevalent value over the second most frequent value. If the percent of unique values was <10% and the frequency ratio was >19, then that predictor was removed (Kuhn & Johnson 2013). Correlations among predictors were calculated and predictors were to be removed if the pair-wise Pearson's correlation coefficient (r) was >0.75 (Kuhn & Johnson 2013), although no colinear predictors were observed. Data were then standardized and centered to have a mean of zero and variance equal to one. Training data (BLM-ESR) and testing data (JFSP) were preprocessed separately to avoid data leakage (Witten et al. 2016). All modeling and model evaluations were done using the "caret" package in R (Kuhn 2008).

Parameterize and Tune Initial Models

We calibrated 30 models (10 models each for the Field-Data, GIS-Data, and All-Data groups) using 10-fold cross-validation

repeated three times on different random subsets of the calibration dataset. A list of the different models and their estimated parameters are listed in Table 2, and a full description of each model can be found in Appendix S1. A model-averaged ensemble prediction was then determined for each of the three data groups by determining the mode of predictions from the 10 parameterized models at each plot in the calibration dataset. The code for calibrating and testing model performance are available via Appendix S2.

Assess Model Performance Rankings and Identify Key Models and Data Groups

We ranked the individual models' predictive power by validating each of the 30 models, and the three model-averaged ensemble predictions, on the independent JFSP test dataset (Fig. 2). We used Cohen's kappa (henceforth "kappa") and balanced accuracy as model performance metrics because both are integrative measures that consider multiple dimensions of model performance instead of just the number of correct classifications (i.e. accuracy). Balanced accuracy is calculated as the mean of sensitivity (true positive rate) and specificity (true false rate) (Kuhn 2008). We note that balanced accuracy is a "penalized" metric; typically resulting in lower values of accuracy than those produced by the standard accuracy calculation, which only reports the total number of correct predictions. Kappa accounts for chance effects and reports a measure of the proportion of all possible classifiers that are predicted accurately (Cohen 1960).

To reduce the total number of models subjected to in-depth evaluation and to limit processing time, we identified two groups of "top" models. Models placed in the first group retained their framework and data-group identities (henceforth the "framework × data group" group) and were in the top 20% of models based on their distance from the *y*-intercept along the kappa and balanced accuracy regression line. This assumes a higher combined kappa and balanced accuracy score indicates a better model. The second group contained the three best model frameworks overall, regardless of data group, based on their combined kappa and balanced accuracy scores (henceforth the "best framework" group).

Variable Importance, Class Imbalance, and Final Model Selection

We ranked variable importance by permuting a mean decrease in classification accuracy per variable during the repeated 10-fold cross-validation procedure used for model parameterization. Variable importance was assessed using the All-Data group from the initial model runs to avoid confounding results across the different groups of predictor variables.

Imbalanced classes (i.e. a greater number of "presences" versus "absences" or vice versa) can affect model parameterization and accuracy and should be considered during model development (Kotsiantis et al. 2007; Kampichler et al. 2010). In this study, sagebrush occupancy classes were indeed imbalanced, with only 37% of plots occupied. Consequently, we randomly subset the framework × data group to produce five new calibration datasets, each with 0.5 sagebrush occupancy rate, to assess class imbalance effects on predictive accuracy. We compared model results derived from these datasets to those obtained from five other random subsets that retained the imbalanced classes.

To test the influence of variable reduction on the fit of models in the best framework group, we built a base model for each of the candidate models that included only the predictor with the highest importance (mean annual precipitation, i.e. a univariate model). Variables were added one at a time according to their relative importance ranking, models were compared, and the best performing model was selected based on kappa and balanced accuracy.

Results

Metrics of model performance varied substantially across modeling frameworks and data groups (Fig. 2). Balanced accuracy ranged from 0.59 for the least predictive model (CART All-Data) up to 0.74 for the most predictive model (random forest with Field-Data) and 0.75 for the model-averaged ensemble prediction from Field-Data. Kappa ranged from 0.14 for GIS CART, to 0.44 for boosted logistic regression parameterized to All-Data, and up to 0.49 for the ensemble prediction to Field-Data. The frequency-density distributions for balanced accuracy and kappa were similar among data groups, with GIS-Data being more skewed away from lower values than the All-Data group, and the Field-Data distribution was dominated by higher values (Fig. 2, upper panels).

Field-Data had the highest mean balanced accuracy (0.69), GIS-Data the lowest (0.64), and the All-Data group had a mean accuracy of 0.67. Kappa was greatest for the Field-Data subset of models (0.35), less for models parameterized with All-Data (0.31), and least for models parameterized with GIS-Data (0.25). Averaged across All-Data groups, random forest, boosted CART, and SVM were the highest performing individual models (Fig. 2, lower panels), and the model-averaged ensemble was the fourth most predictive framework. Arranging the "framework × data groups" based on their balanced accuracy and kappa scores along a regression line showed a relative ranking of overall "best" model performance (Fig. 3). The top performing 20% of models are identified within the circle.

Predictive accuracy was overestimated by up to 23% (average of 16.6% across all models and data groups) if assessed only on within-sample subsets of data (i.e. 10-fold cross-validation) versus validation on the independent dataset (Fig. 4). Overestimated accuracy was greatest in the GIS-Data and All-Data groups (19% and 18.4%, respectively) versus the Field-Data group (3.7%). The range of overestimations was spread evenly among models, with SVM, random forest, and CART models having the greatest overestimates, and Boosted logistic, naïve Bayes, and Bayesian generalized linear models (GLMs) having the least.

The rank of variable importance was consistent across models (Fig. 5A). Variation in relative variable importance

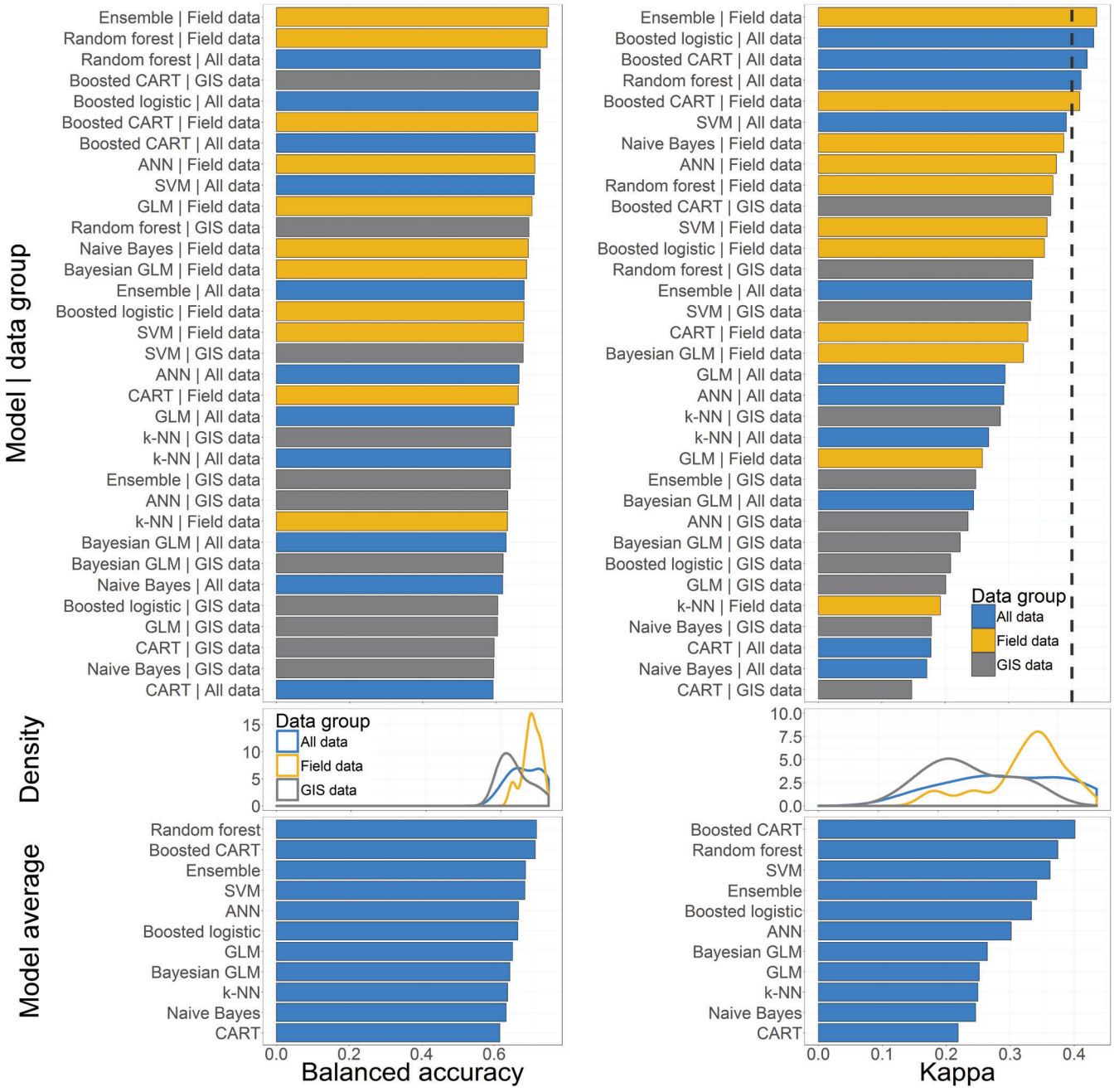

Figure 2. Comparison of model metrics (balanced accuracy and kappa, left and right columns, respectively, for predicting sagebrush occupancy) across modeling frameworks and data groups. Upper panels represent the predictive performance of individual model, ensembles, and data group parameterizations. Middle panels are a probability density function showing the distribution of balanced accuracy and kappa for the three data groups averaged across all modeling frameworks. The lower two panels show the mean balanced accuracy and kappa among data groups for the 10 modeling frameworks tested in this study. The *x*-axis on the lowest panel is applicable to the middle and upper panels as well Dashed line in upper-right panel represents a 0.4 threshold for kappa.

in the All-Data subset was small except for the five multilevel categorical predictors (burn history, LANDFIRE existing vegetation cover, soil taxonomic class, LANDFIRE site potential, and pedoderm class) due to variable effects among factor levels. Mean scaled variable importance of the three best All-Data models (i.e. random forest, SVM, and boosted CART; represented by smaller dark blue circles) were within the variable importance error range determined from all models.

Reduced predictive performance due to model overfitting from unconstrained variable selection was apparent in the candidate models (Fig. 5B). We found that predictive power increased, as up to six predictor variables were added, but accuracy generally decreased when more than six predictors were

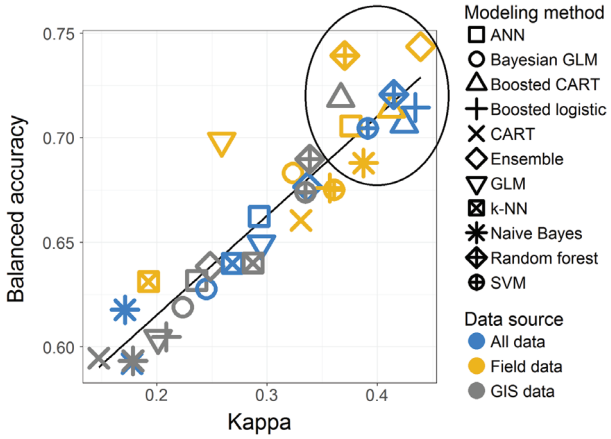

Figure 3. Validation balanced accuracy versus kappa for the 10 models (black/white symbols) of sagebrush occupancy tested, and the ensemble averages and three data groups (color symbols). Balanced accuracy is the mean of sensitivity and specificity whereas kappa accounts for chance effects and reports a measure of the proportion of all possible classifiers that are predicted accurately. The black line represents a least squares regression line. The models and data groups inside of the black circle are the top models in the model × data subgroup.

included. Boosted CART, SVM, and random forest models had the highest balanced accuracy and kappa (0.79 and 0.54, respectively for all models) when parameterized with just three predictor variables: mean annual precipitation, sagebrush seeding treatment, and fertile island cover. Hence, we concluded that these three models were the best and final models. We found no significant effect of class imbalance on predictive accuracy, sensitivity, and specificity by randomly down-sampling the majority class and comparing it to down-sampled calibration sets (for equal sample sizes) that retained class imbalance ($p > 0.15$ for all comparisons).

The predictability of sagebrush occupancy, in terms of the 10 different models agreeing on predictions of presence or absence, varied across gradients in key landscape characteristics (Fig. 6). This was determined by calculating a mean value of the numeric prediction (0,1) from the 30 tested models (i.e. a value of 0 indicates that all models agree on sagebrush absence, and a value of one indicates all models agree on sagebrush presence) and regressing those values against key landscape characteristics. All model predictions agreed that sagebrush *would not* be present at 33% of plots in the independent testing dataset, whereas all model predictions agreed that sagebrush *would* be present at only 1% of the plots. Mean annual precipitation, exotic annual grass cover, and fertile island cover all significantly influenced model agreement ($p < 0.001$ for all). Agreement among model predictions of no seedling presence (i.e. complete restoration seeding failure) was more likely to occur at lower mean annual precipitation, higher exotic annual grass cover, and low fertile island cover. Conversely, high mean annual precipitation, low exotic annual grass, and high intermediate to high fertile island cover were more likely to result in models that agreed on sagebrush presence (i.e. some restoration seeding success).

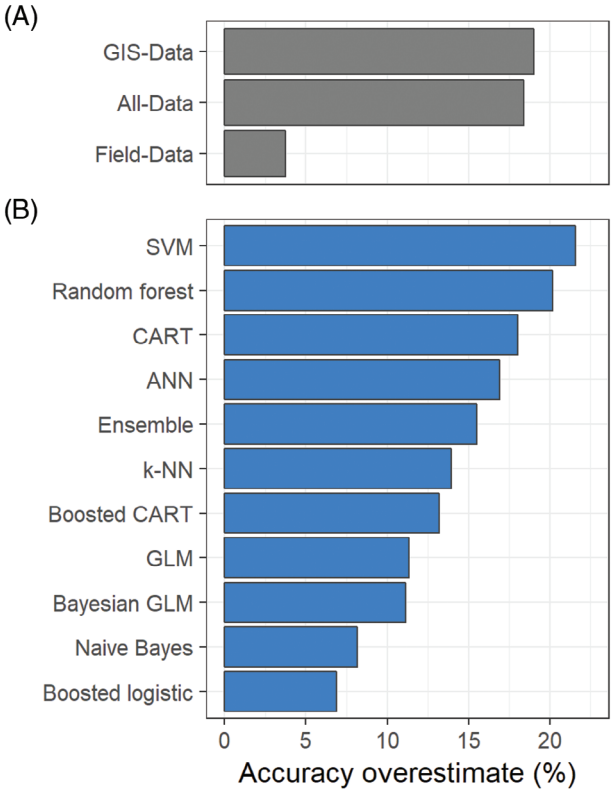

Figure 4. Overestimate of predictive accuracy when models are tested only on out-of-sample subsets (i.e. cross-validation) versus when tested on independent data for the three data groups (A), and the 10 tested models and ensemble prediction (B).

Discussion

Variability Among Modeling Frameworks and Predictor Groups

Our goal was to introduce a model comparison and refinement approach to optimize model predictive ability to help guide the field of restoration ecology as mathematical and statistical modeling are rapidly adopted into this discipline. We illustrated this approach with a case study of postfire sagebrush rehabilitation treatments across a large, heterogeneous area where issues of environmental variability and scale capture the ecological complexity typical of restoration projects and complicate understanding of restoration outcomes. The generalized nature of this framework can be readily adapted to other systems and study designs based on investigator goals and dataset characteristics and thus in this section we explore our findings in a more generalized context. In this regard, we note that many of the models can predict continuous outcomes in addition to the binary categorical outcomes examined in this study (e.g. boosted CART, random forest, artificial neural networks (ANN)). In the process of testing and down-sampling multiple models, we found the predictive accuracy of our models to vary substantially, and the best models to be those with the fewest variables. These results underscore the importance of prioritizing model simplicity even with models that are otherwise understood to handle a large number of predictors well (i.e. MLAs; discussed below).

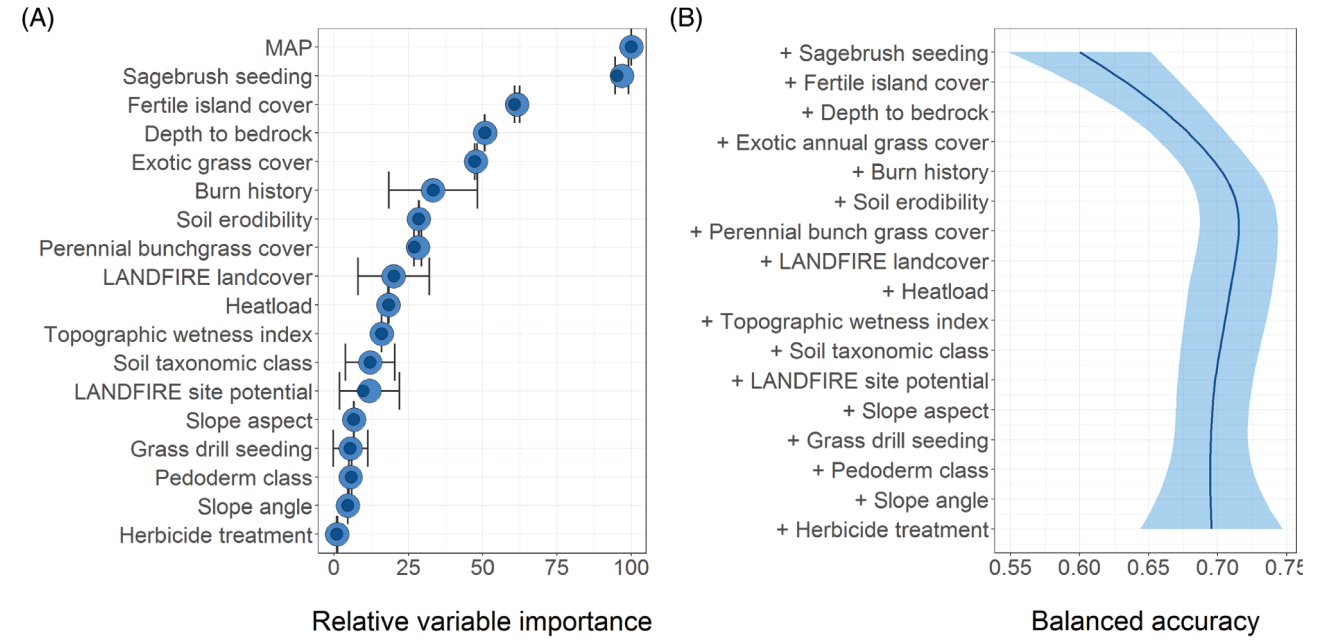

Figure 5. Mean and standard deviation of relative variable importance for predicting sagebrush occupancy across all models in Figure 1 (A; large blue circles) and the mean importance of each variable in the two best individual models (random forest and boosted CART; smaller dark blue circles), and (B) the change in the mean of balanced accuracy among models as parameters were added sequentially to the models in order of their scaled variable importance.

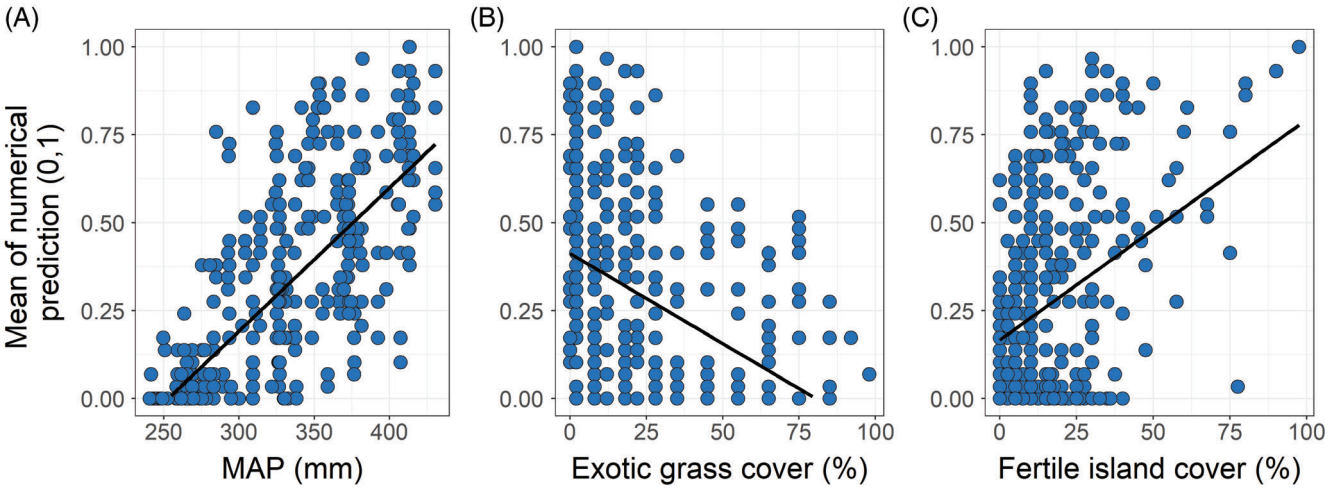

Figure 6. Mean of the numerical model predictions of sagebrush occupancy (0 for absence, 1 for presence) at the test data plots as a function of mean annual precipitation (A), exotic annual grass cover (B), and (C) fertile island cover. A mean of binary prediction of 0 indicates all models agree on sagebrush absence, and a model agreement of 1 indicates all models agree on presence. Values between 0 and 1 are equal to the number of models predicting presence divided by the number of models predicting absence.

Studies that tested the predictive power of different data groups are lacking in the literature. However, we found models parameterized using only Field-Data generally had the highest accuracy, although GIS-Data predictor variables also contributed information to the best models. Despite these results being specific to our system, the variation among data groups has interesting implications for practitioners looking to maximize predictive power across a range of spatial scales and in different ecosystems. For example, the coarse resolution of GIS-Data may be sufficient (or potentially superior) for research projects focused at eco-regional scales where coarser climate and edaphic variation may overwhelm microscale effects. However, for practitioners working at smaller scales, in individual projects, or in heterogeneous landscapes like this study, using field measurements as predictor variables was clearly important for maximizing predictive accuracy. Similarly, some models performed better with coarser GIS-Data than Field-Data (Boosted CART, k-NN), or similarly to those with Field-Data (random forest) and may make ideal candidates for future studies with limited resources to invest into field sampling.

We observed substantial variation in predictive power among models and data groups that we attribute to differences in the mathematical formulation. These differences, such as how variables are selected or a tendency to overfit models, can affect model accuracy. While an assessment of the intricate differences among models and how they may affect predictive accuracy is beyond the scope of this article, for further information we direct the reader to Appendix S1, seminal publications on MLAs and model comparisons (e.g. De'ath & Fabricius 2000; Kotsiantis et al. 2007; Kampichler et al. 2010), and the documentation for the "caret" package (http://topepo.github.io/caret/index.ht). Nonetheless, we found CART to be the least predictive model overall which is contrary to Germino et al. (2018), an inferential study using this dataset, that reported 71% predictive accuracy for CART in predicting sagebrush reestablishment (versus 59–66% in this study). There are two potential reasons for this disparity; first, in the current study, the parameter used to "prune" initial decision trees and reduce model overfitting was randomly selected from a range of potential values, whereas the previous study used a mathematical approach by finding the parameter value with the greatest cross-validated accuracy. Second, the previous study used a subset of predictors that had previously been identified as significant in a linear model to parameterize the CART model, whereas in this study we did not assign any variable importance during initial CART parameterization and assessment. Variable importance can be assigned using most CART model algorithms, and may be essential given the "greedy" approach to partitioning predictor variable homogeneity (i.e. partitioning maximum homogeneity at each step may occlude some interactions and latent relationships; De'ath & Fabricius 2000). However, given the poor initial performance in this study, CART was not a candidate for further model refinement, but considering variable importance should be included in future studies.

On the other hand, random forest and boosted CART, which are derivatives of the CART formulation, produced highly accurate models by parameterizing a suite of models and reporting an average prediction. Additionally, random forest, boosted CART, and SVM have been shown to perform better on high-dimension datasets (Huang et al. 2002; Oommen et al. 2008). Conversely, in agreement with our findings, naïve Bayes and k-NN (second and third least predictive overall) have lower performance on high-dimension datasets due to a sensitivity to irrelevant predictor variables (Kotsiantis et al. 2007).

Although MLAs (e.g. random forest, CART, ANN) can produce highly predictive models, shortcomings may limit their adoption as research and planning tools. Machine learning models can produce overly complex models compared to maximum likelihood procedures where parsimony was prioritized (Halvorsen 2013). Moreover, the "black-box" formulation of MLAs limits interpretation and model complexity can quickly exceed human cognitive abilities (i.e. the "curse of dimensionality"; Bellman 1957). While certain graphical techniques (e.g. partial dependence plots) can aid visualization of simple relationships, the inherent difficulty of interpreting MLAs limits their use for directly assessing ecological relationships and complicates communication of findings.

The Advantages of a Model Comparison Approach and Independent Validation

The rankings in predictive power among the models tested in this study may offer—generalized guidelines. For example, we identified that a suite of candidate models should be tested to identify which will be the most appropriate. We found random forest and boosted CART to have the greatest overall predictive accuracy, both have a record of exceptional accuracy in the literature (e.g. Prasad et al. 2006; Cutler et al. 2007; Elith et al. 2008), and should be considered first in future studies. However, other models such as SVM and boosted logistic models have similar records and may prove to be more accurate when used on coarser datasets or with a large number of potentially irrelevant predictors (Huang et al. 2002; Kotsiantis et al. 2007; Oommen et al. 2008). We acknowledge that a large model comparison procedure, such as that reported in this study, may be untenable for certain projects. We thus encourage future researchers to test as many models as is feasible and to understand that acceptable thresholds for what determines a "good" model will be specific to a given dataset.

A model comparison approach may also identify situations where model-averaged ensemble predictions can be advantageous or not. Model averaging is often used to homogenize across individual model idiosyncrasies, supposedly with greater transferable predictive power (e.g. Buisson et al. 2010; Elith et al. 2011). In this study, the ensemble model using only Field-Data was well above average in terms of predictive accuracy, but it was less predictive than several individual models that had been refined though variable reduction, suggesting that a model-averaging approach may limit predictive power in certain situations.

Predictive accuracy was markedly improved by reducing the number of predictors based on importance rankings. However, predictive accuracy was not impacted by randomly subsampling the calibration data to ensure there was an equal number of presences and absences. Interestingly, both findings are contrary to previous studies (Chawla et al. 2003; Chen et al. 2004; Cutler et al. 2007; Evans et al. 2011). The enhanced predictive power obtained by reducing the number of predictors is especially relevant because it contradicts a common assumption that many MLAs, which automatically select important predictors, are immune to model overfitting (Huang et al. 2002; Cutler et al. 2007). Instead, our findings suggest that the larger All-Data group produced less predictive models because many of the formulations may have selected irrelevant variables.

Another concern is for future studies to report overly optimistic measures of predictive accuracy by not validating models on data that are independent from those used for model calibration. Our results show that model accuracy can be considerably less when tested on independent data (up to 23%; Fig. 4), indicating that even with repeated *k*-fold cross-validation, model overfitting can lead to a reduction in predictive power.

Given that access to spatially and temporarily independent testing datasets is uncommon, we encourage practitioners and researchers to split datasets into calibration and testing sets before data preprocessing and model development. We additionally emphasize the collection of a separate validation dataset should be considered and prioritized during the study planning and design phase (Hooten & Hobbs 2015; Houlahan et al. 2017).

Model Refinement and Improving System Understanding

Predictive model goals differ from those of models developed for hypothesis testing and inference with the former being focused on *demonstrating* understanding through predictive power, whereas the latter is concerned with *developing* system understanding through hypothesis testing (Halvorsen 2012; Houlahan et al. 2017). The framework we propose in this study bridges these goals by maximizing predictive accuracy, while also developing system understanding through variable importance rankings and assessing model agreement in relation to environmental gradients. In this regard, the consistency with which variable importance was ranked across models lends defensible and novel insight into study-system behavior. The development of similar rankings could be key in guiding researchers and land managers to prioritize measurements in future work to maximize their predictive accuracy and ecological understanding.

The degree of model agreement regarding sagebrush presence or absence varied in relation to landscape features, but model agreement of no occupancy was generally greater in areas with low mean annual precipitation, high annual grass cover, and lower fertile island cover. Specific to the findings of this study, these characteristics generally correspond to areas widely regarded as low resistance and resilience landscapes for sagebrush reestablishment after fire (Maestas et al. 2016). Thus, models widely agreeing on the absence of sagebrush in this area are not surprising. However, the variability among models in the middle and upper ranges of precipitation and the lower ranges of exotic annual grass cover underscores the importance of landscape heterogeneity in determining the predictability of restoration outcomes. In addition, future studies may consider that model selection and uncertainty could be constrained more by landscape characteristics unique to a system rather than model formulation or framework, which may also vary as response variables traverse gradients in predictability (Laughlin et al. 2017).

Acknowledgments

This project was funded by the Joint Fire Sciences Program (grant 16-1-03-13) with partial support from the Great Basin Landscape Conservation Cooperative, US Geological Survey (USGS) Fire Program, USGS/BLM SageSuccess Project, and in-kind contribution of USGS data funded by the BLM's Fire Rehabilitation Program. We thank Trevor Caughlin for helpful comments that improved this manuscript substantially and Manuella Huso for modeling discussion. Field-Data were collected by J. Aaronson, J. Albertson, A. Blomberg, V. Callahan, R. Donaldson, E. Donohoe, B. Fischer, L. Hunsaker, C. Kuzis, A. Lague, T. Laird, J. Moran-Davidson, H. Person-Nadal, R. Robinson, C. Thomas, J. Titcomb, M. Toure, M. V. M. Davidson, P. Weinberg, and W. Yates. Partial support for planning and analysis of the data was provided by the BLM/USGS SageSuccess Project. Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

LITERATURE CITED

Applestein C, Germino MJ, Pilliod DS, Fisk MR, Arkle RS (2018) Appropriate sample sizes for monitoring burned pastures in sagebrush steppe: how many plots are enough, and can one size fit all? Rangeland Ecology & Management 71:721–726

Arlot S, Celisse A (2010) A survey of cross-validation procedures for model selection. Statistics Surveys 4:40–79

Bellman R (1957) A Markovian decision process. Journal of Mathematics and Mechanics 6:679–684

Beven K, Kirkby MJ (1979) A physically based, variable contributing area model of basin hydrology. Hydrological Sciences Journal 24:43–69

Breiman L, Cutler A, Liaw A, Wiener M (2011) Package 'randomForest' software. http://stat-www.berkeley.edu/users/breiman/RandomForests (accessed Nov 2017)

Brudvig LA (2011) The restoration of biodiversity: where has research been and where does it need to go? American Journal of Botany 98:549–558

Brudvig LA, Barak RS, Bauer JT, Caughlin TT, Laughlin DC, Larios L, Matthews JW, Stuble KL, Turley NE, Zirbel CR (2017) Interpreting variation to advance predictive restoration science. Journal of Applied Ecology 54:1018–1027

Buisson L, Thuiller W, Casajus N, Lek S, Grenouillet G (2010) Uncertainty in ensemble forecasting of species distribution. Global Change Biology 16:1145–1157

Burkett LM, Bestelmeyer BT, Tugel AJ (2011) A field guide to pedoderm and pattern classes. National Resource Conservation Service, Las Cruces, NM, U.S.A.

Chawla N, Lazarevic A, Hall L, Bowyer K (2003) SMOTEBoost: improving prediction of the minority class in boosting. Pages 107–119. In: Lavrac, N. Gamberger, D, Blockeel, H., Todorovski, L (eds) Proceedings of the Principles of Knowledge Discovery in Databases: PKDD 2003. Springer, Germany

Chen C, Liaw A, Breiman L (2004) Using random forest to learn imbalanced data. University of California, Berkeley

Cohen J (1960) A coefficient of agreement for nominal scales. Educational and Psychological Measurement 20:37–46

Culp M, Johnson K, Michailidis G (2006) ada: an r package for stochastic boosting. Journal of Statistical Software 17:9–36

Cutler DR, Edwards TC, Beard KH, Cutler A, Hess KT, Gibson J, Lawler JJ (2007) Random forests for classification in ecology. Ecology 88:2783–2792

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

Dietze MC (2017) Prediction in ecology: a first-principles framework. Ecological Applications 27:2048–2060

Dimitriadou E, Hornik K, Leisch F, Meyer D, Weingessel A, Leisch MF (2009) Package "e1071". R Software package. http://cran.rproject.org/web/packages/e1071/index. html (accessed Nov 2017)

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

Elith J, Phillips SJ, Hastie T, Dudík M, Chee YE, Yates CJ (2011) A statistical explanation of MaxEnt for ecologists. Diversity and Distributions 17:43–57

Evans JS, Murphy MA, Holden ZA, Cushman SA (2011) Modeling species distribution and change using random forest. Pages 139–159. In: Drew CA, Wiersma Y, Huettmann F (eds) Predictive species and habitat modeling in landscape ecology. Springer, Berlin, Germany

Friedman J, Hastie T, Simon N, Tibshirani R (2016) glmnet: Lasso and elastic-net regularized generalized linear models. R package version 1.9–5. R Foundation for Statistical Computing, Vienna, Austria

Gelman A, Su Y-S, Yajima M, Hill J, Pittau MG, Kerman J, Zheng T, Dorie V (2009) arm: data analysis using regression and multilevel/hierarchical models. R package version 9.01. R Foundation for Statistical Computing, Vienna, Austria

Germino MJ, Barnard DM, Davidson BE, Arkle RS, Pilliod DS, Fisk MR, Applestein C (2018) Thresholds and hotspots for shrub restoration following a heterogeneous megafire. Landscape Ecology 33:1177–1194

Halvorsen R (2012) A gradient analytic perspective on distribution modelling. Sommerfeltia 35:1–165

Halvorsen R (2013) A strict maximum likelihood explanation of MaxEnt, and some implications for distribution modelling. Sommerfeltia 36:1–132

Hooten MB, Hobbs N (2015) A guide to Bayesian model selection for ecologists. Ecological Monographs 85:3–28

Hoover AN, Germino MJ (2012) A common-garden study of resource-island effects on a native and an exotic, annual grass after fire. Rangeland Ecology & Management 65:160–170

Houlahan JE, Mckinney ST, Anderson TM, Mcgill BJ (2017) The priority of prediction in ecological understanding. Oikos 126:1–7

Huang C, Davis L, Townshend J (2002) An assessment of support vector machines for land cover classification. International Journal of Remote Sensing 23:725–749

Kampichler C, Wieland R, Calmé S, Weissenberger H, Arriaga-Weiss S (2010) Classification in conservation biology: a comparison of five machine-learning methods. Ecological Informatics 5:441–450

Kotsiantis SB, Zaharakis I, Pintelas P (2007) Supervised machine learning: a review of classification techniques. Informatics 31:249–268

Kuhn M (2008) Caret package. Journal of Statistical Software 28:1–26

Kuhn M, Johnson K (2013) Applied predictive modeling. Springer, Berlin, Germany

Laughlin DC, Strahan RT, Moore MM, Fulé PZ, Huffman DW, Covington WW (2017) The hierarchy of predictability in ecological restoration: are vegetation structure and functional diversity more predictable than community composition? Journal of Applied Ecology 54:1058–1069

Maestas JD, Campbell SB, Chambers JC, Pellant M, Miller RF (2016) Tapping soil survey information for rapid assessment of sagebrush ecosystem resilience and resistance. Rangelands 38:120–128

McCune B, Keon D, Marrs R (2002) Equations for potential annual direct incident radiation and heat load. Journal of Vegetation Science 13:603–606

Olden JD, Lawler JJ, Poff NL (2008) Machine learning methods without tears: a primer for ecologists. The Quarterly Review of Biology 83:171–193

Oommen T, Misra D, Twarakavi NK, Prakash A, Sahoo B, Bandopadhyay S (2008) An objective analysis of support vector machine based classification for remote sensing. Mathematical Geosciences 40:409–424

Peters RH (1991) A critique for ecology. Cambridge University Press, Cambridge, UK.

Peters DP, Havstad KM, Cushing J, Tweedie C, Fuentes O, Villanueva-Rosales N (2014) Harnessing the power of big data: infusing the scientific method with machine learning to transform ecology. Ecosphere 5:1–15

Prasad AM, Iverson LR, Liaw A (2006) Newer classification and regression tree techniques: bagging and random forests for ecological prediction. Ecosystems 9:181–199

Prism-Climate-Group (2004) Oregon State University. http://prism.oregonstate.edu (accessed Nov 2017)

Ripley B (2013) Feed-forward neural networks and multinomial log-linear models. nnet package, version 7.3-6. http://www.stats.ox.ac.\ignorespacesuk/pub/MASS4 (accessed Nov 2017)

Rollins MG (2009) LANDFIRE: a nationally consistent vegetation, wildland fire, and fuel assessment. International Journal of Wildland Fire 18:235–249

Schliep K, Hechenbichler K, Lizee A (2010) kknn: weighted k-nearest neighbors. R package version 1.0-8. R Foundation for Statistical Computing, Vienna, Austria

Soil Survey Staff (2017) Soil survey geographic (SSURGO) database. Natural Resources Conservation Service, United States Department of Agriculture

Therneau TM, Atkinson B, Ripley MB (2010) The rpart package. https://cran.rproject.org/web/packages/rpart/index.html (accessed Nov 2017)

Tuszynski J (2012) caTools: Tools: moving window statistics, GIF, Base64, ROC AUC, etc. R package version 116. http://CRAN.R-project.org/package=caTools (accessed 1 Apr 2014)

Witten IH, Frank E, Hall MA, Pal CJ (2016) Data mining: practical machine learning tools and techniques. Morgan Kaufmann series. Elsevier, Cambridge, MA, U.S.A.

Supporting Information

The following information may be found in the online version of this article:

Appendix S1. Expanded description of models included in the model-comparison framework and references
Appendix S2. Data set and R code used to calibrate and test model performance
Table S1. Overview of variable factor levels and those retained in analysis
Figure S1. Map of sampling locations
Figure S2. Summary of continuous data inputs for modeling

Coordinating Editor: Leighton Reid

Received: 18 September, 2018; First decision: 9 December, 2018; Revised: 17 February, 2019; Accepted: 19 February, 2019; First published online: 28 March, 2019