

Random forests and stochastic gradient boosting for predicting tree canopy cover: comparing tuning processes and model performance¹

Elizabeth A. Freeman, Gretchen G. Moisen, John W. Coulston, and Barry T. Wilson

Abstract: As part of the development of the 2011 National Land Cover Database (NLCD) tree canopy cover layer, a pilot project was launched to test the use of high-resolution photography coupled with extensive ancillary data to map the distribution of tree canopy cover over four study regions in the conterminous US. Two stochastic modeling techniques, random forests (RF) and stochastic gradient boosting (SGB), are compared. The objectives of this study were first to explore the sensitivity of RF and SGB to choices in tuning parameters and, second, to compare the performance of the two final models by assessing the importance of, and interaction between, predictor variables, the global accuracy metrics derived from an independent test set, as well as the visual quality of the resultant maps of tree canopy cover. The predictive accuracy of RF and SGB was remarkably similar on all four of our pilot regions. In all four study regions, the independent test set mean squared error (MSE) was identical to three decimal places, with the largest difference in Kansas where RF gave an MSE of 0.0113 and SGB gave an MSE of 0.0117. With correlated predictor variables, SGB had a tendency to concentrate variable importance in fewer variables, whereas RF tended to spread importance among more variables. RF is simpler to implement than SGB, as RF has fewer parameters needing tuning and also was less sensitive to these parameters. As stochastic techniques, both RF and SGB introduce a new component of uncertainty: repeated model runs will potentially result in different final predictions. We demonstrate how RF allows the production of a spatially explicit map of this stochastic uncertainty of the final model.

Key words: tree canopy cover, predictive mapping, classification and regression trees, random forest, stochastic gradient boosting.

Résumé : Dans le cadre de l'élaboration de la couche cartographique du couvert forestier de la National Land Cover Database 2011, un projet pilote a été lancé pour tester l'utilisation de la photographie haute résolution couplée à de multiples données accessoires pour cartographier la distribution du couvert forestier dans les états contigus des États-Unis. Deux techniques de modélisation stochastique, les forêts aléatoires (FA) et le « gradient boosting » aléatoire (GBA) sont comparées. Les objectifs de cette étude consistaient : premièrement à explorer la sensibilité des deux techniques face aux choix pour le réglage des paramètres; et deuxièmement à comparer la performance des deux modèles finaux en évaluant l'importance des variables prédictives et leurs interactions, les mesures d'exactitude globale dérivée d'un dispositif de test indépendant, de même que la qualité visuelle des cartes du couvert forestier qui sont produites. L'exactitude des prévisions des deux techniques était remarquablement similaire dans les quatre régions pilotes. Dans les quatre régions, l'EQM du dispositif de test indépendant était identique à trois décimales près; le plus grand écart était au Kansas où la technique FA produisait un EQM de 0,0113 tandis que la technique GBA produisait un EQM de 0,0117. Avec des variables prédictives corrélées, la technique GBA avait tendance à concentrer l'importance des variables sur moins de variables alors que la technique FA avait tendance à répartir l'importance parmi davantage de variables. La technique FA est plus simple à appliquer que la technique GBA étant donné qu'elle compte à la fois moins de paramètres qui ont besoin de réglage et qu'elle est aussi moins sensible à ces paramètres. En tant que techniques stochastiques, tant la technique FA que la technique GBA introduisent une nouvelle composante d'incertitude : des simulations répétées vont potentiellement produire différentes prédictions finales. Nous illustrons comment la technique FA permet de produire une carte spatialement explicite de cette incertitude stochastique du modèle final. [Traduit par la Rédaction]

Mots-clés : couvert forestier, cartographie prédictive, classification et arbres de régression, forêts aléatoires, « gradient boosting » aléatoire.

1. Introduction

The tree canopy cover in a given area is a primary structural characteristic of both forested and nonforested landscapes. Understanding and quantifying the spatial distribution of tree canopy cover is relevant to many applications, including forest management (Jennings et al. 1999), fire modeling (Rollins and

Frame 2006), air pollution mitigation (Nowak et al. 2006), stream and water temperatures (Webb and Crisp 2006), and carbon storage (Kellendorfer et al. 2006). Because of the importance of tree canopy cover, Homer et al. (2007) developed a 30 m geospatial dataset of percent tree canopy cover in the United States as part of the 2001 National Land Cover Database (NLCD, <http://www.mrlc>.

Received 20 December 2014. Accepted 29 July 2015.

E.A. Freeman and G.G. Moisen. USDA Forest Service, Rocky Mountain Research Station, 507 25th Street, Ogden, UT 84401, USA.

J.W. Coulston. USDA Forest Service, Southern Research Station, 1710 Research Center Drive, Blacksburg, VA 24060, USA.

B.T. Wilson. USDA Forest Service, Northern Research Station, 1992 Folwell Avenue, Saint Paul, MN 55108, USA.

Corresponding author: Elizabeth A. Freeman (e-mail: efreeman@fs.fed.us).

¹This article is one of a series of papers presented at the session "Quantifying uncertainty in forest measurements and models: approaches and applications" held during the XXIV IUFRO World Congress 2014 "Sustaining Forests, Sustaining People: The Role of Research", Salt Lake City, Utah, USA, 5–11 October 2014.

gov/). As part of the development of the updated 2011 NLCD tree canopy cover layer, a national pilot project was launched to test the use of high-resolution photography acquired through the National Agriculture Imagery Program (NAIP) coupled with extensive ancillary data layers through alternative sampling and modeling methodologies (Coulston et al. 2012).

Nationwide mapping of tree canopy cover posed numerous technical questions: alternative means to observe tree canopy cover (Frescino and Moisen 2012), the relationship between photo-based tree canopy cover and canopy modeled from extant forest inventory plots (Toney et al. 2012), repeatability in photo-interpretation (Jackson et al. 2012), efficient sampling strategies (Tipton et al. 2012), and choice of appropriate subpopulations over which to construct predictive models (Moisen et al. 2012). Related to modeling methodologies, Coulston et al. (2012) compared the parametric modeling technique beta regression with the nonparametric technique of random forest (RF) and found that the RF modeling technique outperformed the beta regression approach, based on pseudo R^2 , root mean squared error, and slope versus intercept of the observed versus predicted regression line. In addition, they found that the degree to which RF outperformed beta regression was related to model quality. In regions with high-quality models, RF only slightly outperformed beta regression, whereas in regions with low-quality models, RF offered a more marked improvement. Now, in this study, two nonparametric modeling techniques, both involving an ensemble of classification and regression trees, are compared: RF and stochastic gradient boosting (SGB).

There has been little direct comparison of RF and SGB for continuous response models, although previous work by Chirici et al. (2013) looking at categorical response models found that SGB outperformed RF, particularly for the rare categories.

RF (Breiman 2001) has received considerable attention in the ecological literature. Prasad et al. (2006) gave an overview of the use of RF for ecological prediction. Baccini et al. (2008) used RF to map aboveground biomass in tropical Africa from satellite imagery. Chan and Paelinckx (2008) compared RF with Adaboost for mapping ecotopes from airborne hyperspectral imagery and found both models to work well, with RF having the advantage of faster training time and greater stability and robustness, while Adaboost had a slight advantage in accuracy. Evans and Cushman (2009) used RF to produce continuous probability of occurrence maps for four conifer species in northern Idaho, USA, based on a combination of climatic, topographic, and spectral predictor variables. Gislason et al. (2006) found that RF performed well compared with basic classification and regression tree (CART) models, as well as other bagging and boosting models, for landcover classification based on Landsat and topographic predictors. Lawrence et al. (2006) found that RF outperformed CART models for mapping invasive plants using hyperspectral imagery and also found that the out-of-bag (OOB) accuracy assessments provided by RF were reliable when compared with withheld test data. Powell et al. (2010) compared RF with reduced major axis regression and gradient nearest neighbor imputation for mapping biomass from Landsat satellite imagery and found that RF performed favorably in terms of RMSE, although poorly in terms of variance ratio.

SGB (Friedman et al. 2000; Friedman 2001, 2002) is now gaining recognition in ecological modeling. De'ath (2007) summarized the benefits that SGB can offer ecologists and provides an introduction to its use with ecological datasets, while Elith et al. (2008) provided guidelines for its use in ecological modeling. However, little attention has been paid to tuning SGB models in ecological applications. For example, Güneralp et al. (2014) and Filippi et al. (2014) compare SGB, multivariate adaptive regression splines (MARS), and cubist for mapping aboveground floodplain biomass and found that both SGB and MARS outperformed cubist. Both papers examined mapping at a relatively small spatial scale — their map area was a single river bend of 0.219 km². Their general results may be applicable at all scales, but they did not look at

the practical issues of tuning SGB models for large-scale maps. Güneralp et al. (2014) sets tuning parameters empirically, whereas Filippi et al. (2014) does not examine tuning in detail but states that they tested other learning rates and found poorer performance.

Leathwick et al. (2006) found SGB to offer superior predictive performance to generalized additive models (GAM) for predicting and mapping fish species richness. They looked at three values for the interaction depth parameter and left the other parameters at their default values. Sankaran et al. (2008) used SGB not to produce maps, but rather to investigate the relationship between predictor variables in regulating the woody cover in African savannas. They used the default shrinkage parameter and cross validation to determine optimal values of interaction depth and number of trees. Lawrence et al. (2004) compared SGB and classification tree analysis (CTA) for classification of remotely sensed imagery for use in forestry and found that SGB improved accuracy, although the strength of the improvement was dataset dependent. They do not describe how their SGB parameters were set. Moisen et al. (2006) compared the performance of SGB, GAM, and proprietary tree-based methods for predicting tree species presence and basal area in Utah and found that for the majority of species, SGB models were the most accurate at predicting species presence and competitive with the other techniques for predicting basal area. Parameters were set based on a related dataset, but the tuning process is not described. Baker et al. (2006) found SGB more effective than CTA for mapping wetlands and riparian areas from Landsat data supplemented with topographic and soils data, although model parameter selection is not mentioned. Pittman et al. (2009) found that SGB worked well to predict the diversity and abundance of fish and corals from underwater Lidar bathymetry with parameters left at their default values.

In addition to the issues of model tuning, stochastic techniques such as RF and SGB also introduce a new component of uncertainty: repeated model runs will potentially result in different final predictions. This uncertainty can be reduced by careful model tuning, and in the case of RF models, it is possible to produce a spatially explicit map based estimate of the stochastic uncertainty remaining in the final model. In RF, the individual trees are independent, allowing the standard deviation of the individual-tree predictions to be calculated for each pixel in the map, providing a measure of the stochastic variability in the map's predictions.

Stochastic models such as RF and SGB are powerful modelling techniques that have been shown to provide superior predictive performance to parametric methods in a variety of applications (e.g., Moisen and Frescino 2002, Prasad et al. 2006, Powell et al. 2010), as well as for the specific datasets examined here (Coulston et al. 2012). RF has been used extensively in ecological and remote sensing applications, whereas SGB is beginning to gain recognition. However, there has been little direct comparison of these two techniques for continuous ecological data, which this paper provides. In addition, the nonparametric modeling literature calls attention to the importance of tuning in the modeling process. Yet, in the applied ecological and natural resource literature, particularly predictive modeling of forest attributes, this important phase of the modeling process is often ignored. Consequently, we explore the sensitivity of RF and SGB to choices in tuning parameters in modeling tree canopy cover over four diverse study regions in the conterminous United States (US). Second, the study sought to compare the performance of the two final models in each study area by assessing the importance of, and interaction between, predictor variables under each modeling technique, the global accuracy metrics derived from an independent test set, and the visual quality of the resultant maps of tree canopy cover. Finally, for the RF model, we produce a map of the stochastic uncertainty remaining in the final tuned model and illustrate the value of this metric for map users by examining in detail two areas of this map that show particularly high levels of uncertainty.

2. Materials and methods

2.1. Study regions

Four study regions in the US were used in this pilot study: Georgia (GA), Kansas (KS), Oregon (OR), and Utah (UT). These locations provide a contrast between high crown cover (GA and OR) and lower crown cover (KS and UT). Kansas, in particular, posed a challenge to model tuning as over half of the study region has zero tree canopy cover. These locations also provide a contrast in the types of tree cover present. Georgia has a higher proportion of broad-leaf tree species, whereas Oregon has a high proportion of conifer species. Utah includes large areas of pinyon-juniper woodlands (Table 1).

Each study region was approximately the size of one Landsat scene and were selected to cross local ecological gradients. For example, the GA study area ranged from the Piedmont region in the south, through the Atlanta metropolitan area, to the heavily forested Appalachian Mountains in the north.

2.2. Data

An intensive (approximately 1 km × 1 km) grid of photo-interpretation plots was established over the four pilot regions. The imagery is provided by the National Agriculture Imagery Program (NAIP) (U.S. Department of Agriculture (USDA) 2009) collected during the growing season in 2009. The grid for locating the photo plots is adopted from the Forest Inventory and Analysis (FIA) sample design (Bechtold and Patterson 2005) of a quasi-systematic sample based on White et al. (1992). This design is assumed to produce a random equal probability sample (McRoberts et al. 2005). The FIA sample design has a nominal sampling intensity of approximately one sample location per 2400 ha across all land covers and types. For this pilot study, the FIA sample design has been intensified 4× (1 plot per 600 ha) as described by White et al. (1992).

Each photo plot consisted of a 105-point triangular grid distributed in a 90 m × 90 m square area surrounding the sample location. Each dot was characterized as “tree canopy” or “no tree canopy”. The response variable of percent tree canopy cover was defined as the proportion of tree canopy dots identified on the photo plot. The design-based estimators of proportion canopy cover in each photo plot, mean proportion canopy cover in each study region, and standard error of the estimate were obtained following Cochran (1977) (Table 2).

Predictor variables included transformed aspect, slope, elevation, topographic positional index (Moore et al. 1991), Bailey’s ecoregions (Bailey 1995), land cover and tree canopy cover from the 2001 NLCD (Homer et al. 2004), and Landsat-5 reflectance bands (Table 3). The Landsat data were also leaf-on and from either 2008 or 2009, depending on cloud cover. Because many of the predictor variables originated from 30 m pixel resolution products, assignment to each 90 m × 90 m plot was accomplished by taking a focal mean over a 3 pixel × 3 pixel window for continuous variables and a focal majority for the categorical variables. See Coulston et al. (2012) for more details on data used in this study.

This study is part of a larger mapping project whose goal is updating and improving the 2001 NLCD. This drove the selection of a number of the predictor variables used in this study.

Because we have relatively large data sets (approximately 4000 data points per region), we were able to use independent tuning and test data to build and evaluate both RF and SGB models. We randomly assigned 25% of the data from each region as a tuning set and 25% as independent test data, leaving 50% for model training, as suggested by Hastie et al. (2009).

2.3. Models

CART models (Breiman et al. 1984) are flexible and robust tools that are well suited to the task of modeling the relationship

Table 1. Percentage of photo plots per National Land Cover Database (NLCD) land cover class for the four study regions.

Code	Land cover class	Percentage of photo plots			
		Georgia	Kansas	Oregon	Utah
11	Open water	2	2	1	0
12	Perennial ice/snow	0	0	0	0
21	Developed open space	11	4	2	1
22	Developed low intensity	6	2	1	0
23	Developed medium intensity	2	0	0	0
24	Developed high intensity	1	0	0	0
31	Barren land	1	0	1	4
41	Deciduous forest	39	8	0	4
42	Evergreen forest	19	0	45	35
43	Mixed forest	1	0	2	2
52	Shrub/scrub	1	0	36	45
71	Grassland/herbaceous	3	41	2	3
81	Pasture/hay	12	17	5	3
82	Cultivated crops	0	24	4	1
90	Woody wetlands	2	1	0	0
95	Emergent herbaceous wetlands	0	0	1	0

Table 2. Mean percent tree canopy cover (TCC) and the standard error of the mean (SE(TCC)) for each study area for NLCD2001 forest land cover urban land cover and across all land cover classes based on photo interpretation. (Table from Coulston et al. (2012), reproduced with permission from the American Society for Photogrammetry & Remote Sensing, Bethesda, Maryland (asprs.org)).

Study region	Land cover class	Percent tree canopy cover	
		Mean	SE(TCC)
Georgia	Forest	84.1	0.45
	Urban	41.1	0.94
	All	66.0	0.53
Kansas	Forest	71.0	1.57
	Urban	14.5	1.28
	All	12.8	0.40
Oregon	Forest	66.5	0.60
	Urban	26.4	2.33
	All	41.6	0.51
Utah	Forest	52.0	0.68
	Urban	9.1	1.68
	All	27.4	0.47

Table 3. Predictor variable for models.

Predictors	Description
NORTHNESS	Northness — cos(aspect)
EASTNESS	Eastness — sin(aspect)
SLOPE	Slope
CTI	Compound topographic index
ELEV	Elevation
ECOREG	Bailey’s ecoregions
LC	Land cover class from 2001 NLCD
TCC2001	Tree canopy cover from 2001 NLCD
BAND1	Landsat-5 band 1 — blue
BAND2	Landsat-5 band 2 — green
BAND3	Landsat-5 band 3 — red
BAND4	Landsat-5 band 4 — near IR
BAND5	Landsat-5 band 5 — shortwave IR
BAND7	Landsat-5 band 7 — shortwave IR

Note: For predictor variables originated from 30 m products, assignment to each 90 m plot was accomplished by taking a focal mean or focal majority over a 3 × 3 window for continuous variables and focal majority for the categorical variables. NLCD, National Land Cover Database; IR, infrared. See Coulston et al. (2012) for more details on the predictor variables.

between a response and a set of explanatory variables for the purposes of making spatial predictions in the form of a map. These are intuitive methods, often described in graphical or biological terms. A CART model begins at its root. An observation passes down the tree through a series of splits, or nodes, at which a decision is made as to which direction to proceed based on values of the explanatory variables. Ultimately, a terminal node, or leaf, is reached and a predicted response is given, which is typically the mean of observations in the terminal node for a continuous response, or a plurality vote for a categorical response. See De'ath and Fabricius (2000) for a thorough explanation and Moisen (2008) for a simple overview.

Although CART models are powerful tools by themselves, much work has been done in the data-mining and machine-learning fields to improve the predictive ability of these models by combining separate tree models into what is often called a committee of experts, or ensemble. Two such ensemble techniques considered here are RF and SGB models.

As discussed earlier, RF is receiving increasing attention in the ecological and remote sensing literature. In this technique, a bootstrap sample of the training data is chosen. At the root node, a small random sample of explanatory variables is selected and the best split is made using that limited set of variables. At each subsequent node, another small random sample of the explanatory variables is chosen, and the best split is made. The tree continues to be grown in this fashion until it reaches the largest possible size and is left unpruned. The whole process, starting with a new bootstrap sample, is repeated 500 or more times. The final prediction is a vote (for categorical responses) or average (for continuous variables) from the predictions of all of the trees in the collection. Because each tree is built from a subsample of the training data, the unsampled portion of the data can be used to produce OOB model predictions for that tree. In addition, these independent trees allow a pixel by pixel estimate of the variability in the predictions of the individual trees in the final model. The standard deviation of these predictions can be calculated for each pixel. This is not to be misconstrued as a prediction interval, but instead is a useful measure of uncertainty in the resultant maps.

Unlike RF, which is an ensemble of independent trees, SGB sequentially builds many small classification or regression trees sequentially from "pseudo"-residuals (the gradient of the loss function) of the previous tree. At each iteration, a tree is built from a random subsample of the "pseudo"-residuals (selected without replacement), producing an incremental improvement in the model. In SGB, the trees are not grown to completion (as in RF); instead, the maximum tree size is specified by a model parameter.

2.4. Software

Analysis was conducted in the R software environment (R Development Core Team 2008) using the package ModelMap (Freeman and Frescino 2009). Many of the diagnostic and graphical tools available in the current version of this package were developed concurrently with this study to address questions that occurred while comparing these models. ModelMap constructs predictive models of continuous or discrete responses by calling the R packages randomForest (Liaw and Wiener 2002) and gbm (Ridgeway et al. 2013), respectively. These models are then applied to image files of predictors to create detailed prediction surfaces.

2.5. Tuning process

2.5.1. Tuning RF

RF, as implemented by the R package randomForest, only requires the user to make decisions about two tuning parameters. The first, *mtry*, controls the number of predictor variables randomly sampled to determine each split. RF models are relatively insensitive to the choice of *mtry* (Breiman 2001; Liaw and Wiener

2002), although higher values of *mtry* tend to work better in cases where only a few of the predictors contribute to the model and there are a lot of "noise" predictors containing no useful information (Liaw and Wiener 2002; Prasad et al. 2006). The second tuning parameter, *ntrees*, controls the total number of independent trees. The number of trees required to stabilize variable importance (Liaw and Wiener 2002) and variable interaction (Evans and Cushman 2009, referencing personal communication of Adele Cutler) may be larger than the number required to stabilize prediction accuracy.

For *mtry*, a suggested starting point for tuning continuous response models is the number of predictor variables divided by three, followed by checking half this number and twice this number. Our models had 14 predictor variables, and thus we considered three possible values for *mtry*: 2, 4, and 8. Using the 50% training data set, we built 20 models for each value of *mtry*, with 2500 trees each.

From each model, predictions were made on the 25% tuning data set from subsets of increasing numbers of trees (100 trees, 200 trees, ..., 2500 trees). Three error measures (mean squared error (MSE) and Pearson and Spearman correlations) were plotted against number of trees, with the *mtry* value of each model indicated by line color.

2.5.2. Tuning SGB

SGB, implemented by the R package gbm (Ridgeway et al. 2013) requires the user to make choices about a larger number of tuning parameters, including shrinkage, bagging fraction, interaction depth, and number of trees. Note that the names used for these parameters differ throughout the literature. This is just differing terminology for describing the same parameters, but these differing terminologies can be daunting for new users. Even the name of the technique itself varies in the literature. Table 4 provides a cross reference for translating among the terminologies used by different authors.

Tuning SGB models is complicated by the fact that changing any one of the parameters can affect the optimal values of the other parameters. Shrinkage, also known as the learning rate, controls the influence of each successive tree on the final predictions. A lower shrinkage (i.e., a slower learning rate) increases the number of trees required and thus increases computing time but also reduces the chance of overfitting. Bagging fraction, also known as sampling fraction, controls the fraction of the training data randomly selected to build each tree. Smaller bagging fractions reduce the chance of overfitting but result in increased variability between model runs, i.e., increased model uncertainty (Friedman 2002). Interaction depth, also known as tree size or tree complexity, controls the maximum size of each tree. As in RF, number of trees controls the total number of trees. Unlike in RF, using too many trees in SGB does result in overfitting and poorer model performance on independent test data. Ridgeway (2007) recommended balancing shrinkage and number of trees to result in models with between 3000 and 10 000 trees and shrinkage rates between 0.01 and 0.001. Elith et al. (2008) recommended models have at least 1000 trees.

As in RF, we built models using the 50% training data and then used accuracy statistics calculated on the 25% tuning data to optimize model parameters. We started by building 10 models of 6000 trees each, for combinations of shrinkage (0.008, 0.004, 0.002, 0.001), bagging fraction (0.1 to 0.9), and interaction depth (1, 2, 4, 8), for each of the four regions. We calculated the three error measures (MSE and Pearson and Spearman correlations) using the tuning data and then averaged these error measures over the 10 models for each combination of model parameters. We then followed this with a separate fine tuning for each individual region. For example, in pilot regions that seemed to still be improving at an interaction depth of 8, we tried interaction depths of 10 and 12. After optimizing bagging fraction and interaction depth,

Table 4. Terminology for stochastic gradient boosting (SGB) model parameters as used by various authors.

Friedman (2002)	Elith et al. (2008)	Ridgeway (2007)	gbm package
Stochastic gradient boosting	Boosted regression trees	Generalized boosted models	Generalized boosted models
Error distribution	Response type	Distribution	Distribution
M — iterations	nt — number of trees	T — number of iterations	n.trees
L — tree size/number of terminal nodes	tc — tree complexity/number of nodes	K — interaction.depth	interaction.depth
ν — shrinkage	lr — learning rate	λ — learning rate	shrinkage
f — sampling fraction	Bag fraction	p — subsampling/bagging rate	bag.fraction

we selected a shrinkage that resulted in a model that met our goal of having the optimum number of trees lie between 3000 and 5000 to balance model accuracy with reasonable computational efficiency.

2.6. Model comparisons

One final RF and one final SGB model were run for each of the four pilot regions using the 50% training data and tuning parameters optimized through the tuning process described above. Comparisons were then made in each pilot area between the two final models by assessing the predictor variables for relative variable importance and interaction effects under each modeling technique, the map accuracy metrics derived from the 25% test set, and the quality of the resultant maps of tree canopy cover.

2.6.1. Role of predictor variables

Unlike simpler model structures such as linear models, tree-based ensemble models do not have a straightforward formula linking predictor variables to model predictions. There are, however, several techniques that can be used to shed some light on the underlying relationships.

Variable importance

Both RF and SGB provide estimates of the relative importance of each predictor variable to the final model. The randomForest package offers two options for calculating variable importance, first by permuting OOB data, and second by calculating the decrease in node impurities from splitting on the variable. The gbm package calculates the relative influence of each variable in reducing the loss function. With continuous response (modeled with a Gaussian loss function), this is the reduction of squared error attributable to each variable. For this study, we compare the permutation importance for our RF models with the relative influence of our SGB models.

Variable interactions

We also examined two-way interactions between predictor variables graphically using the three-dimensional partial dependency plots presented in Elith et al. (2008). The “model.interaction.plot” function from the ModelMap package further develops these plots to include the ability to investigate both continuous and categorical predictor variables. In these plots, the predictor variables are examined two at a time. An x - y grid is created of possible combinations of predictor values over the range of both variables. The remaining predictor variables are fixed at either their means (for continuous predictors) or their most common value (for categorical predictors). Model predictions are generated over this grid and plotted as the z axis.

2.6.2. Accuracy measures

Histograms were made of the distribution of tree canopy cover (TCC) for the observed test data within each pilot area and compared with the histograms of the RF and the SGB predictions. In these histograms, the vertical bars represent the number of photo plots in the test set that have a given value of percent canopy cover. Comparing the observed distribution of plots as a function of crown cover with that of the model predictions allows an assessment of how well the model predictions capture the range of

the observed data, in particular, how well they predict the extremes of the data — does each model predict an accurate proportion of very low and very high canopy cover?

In addition, for each pilot area, we examined final model accuracy in terms of MSE and Pearson and Spearman correlation coefficients, the difference between the mean of the observed TCC and the mean of the predicted TCC, and the slope and intercept of the observed versus predicted TCC regression line. Pearson correlation coefficient is a measure of the linear relationship, whereas Spearman correlation coefficient is a measure of the monotonic relationship (Hauke and Kossowski 2011). Although it is common to use Pearson correlation for ordinal data and Spearman correlation for ranked data, Spearman correlation can also be useful even for ordinal data if there are outliers or to identify a nonlinear but monotonic relationship.

2.6.3. Map quality

We used the final RF and SGB models for Utah to produce maps for that region, with predictions for each 90 m pixel. The maps were generated using the same scales and colors and were examined visually for spatial structures that differed between the RF and SGB predictions. Histograms of number of pixels predicted by percent predicted TCC over all map pixels in the Utah pilot region were compared for the final RF and SGB models.

A map of the stochastic uncertainty remaining in the final RF model for Utah was also created. This map was built by calculating the standard deviation for each pixel from the predictions of each of the independent randomly generated trees that compose the RF model. If the individual trees are in agreement, the uncertainty is low. If the trees are not in agreement, with some trees predicting low TCC and others predicting high TCC, then the uncertainty is higher. In SGB, the trees are not independent, thus this uncertainty map is not available.

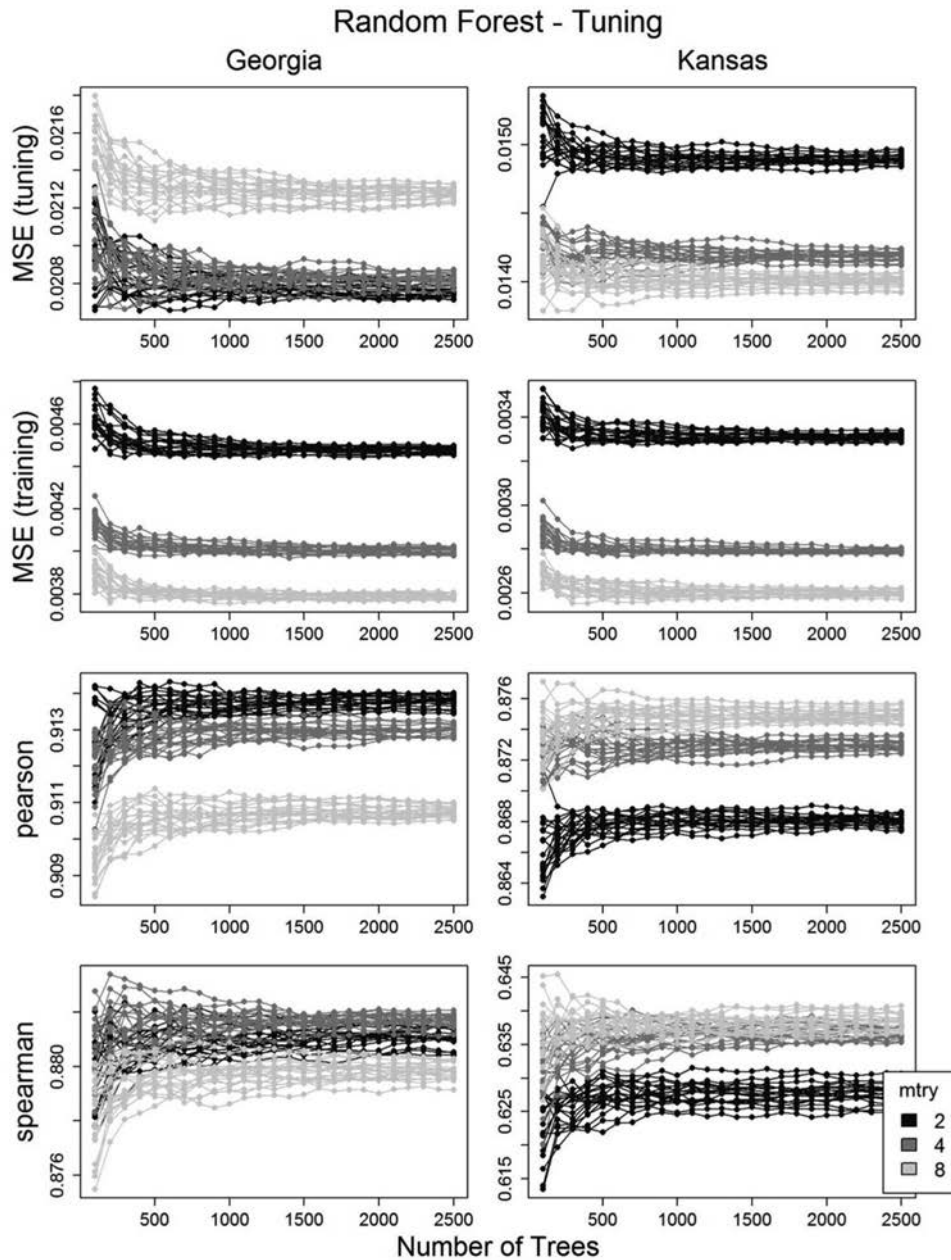
3. Results

3.1. Tuning process

3.1.1. Tuning RF models

As suggested by Breiman (2001) and Liaw and Wiener (2002), RF proved to be relatively insensitive to the choice of $mtry$ (Fig. 1). Figure 1 illustrates how accuracy metrics are affected by the $ntree$ and $mtry$ parameters for two of our study regions. The lines represent each independent model run, with the shade of the lines indicating the value of $mtry$ used for that particular model run. The vertical spread of a given color indicates the variation between independent model runs, in other words, the stochastic uncertainty. This uncertainty is highest when the models have relatively few trees and decreases as more trees are added. The influence of $mtry$ can be seen by the distance between the three shades of the lines once the model runs have stabilized at higher values of $ntree$. In this figure, Georgia shows greater stochastic uncertainty for a given value of the $mtry$ parameter (a greater spread within the lines of a given color), particularly for low values of $ntree$, whereas Kansas shows a slightly greater sensitivity to the $mtry$ parameter (a greater distance between the three colors), particularly at higher values of $ntree$.

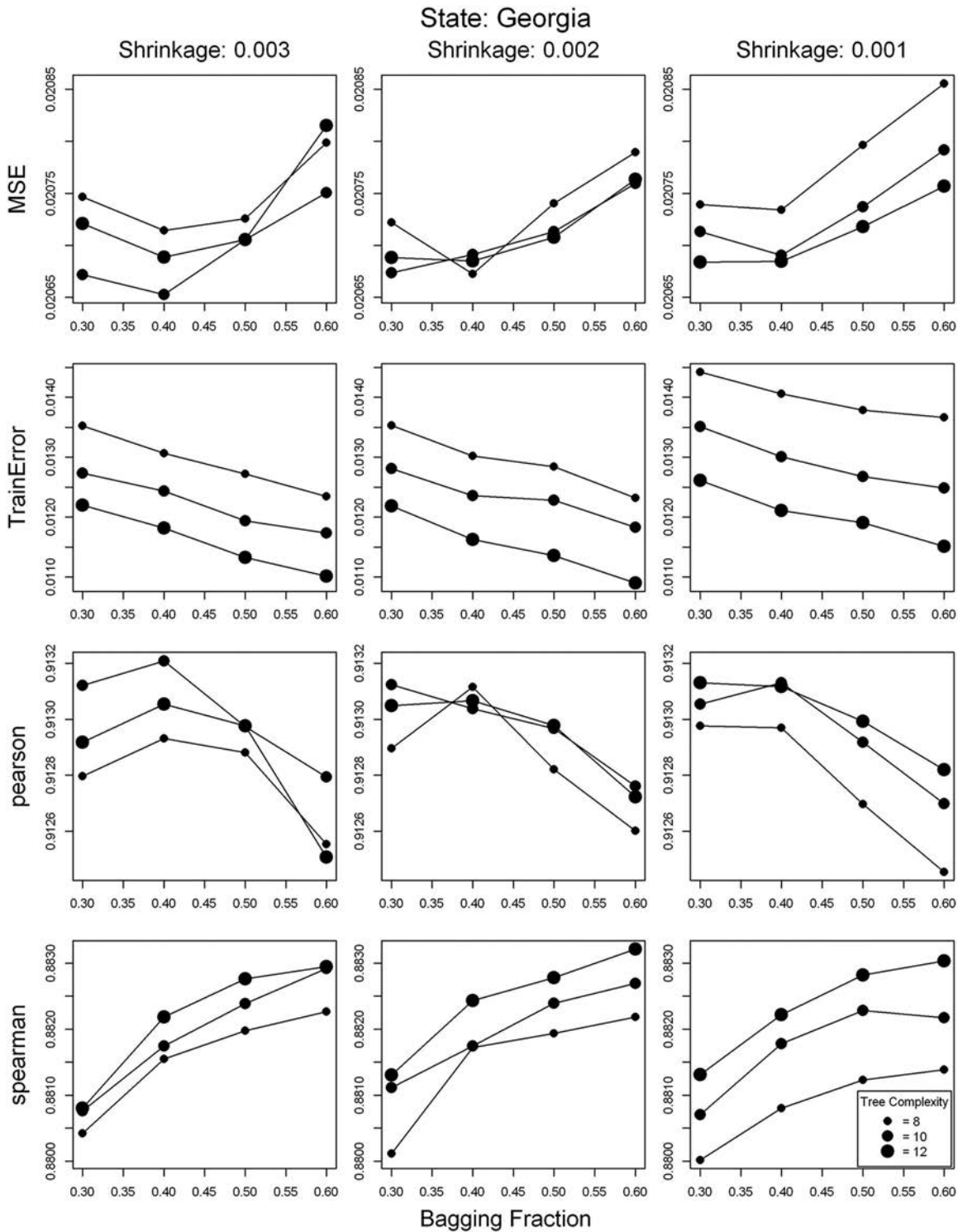
Fig. 1. Tuning RF — effect of *mtry* and *ntree* on RF models for Georgia and Kansas. Each line indicates one model as the number of trees is increased. In each region, 20 models were fit for each of three values of *mtry* (line shading indicates different *mtry* values). The error is plotted as a function of the number of trees in each model. As *ntree* increases, the stochastic uncertainty of the models decreases and the lines for a given value of *mtry* converge. Good models have a low MSE and high Pearson and Spearman correlations. MSE is shown both on the tuning data and on the training data. In Georgia, *mtry* values of 2 and 4 had lowest MSE, with *mtry* of 2 having highest Pearson correlation and *mtry* of 4 having highest Spearman correlation. In Kansas, *mtry* of 8 was best by all three measures. These RF models do not show evidence of overfitting with increasing numbers of trees, as both the tuning and the training MSE initially decrease and then remain stable as trees are added to the models.



In most regions, MSE and correlation stabilized at between 1000 and 1500 trees, and there was no evidence of overfitting with the larger number of trees. As trees are added to the models, both the training and the tuning MSEs initially decrease and then stabilize at higher numbers of trees (overfitting would have led to the tuning data MSE increasing with higher numbers of trees and the training MSE continuing to decrease). Keeping in mind that stabilizing variable importance (addressed later) may require more trees than stabilizing model accuracy (Liaw and Wiener 2002), we used 2000 trees for our final models. The four study regions exhibited similar effects.

The RF models were relatively insensitive to the *mtry* parameter. Of the four study regions, Kansas showed the greatest sensitivity to *mtry*, but even in that region, the effects of varying *mtry* were minor: when the tuning set error statistics were compared for the three values of *mtry*, MSE ranged from 0.014 to 0.015, Pearson correlation ranged from 0.868 to 0.865, Spearman correlation ranged from 0.63 to 0.64, and the predicted mean ranged from 0.129 to 0.130. An *mtry* of 4 performed best overall in Georgia and Oregon, whereas an *mtry* of 8 performed best in Kansas and Utah, but in all four regions, the improvements from tuning *mtry* were very slight.

Fig. 2. Tuning SGB — effect of interaction depth (tree complexity) and bagging fraction on SGB models of Georgia. Points represent best number of trees for each combination of parameters. Final model was constructed with shrinkage = 0.002, bagging fraction = 0.40, and interaction depth = 10.

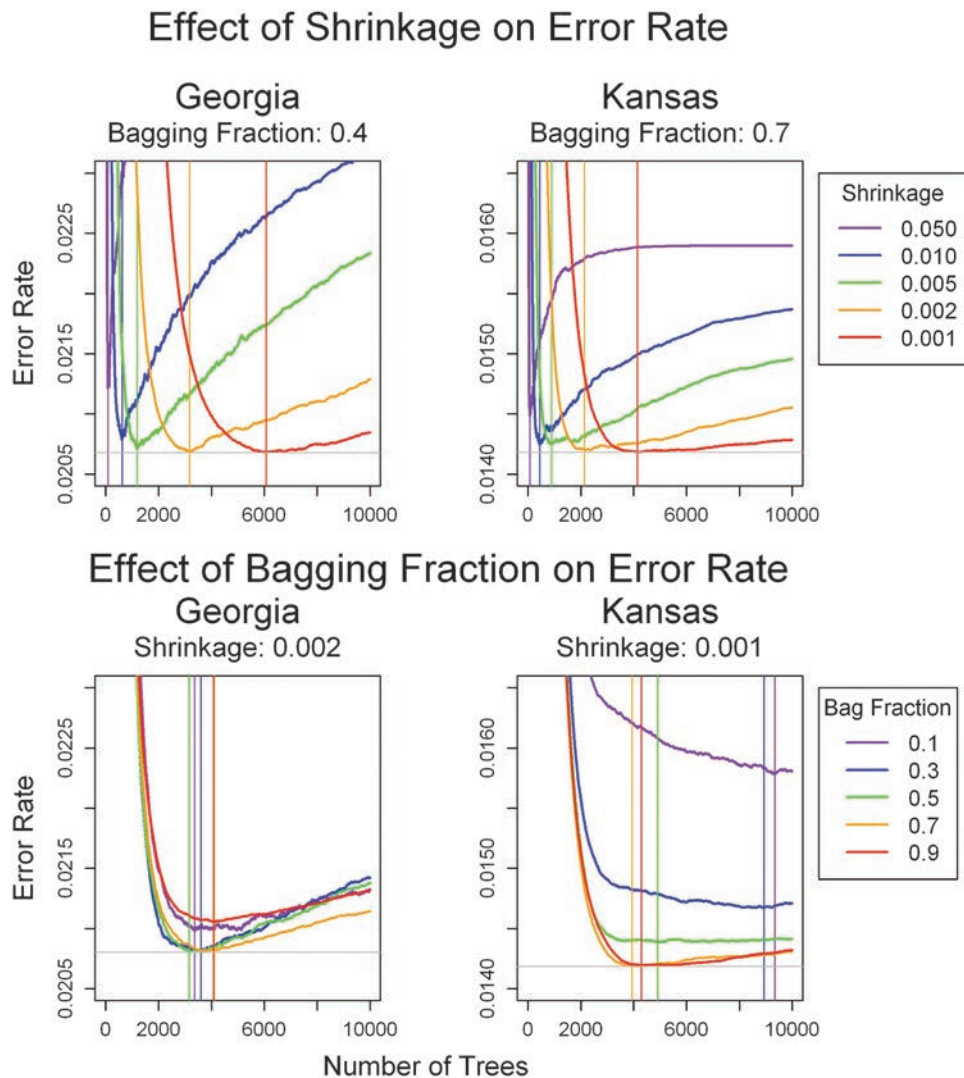


3.1.2. Tuning SGB models

For each value of shrinkage, we plotted the average error measures (from the optimal number of trees for that combination of parameters) as a function of bagging fraction, with interaction depth represented by point size (Fig. 2). Better models have lower MSE and higher correlation.

The SGB models do show overfitting for higher numbers of trees, and the number of trees that leads to overfitting varies with the values of the other parameters. The tuning error initially decreases as trees are added, but then as more trees are added, the tuning error begins to rise again. The number of trees used in the final SGB models was chosen to minimize this overfitting.

Fig. 3. Tuning SGB — effects of shrinkage (learning rate) and bagging fraction on tuning set error rates in Georgia and Kansas. All other model parameters optimized for each region. Notice that the same shrinkage rate requires more trees to reach the best model performance in Georgia than in Kansas. Our goal of 3000–5000 trees required a shrinkage rate of 0.002 in Georgia and 0.001 in Kansas. In Georgia, the best bagging fraction is between 0.3 and 0.5 (we used 0.4 in the final model). In Kansas, 0.7 and 0.9 tie for best bagging fraction (we used 0.7 is the final model). Notice also that the error line for models with low bagging fractions (0.1 and 0.3) jitters slightly due to increased stochastic uncertainty. The SGB models show evidence of overfitting with increasing number of trees, particularly with higher values of shrinkage. The tuning error initially decreases as trees are added, but then as more trees are added, the tuning error begins to rise again. This illustrates why it is essential with SGB models to optimize the number of trees used for the final model.

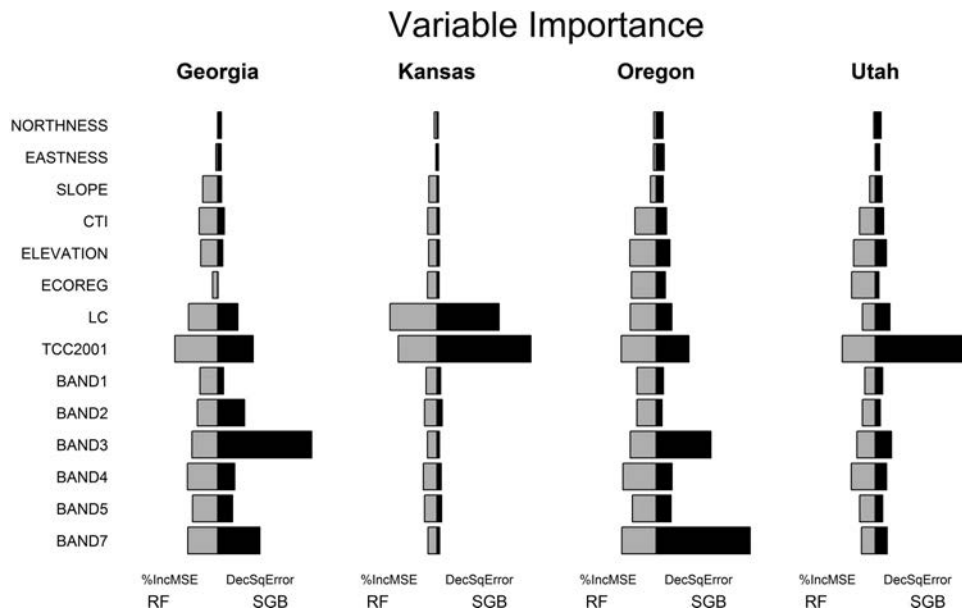


With larger values of shrinkage (shrinkage = 0.05 and 0.01), models begin to overfit before reaching optimum model accuracy (Fig. 3). However, we found that for the smaller values of shrinkage (shrinkage ≤ 0.005), the models all reached similar accuracy before overfitting was observed. For these smaller shrinkage values, we found the relationship between model quality, bagging fraction, and interaction depth was generally stable as shrinkage was changed, given that the number of trees was increased to compensate for the lower shrinkage (slower learning rate). It just took more trees to reach optimal model accuracy at lower (slower) shrinkages (Fig. 3). Therefore, we began by picking the optimal combination of bagging fraction and interaction depth and then selected the shrinkage that resulted in 3000–5000 trees. Notice that the same shrinkage rate requires more trees to reach the best model performance in Georgia than in Kansas (Fig. 3). Our goal of 3000–5000 trees required a shrinkage rate of 0.001 in Kansas, whereas Georgia, Oregon, and Utah required the slightly faster

shrinkage rate of 0.002 to have the final number of trees in the preferred range.

We expected to find that the best models had bagging fractions near 0.5. Ridgeway (2007) suggests starting with bagging fractions near 0.5, and Elith et al. (2008) found bagging fractions between 0.5 and 0.75 worked best. Friedman's data (Friedman 2002) was best modeled with a bagging fraction of 0.4. In some of our regions, bagging fractions near 0.5 did give the best results. For example, in Georgia, the test MSE and the Pearson correlation were best at a bagging fraction of 0.4, whereas Spearman correlation was optimized at the slightly higher bagging fraction of 0.6 (Fig. 3). Utah performed best with a bagging fraction of 0.5. In Oregon, bagging fraction had little effect on model quality, but a quite low bagging fraction of 0.2 seemed to very slightly improve model fit. Smaller bagging fractions introduce more stochasticity into the model and, therefore, can counteract overfitting (Friedman 2002). This increased stochasticity can be observed in Fig. 3, where

Fig. 4. Variable importance for the RF and SGB models of the four pilot regions (RF on the left in gray, and SGB on the right in black). RF importance is measured by the percent increase in MSE (%IncMSE) with random permutation of each variable. SGB importance is measured by the decrease in squared error (DecSqError) attributed to each variable in the gradient of the loss function. Variable importance of each model is scaled to sum to 1.



the lines representing very low bagging fraction show a jitter. In contrast, Kansas (where 53% of plots have zero TCC) did better with a larger bagging fraction of 0.7 (Fig. 3). This is probably due to the fact that with a higher bagging fraction of 0.7, it is more likely that each tree will be based on a subset containing at least some nonzero responses.

To balance computational complexity with model improvement, we settled on an interaction depth of 10 for the four regions. Some of our regions showed slight improvement with even higher interaction depths, but we seemed to be reaching a point of diminishing returns. For example, increasing interaction depth from 10 to 12 had much less of an effect than the change from 8 to 10. An interaction depth of 10 was a compromise between model improvement, computation time, and risk of overfitting. We then selected a shrinkage rate slow enough to maintain the desired 3000–5000 trees. The numbers of trees for the final models ranged between 3500 and 4500.

3.2. Role of predictor variables

3.2.1. Predictor variable importance

We expected the most important predictors to be similar between the SGB and the RF models. This proved true in two of our regions (Kansas and Utah), but in the other two regions (Georgia and Oregon), RF and SGB differed in their choice of most important variable (Fig. 4). In both of these regions, the most important SGB predictor was a remote-sensing band (band 3 in Georgia, band 6 in Oregon). These remote-sensing bands were highly correlated with several of the other bands (between-band correlation up to 0.92 in Georgia and 0.93 in Oregon) and moderately negatively correlated with TCC2001 (correlation with TCC2001 of -0.55 in Georgia and -0.75 in Oregon). RF models spread the importance among the correlated predictors, whereas SGB models concentrated the importance in a single band.

For both RF and SGB models, most of the variable importance is split between two predictors in Kansas. Unlike SGB importance, in RF, when the importance is concentrated in a small number of variables, it does suggest that the other predictors are noise variables. In such cases, using a higher value of *mtry* can often improve the model, as it is more likely that the randomly selected

variables for each split will contain at least one non-noise variable. This may explain why Kansas was the pilot region that performed better with a higher than default value for *mtry* (*mtry* = 8 instead of *mtry* = 4).

3.2.2. Predictor variable interactions

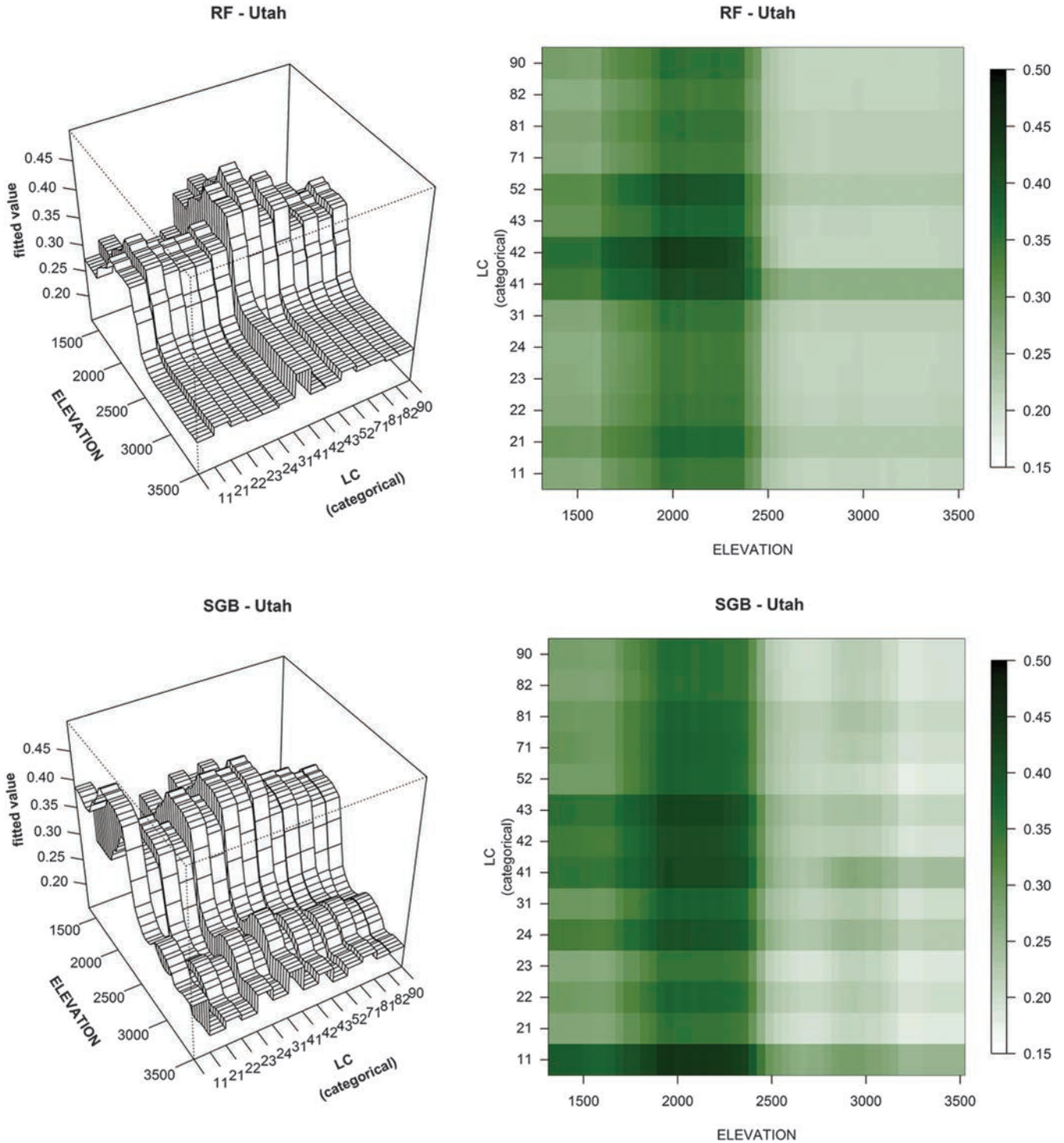
We did not find dramatic interaction effects in this dataset, but we did find subtle interactions. For example, in the RF model for Utah, there are some interactions between land cover class and elevation (Fig. 5); in particular, there are some interesting difference in the effect of elevation in land cover classes 41 (deciduous forest) and 42 (evergreen forest). This figure shows the predicted TCC for each land cover class across the range of elevations found in the training dataset for Utah. In class deciduous forest, TCC is moderate at the lowest elevations, increases at mid elevations, and drops slightly to moderate levels at high elevations. In class evergreen forest, TCC starts slightly higher than class 41 at low elevations, rises at mid elevations, but then drops to near zero at high elevations.

Also, in Fig. 5, notice the differences between the RF model and the SGB model. Although overall accuracy of the two models is similar, the interaction plots highlight subtle differences in the relationships of TCC to the predictor variables. For example, although both RF and SGB models show a peak crown cover at low to mid elevations (1600 m to 2400 m), the SGB model also has a small bump in crown cover at a higher elevation (3000 m).

Another example of differing predictor effects in Utah is illustrated in Fig. 5 by the predicted TCC in land cover class 11 (open water). The SGB model is predicting high TCC for points landing in open water in Utah, particularly at low to mid elevations, although the difference can be seen to some extent at all elevations. A possible explanation may be seen in the variable importance plot for Utah (Fig. 4). The SGB model concentrated importance in TCC2001, whereas the RF model spread the importance between multiple predictors.

The interaction plots examine the effects of the two selected predictors with the remaining variables fixed at their mean value (or most common value for categorical predictors). Therefore, Fig. 5 shows the theoretical predictions for pixels at each land

Fig. 5. Interaction plots for elevation (ELEVATION) and land cover class (LC) for final Utah models. These figures show the effect of changes in two predictor variables on predicted TCC, with all other variables held at their mean (or majority). To see an example of interaction, look at the RF model and the effect of elevation in land cover classes 41 (deciduous forest) and 42 (evergreen forest). In class 41, TCC is moderate at the lowest elevations in the Utah region, increases at mid elevations, and drops slightly to moderate levels at high elevations. In class 42, TCC starts slightly higher than class 41 at low elevations, again rises at mid elevations, but then drops to near zero at high elevations. Also notice the differences between the RF model and the SGB model. Although overall accuracy of the two models is similar, the predicted TCC in class 11 (open water) is very different.

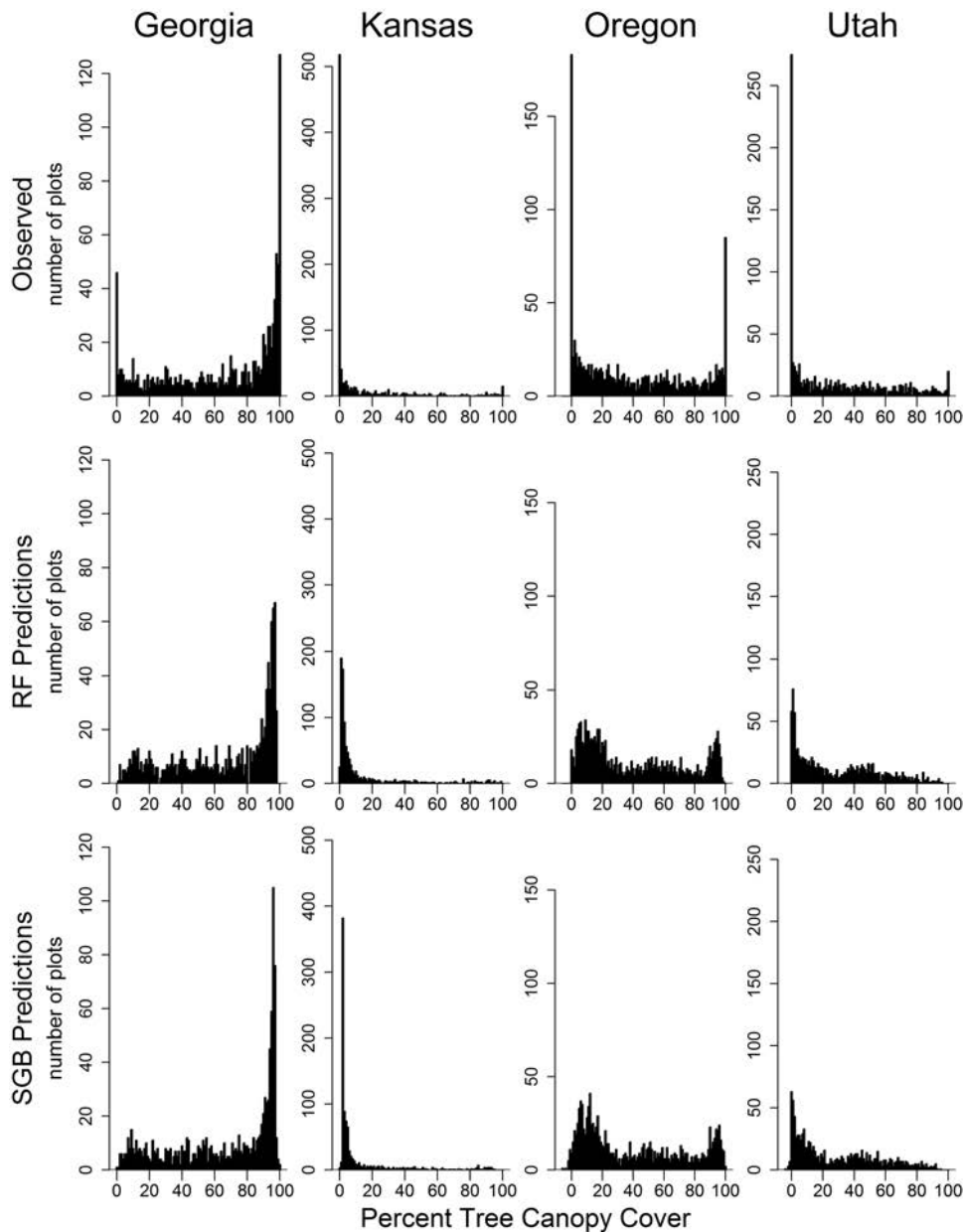


cover class and elevation with average values of TCC2001 (and all other predictors.) In the training data, all points with land cover class 11 have TCC2001 of zero, whereas the overall mean value of TCC2001 is 16.8. So the interaction plot is showing how the RF and

SGB models would extrapolate a combination of variables not found in the training data.

In this case, the RF model extrapolation, where importance is shared among multiple predictors, is more sensible than that of

Fig. 6. Histograms illustrating number of photo plots in the independent test sets by percent tree canopy cover (1% classes). The observed plot distribution varies across the four regions. Georgia and Oregon include plots across the full range of possible canopy cover, with spikes in the number of plots at 0% and 100%. Utah is similar, except that it has a higher proportion of low canopy cover plots and fewer plots with 100% canopy cover. In Kansas, the majority of the plots have very low canopy cover, with only a small number of plots at higher canopy covers. Both RF and SGB models had difficulty capturing the observed spikes in number of plots at 0% and 100%.



the SGB model, where a single variable is driving the predictions. Because TCC2001 is based on canopy cover predictions from a previous model, if that model had erroneously classified some water as forest, the error would disproportionately affect the SGB model, as the SGB model for Utah is relying almost exclusively on that single variable (Fig. 4).

3.3. Accuracy measures

3.3.1. Histograms of plot frequency by percent canopy cover

We created histograms of plot frequency by percent canopy cover (1% classes) for the photo plots in the independent test set. The height of the vertical bars represents the number of plots found in each 1% of TCC. These histograms were created for both the observed TCC and the TCC predicted by the RF and SGB models.

For all four regions, the RF and SGB predictions for the test set had plot frequency histograms similar to the observed data, and when the predictions failed to match the observed histograms, both RF and SGB failed in similar ways (Fig. 6).

The observed distribution of plots across low, medium, and high canopy cover varies across the four regions. The observed data in several regions had spikes at zero and (or) 100% canopy cover. Georgia and Oregon include plots across the full range of possible canopy cover, with spikes in the number of plots at 0% and 100%. Utah is similar, except that it has a higher proportion of low canopy cover plots and fewer plots with 100% canopy cover. In Kansas, the majority of the plots have very low canopy cover, with only a small number of plots at higher canopy cover.

Both models miss the canopy cover spike at zero in Oregon and severely underestimate it in Utah. Both models do capture the spike at zero canopy cover in Kansas, although they do not predict quite as much low canopy cover as observed in the test data and tend to place the spike at very low canopy cover rather than at zero canopy cover. In Kansas, SGB places the spike at 2% cover and RF spreads the spike between zero and 3% canopy cover.

Both models do an even worse job at capturing the observed spikes at 100% canopy cover. In Oregon, the 100% canopy cover spike is reduced to a slight bump in frequency of plots with high canopy cover. The models do better in Georgia, but the spike is still reduced slightly in amplitude and shifted down from 100% canopy cover and spread between 95% and 98% cover for RF and 95% and 97% cover for SGB.

3.3.2. Error statistics from the independent test set

There was very little difference in model performance between RF and SGB as measured by global accuracy metrics (Table 5). In all four study regions, the independent test set MSE was identical to three decimal places, with the largest difference in Kansas where RF gave an MSE of 0.0113 and SGB gave an MSE of 0.0117. Pearson correlation coefficient was identical to two decimal places, with Kansas again showing the largest difference in that RF had a correlation coefficient of 0.905 and SGB had 0.901. The Spearman correlation coefficient for RF was slightly worse than that for SGB in Kansas, but only by 0.01. The largest difference in the predicted mean TCC was in Georgia, but even there the difference was only 0.004. The largest difference in slope of the regression line was 0.03 in Oregon, and the largest difference in intercept was 0.01 in Georgia, Oregon, and Utah.

Not only were the differences between RF and SGB models small, but also there was no clear pattern to which type of model performed best. In terms of MSE, RF was slightly better in Kansas and Oregon, whereas SGB was slightly better in Georgia and Utah. In terms of Pearson correlation, RF was slightly better in Kansas, whereas SGB was slightly better in Georgia, Oregon, and Utah. In terms of Spearman correlation, RF was slightly better in Utah, whereas SGB was slightly better in Kansas, Georgia, and Oregon. Note that some of these differences were so small as to be negligible. For example, in Utah, the RF model had an MSE of 0.02976 and the SGB model had an MSE of 0.02977.

3.4. Map quality

In Utah, the final models were used to produce detailed TCC maps with predictions for every 30 m pixel (Figs. 7 and 8). These figures illustrate the NAIP09 imagery (USDA 2009) for portions of the Utah region, with the corresponding RF- and SGB-predicted TCC and the RF uncertainty.

Histograms of number of pixels predicted by percent canopy cover for the Utah maps were nearly identical for both RF and SGB.

The biggest difference between the RF and SGB maps was that the SGB model extrapolated beyond the values of TCC found in the training data. The training data for UT had TCC ranging from 0% to 100%. The RF map predictions ranged from -0.0001% to 97%, approximately within the range of the training data. In contrast, the SGB map predictions ranged from -13% to 106%. In producing the final maps, we treated predicted TCC values less than 0% as 0% and TCC values greater than 100% as 100%.

Overall, the maps for both models look similar, with RF picking up slightly more detail in the regions of low crown cover. This highlights the need for potentially masking out nonforest areas to eliminate spectrally dark anomalies in rangelands from being predicted erroneously as trees.

The map of RF uncertainty highlighted small anomalous areas of high uncertainty. These are localized areas where the predictions from individual trees varied widely, with some of the trees in the RF predicting low TCC and other trees predicted high TCC.

Table 5. Independent test set error statistics^a from final models: stochastic gradient boosting (SGB) and random forests (RF).

Study region	Error statistics	Model	
		SGB	RF
Georgia	MSE	0.0185	0.0188
	Pearson	0.919	0.917
	Spearman	0.882	0.881
Kansas	MSE	0.0117	0.0113
	Pearson	0.901	0.905
	Spearman	0.673	0.662
Oregon	MSE	0.0246	0.0246
	Pearson	0.896	0.896
	Spearman	0.850	0.880
Utah	MSE	0.0298	0.0298
	Pearson	0.833	0.833
	Spearman	0.850	0.850

^aError statistics: mean square error (MSE), Pearson correlation (a measure of the linear relationship), and Spearman correlation (a measure of monotonic nonlinear relationship).

Two such regions are examined in greater detail in Figs. 7 and 8. Figure 7 illustrates a lava bed, and Fig. 8 shows a wetland.

4. Discussion

A number of lessons were learned through the course of this study. First, we need to emphasize the importance of having an independent tuning dataset, particularly for SGB models. We increased our understanding of the selection of tuning parameters for RF and SGB models. We also learned some interesting things about the effect of correlated predictor variables on variable importance measures and how these effects differ between RF and SGB.

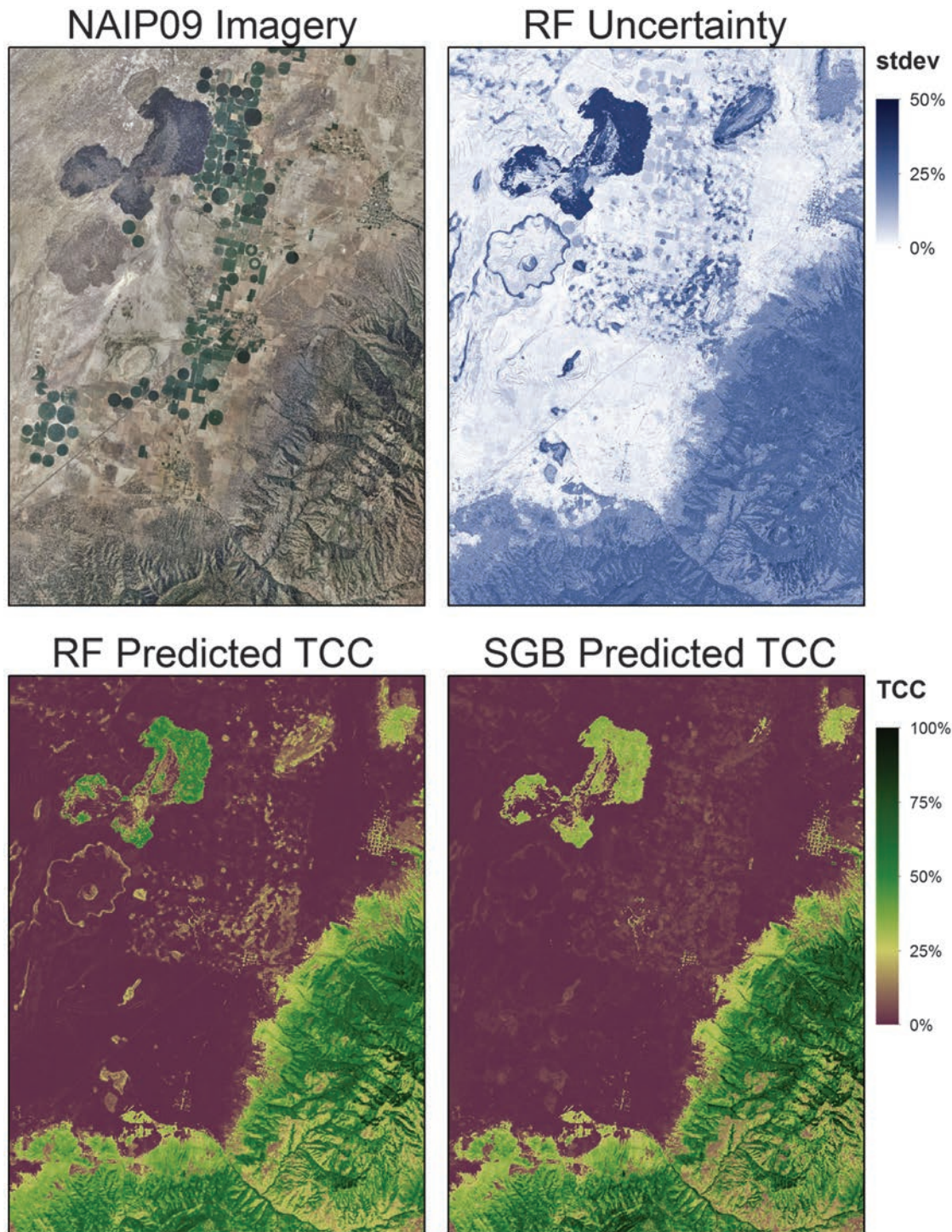
4.1. Importance of an independent tuning data set

Tuning is a sometimes overlooked step in the model building process, both for traditional parametric models and for newer nonparametric models. Parametric models such as beta regression depend on the choice of link function, preliminary removal of highly correlated variables, and selection of final variables to include in both the model and, optionally, the precision equation. Nonparametric models do not require the elimination of nonsignificant variables, but they do require optimizing model parameters, SGB more so than RF.

Model production is a two-step process: first, building potential models with a range of values of model parameters; and second, selecting the combination of model parameters that gives the best performance as a final model. It is obvious that assessing a model over the data used to train the model will not give a true estimate of performance on new independent data, as there will be no way to distinguish overfitting from simple good performance. It is less obvious but equally hazardous to base a final model assessment on the data used for parameter selection. The selected model could potentially perform best on that particular data set but not generalize to new independent data. Therefore, particularly when working with complex models with large numbers of parameters that need to be tuned, it is important to set aside both a tuning set and a separate independent test set to be used only for final model assessment.

If the final model can be built using only the training data, then a tuning set is not necessary. If decisions about the model are based on how well the model predicts over a dataset, then that dataset can no longer be used to judge final model performance. With RF models, the default parameters generally perform well (Liaw and Wiener 2002); therefore, it is possible to accept the default RF parameters and use the entire dataset as training data and use OOB for final model evaluation. If a particular study does

Fig. 7. Detailed maps of the northwest portion of the Utah region, showing NAIP09 imagery, RF uncertainty, and RF and SGB predictions for tree canopy cover (TCC). The dark patch in the upper left of the photo is a lava bed. Both RF and SGB mistook it for an area of moderate TCC (RF predicted slightly higher cover than SGB). The RF uncertainty map shows that although the mean prediction from the 2000 trees in the model was for moderate TCC, there was a very high level of uncertainty in these predictions, with some trees predicting low TCC and other trees predicting much higher TCC.

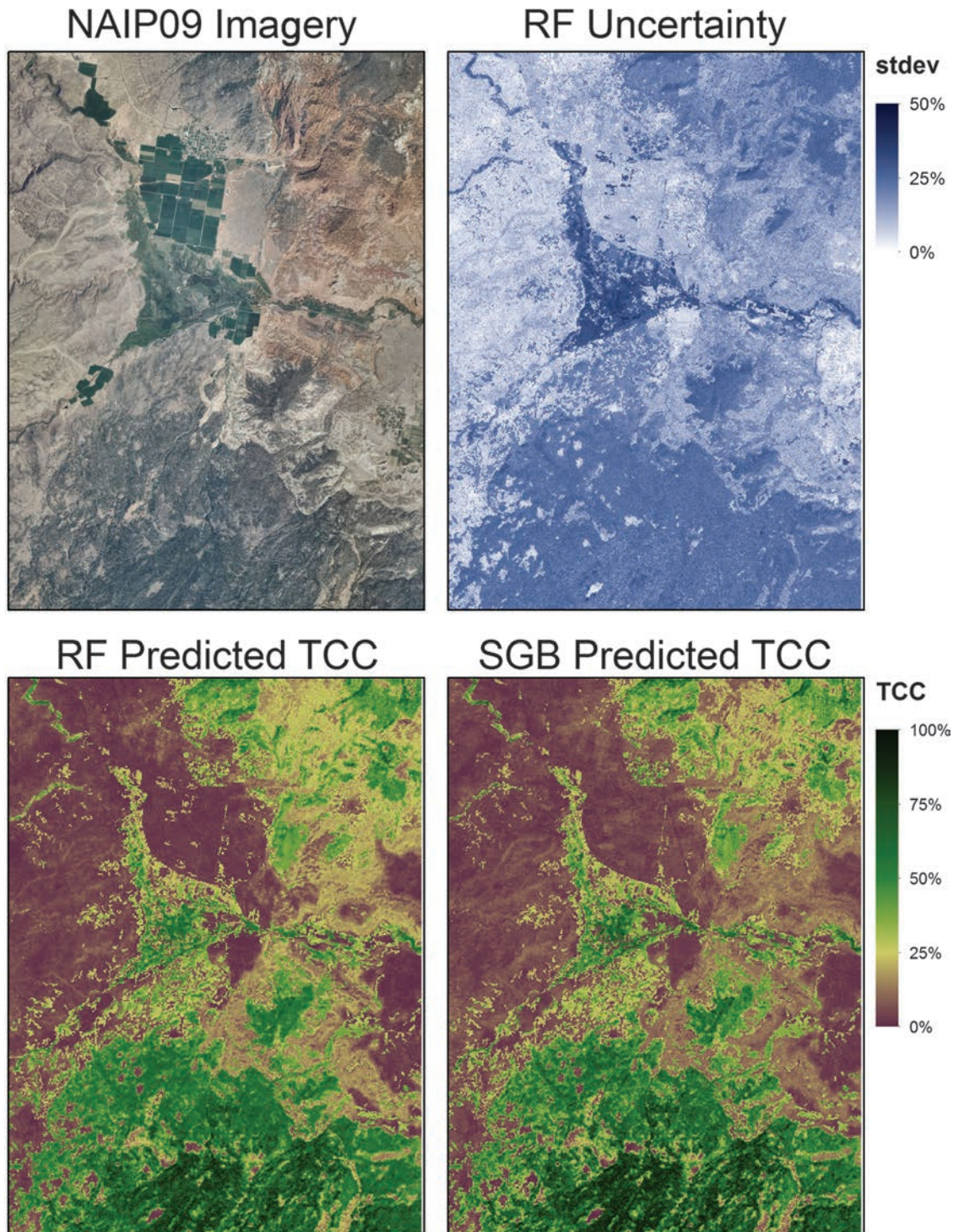


require tuning the RF parameters, then OOB errors can be used for tuning, and an independent test set can be used for final model evaluation. With SGB models, there is not an option for OOB evaluation, and SGB models tend to be more sensitive to model parameters, so having separate training, tuning, and test sets is more important. With all types of models, the data that has been used to optimize the model cannot be used for the final model evaluation.

4.2. Tuning process

SGB models depend on a large number of parameters, and model performance can vary greatly depending on the values chosen for these parameters. SGB models are also vulnerable to overfitting. If final model assessments are made on the same dataset as was used to select model parameters, the true error may be considerably underestimated.

Fig. 8. Detailed maps of the eastern portion of the Utah region, showing NAIP09 imagery, RF uncertainty map, and RF and SGB predictions for tree canopy cover (TCC). The dark triangular patch in the middle of the photo is a wetland. Both RF and SGB mistook it for an area of moderate to high TCC. The RF uncertainty map shows that although the mean prediction from the 2000 trees in the model was for moderate to high TCC, there was a very high level of uncertainty in these predictions, with some trees predicting low TCC and other trees predicting much higher TCC. The adjoining agricultural area was correctly identified as low TCC by both models, most likely due to the TCC2001 and the land cover class predictor layers.



For SGB models, slower learning rates (smaller shrinkages) usually improve model performance but are computationally more expensive. There is also a point of diminishing returns at which dropping the shrinkage has less and less of an effect on model performance (Ridgeway 2007). Shrinkage is inversely related to

number of trees, so smaller values of shrinkage (slower learning rates) require more trees. Authors agree on the inverse relationship, but opinions are divided on whether this relationship scales evenly. For example, dividing the learning rate by two might or might not double the number of required trees (De'ath 2007;

Ridgeway 2007; Elith et al. 2008). For most of our regions, a shrinkage of 0.002 resulted in models with our desired number of trees (3000 to 5000 trees), although Kansas required a lower (slower) shrinkage of 0.001. Kansas is the region with the highest proportion (53%) of study sites with tree canopy cover of zero.

It has been previously found that for presence–absence data, species with very high or very low prevalence require lower values of shrinkage (Elith et al. 2008). When the majority of the data are a single value, less information is available for each tree. In an extreme case, the random subset selected for a tree could all have identical responses. A slower shrinkage allows a greater number of trees to contribute to the final model before overfitting begins to occur. For continuous response variables with high proportions of zero or 100% values, a variation of this effect could influence results in regions with large nonforest areas such as Kansas.

Very low prevalence could have a similar effect on bagging fraction. When a majority of the data has a value of zero, larger bagging fractions help make certain that the points selected for each tree contain response values other than zero. Our results are consistent with this reasoning, as we found that a higher bagging fraction (0.7) improved model predictions in Kansas.

Smaller bagging fractions introduce more stochastic uncertainty into the model and, therefore, can help counteract overfitting (Friedman 2002). Oregon performed best with a lower bagging fraction (0.2), which could suggest that Oregon is somewhat more prone to overfitting. On the other hand, it is also possible that the wide range of tree canopy cover present in Oregon meant that all values of TCC were represented, even in the smaller bagging fractions, and thus larger bagging fractions offered less of an advantage in Oregon compared with regions such as Kansas where the majority of the study sites had identical response values of zero and larger bagging fractions may have been required to assure that all subsamples contained at least some nonzero response data.

There are also potential drawbacks to increased stochasticity that need to be kept in mind when using models with low bagging fraction. (Note that faster shrinkages can also cause increased stochasticity.) For example, increased stochasticity can result in higher between-model variability (Elith et al. 2008). Overall model performance may be similar between repeated model runs, but predictions for individual locations may show high variation.

All of our regions performed best with a fairly high interaction depth. Shrinkage is inversely related to interaction depth (Elith et al. 2008). A model with more complex trees will require fewer total trees. On the other hand, complex trees make the model more vulnerable to overfitting. Therefore, it is recommended that as the tree complexity is increased, the learning rate (shrinkage) is decreased so that the model fits slower and still requires a higher number of trees. Elith et al. (2008) also notes that larger datasets can take better advantage of more complex trees. Our dataset was large enough that even our 50% training data contained nearly 2000 data points per region.

In contrast, RF models have only two parameters and are relatively insensitive to the choice of these parameters (Liaw and Wiener 2002). The *mtry* parameter usually performs well at the suggested default (one-third the number of predictor variables for regression models). Our data supports this. Two of our study regions did best with the suggested default *mtry* of 4, and in the other two, the slight improvement from the optimized *mtry* of 8 was only seen when the accuracy measures were taken to three decimal places. In most cases, RF is also less vulnerable to overfitting than SGB, so the only limit on number of trees is computation time (Breiman 2001). Therefore, with RF, tuning is less imperative than with SGB. This, combined with the possibility of using OOB estimates of model quality, means that when faced with a small dataset, RF can be used successfully without setting aside tuning or test data: simply building a RF model on the full data set with

the default parameters, and using OOB estimates of model performance.

4.3. Comparing final model performance

Our examination of relative variable importance elucidated the differences in how RF and SGB make use of correlated predictor variables. SGB had a tendency to concentrate variable importance in fewer variables, whereas RF tended to spread importance among more variables. In RF, each tree is independent, so if predictor variables are highly correlated, importance tends to be divided between the variables, with one variable important to some of the trees and the other variables important in other trees. In SGB, each successive tree builds on the previous tree, so if variables are correlated, the first variable that is randomly selected is the most important, and even if other correlated variables are chosen in later trees, there is less information that they can contribute.

Both RF and SGB models had difficulty capturing spikes in crown cover at the extremes of the distributions, either 0% or 100%. Averaging inherent in both of these modeling techniques will smooth the tails. This also indicates that global accuracy measures should be used with caution and more information can be gained by examining the observed and predicted distributions.

Because this study is part of a larger project to update and improve the 2001 NLCD product, we included tree canopy cover from the 2001 map as a predictor layer in our model. From the variable importance graphs in Fig. 4, it is clear that TCC2001 was an important variable for both the RF and the SGB models. We did experiment with models that did not include TCC2001, and surprisingly (given this predictor's high importance in models built from the full set of predictors) model performance dropped only slightly, with other predictor variables increasing in importance when TCC2001 was not available. Kansas showed the strongest loss of model quality from leaving out the TCC2001 predictor, and even there, MSE from the RF model only increased from 0.0113 to 0.0156, and the MSE from the SGB model increased from 0.0117 to 0.0157.

4.4. Comparing maps

The RF uncertainty map for Utah is interesting in that several small areas of unusually high uncertainty were found. Figures 7 and 8 examine two of these anomalies in detail. On closer inspections, both of these proved to be localized anomalies on the flat valley floors where both the RF and SGB models had predicted moderate to high crown cover. This was unusual as in the Utah region flat, low-elevation areas are most commonly sagebrush or other shrubland.

The RF uncertainty map is the standard deviation of the individual-tree predictions from the final RF model. High values of this standard deviation mean indicate a lack of agreement between the trees. Each tree in a RF model is built from a different randomly selected subset of the predictor variables, as well as a different bootstrap sample of the training locations. A high level of uncertainty thus indicates that either particular predictors or particular training locations are leading to high variation in the estimates for the response variable.

The black, irregularly shaped area in the northwest portion of Fig. 7 is a lava bed. The bright green circles to the east of the lava bed are agricultural. Both RF and SGB mistook the lava bed for an area of moderate TCC (RF predicted slightly higher cover than SGB). The RF uncertainty map shows that although the mean prediction from the 2000 trees in the model was for moderate TCC, there was a very high level of uncertainty in these predictions, with some trees predicting low TCC and other trees predicting much higher TCC.

The triangular area in the center of Fig. 8 is a wetland, the Ke Bullock Waterfowl Management Area. The rectangular green areas adjoining the wetland to the north and east are agricultural.

Both RF and SGB mistook the wetland for an area of moderate to high TCC. Again, the RF uncertainty map shows that although the mean prediction from the 1500 trees in the RF model was for moderate to high TCC, there was a very high level of uncertainty in these predictions.

In the final maps for Utah (Figs. 7 and 8), it is also interesting to look at the predictions and uncertainty for agricultural areas. In Fig. 7, the RF uncertainty map shows some moderate uncertainty in the northern portion of the adjoining agricultural areas. Although the land cover class (LC) and TCC2001 predictors would indicate that these agricultural areas are nonforest, because of the structure of RF models, some of the trees would be constructed from subsets of the predictors containing neither LC nor TCC2001, leading to higher levels of between-tree RF uncertainty. In contrast, the southern portion of the agricultural area in Fig. 7 as well as the agricultural area in Fig. 8 were correctly identified as low TCC by both models, with low values of RF uncertainty.

In addition to the lava bed, the uncertainty layer in Fig. 7 also points out a rectangular irregularity that can be traced back to the inconsistency of the national digital elevation model (DEM) layer, which composites information from both Lidar and large-scale photography, depending on which is available in different parts of the country. As a result, all of the predictors that are derivatives from the DEM (e.g., slope, elevation, aspect, and compound topographic index (CTI)) witness these irregularities and hence influence the uncertainty in the model. These irregularities may explain why the models ability to correctly discern agriculture varied in different areas of the map. It was the pattern in the uncertainty layer here that first called attention to the more subtle patterning in the predictions themselves.

5. Conclusions

RF and SGB are both powerful tree-based modeling techniques. We found that for our continuous response models, the performance of both techniques was remarkably similar on all four of our pilot regions, by all the accuracy measures that we examined. Therefore, the choice of model type may come down to ease of use.

In RF, all of the trees are independently built, whereas in SGB, each successive tree builds on the previous trees. Both provide importance measures for the predictor variables. RF is more user friendly than SGB, as it has fewer parameters to be set by the user and is less sensitive to tuning of these parameters. RF is also less prone to overfitting than SGB. In RF, using additional trees increases the time and computations but does not usually lead to loss of model quality.

RF has an OOB option for model evaluation without the necessity of setting aside an independent test set. Combine this with RF's lack of sensitivity to model parameters and it is possible to build and evaluate a model from the full dataset, without setting aside tuning or test data. This can be an important advantage over SGB, particularly for small datasets.

In contrast to RF, SGB has many parameters needing tuning, and these parameters have stronger effects on model quality. Also, overfitting is much more likely with SGB, and the number of trees that will lead to overfitting changes with the values of the other parameters.

Additionally, RF offers the possibility of a map of the stochastic uncertainty remaining in the final model, which we found valuable for identifying possibly anomalous areas in the final map.

As a result of these and other analyses, RF models are currently being used by the U.S. Forest Service Remote Sensing Applications Center (<http://www.fs.fed.us/eng/rsac/>) to produce the 2011 NLCD percent tree canopy cover dataset for the conterminous US. This dataset is publicly available at the NLCD website (<http://www.mrlc.gov/>). Percent canopy cover datasets for coastal Alaska will be finished in spring of 2015, and it is planned to have interior

Alaska, Hawaii, Virgin Islands, and Puerto Rico completed near the end of 2015.

Acknowledgements

The authors would like to acknowledge the U.S. Forest Service Forest Inventory and Analysis program for support. The authors also thank the staff at the U.S. Forest Service Remote Sensing Applications Center.

References

- Baccini, A., Laporte, N., Goetz, S.J., Sun, M., and Dong, H. 2008. A first map of tropical Africa's above-ground biomass derived from satellite imagery. *Environ. Res. Lett.* 3(4): 045011. doi:10.1088/1748-9326/3/4/045011.
- Bailey, R.G. 1995. Description of the ecoregions of the United States. USDA Forest Service Misc. Pub. 1391.
- Baker, C., Lawrence, R., Montagne, C., and Patten, D. 2006. Mapping wetlands and riparian areas using Landsat ETM+ imagery and decision-tree-based models. *Wetlands*, 26(2): 465–474. doi:10.1672/0277-5212(2006)26[465:MWARAU]2.0.CO;2.
- Bechtold, W.A., and Patterson, P.L. (Editors). 2005. The Enhanced Forest Inventory and Analysis Program — national sampling design and estimation procedures. USDA Forest Service, Southern Research Station, Asheville, North Carolina, General Technical Report SRS-80.
- Breiman, L. 2001. Random forests. *Mach. Learn.* 45: 5–32. doi:10.1023/A:1010933404324.
- Breiman, L., Friedman, R.A., Olshen, R.A., and Stone, C.G. 1984. Classification and regression trees. Wadsworth.
- Chan, J.C.W., and Paelinckx, D. 2008. Evaluation of Random Forest and Adaboost tree-based ensemble classification and spectral band selection for ecotype mapping using airborne hyperspectral imagery. *Remote Sens. Environ.* 112(6): 2999–3011. doi:10.1016/j.rse.2008.02.011.
- Chirici, G., Scotti, R., Montagni, A., Barbati, A., Cartisano, R., Lopez, G., Marchetti, M., McRoberts, R.E., Olsson, H., and Corona, P. 2013. Stochastic gradient boosting classification trees for forest fuel types mapping through airborne laser scanning and IRS LISS-III imagery. *Int. J. Appl. Earth Obs. Geoinf.* 25: 87–97. doi:10.1016/j.jag.2013.04.006.
- Cochran, W.G. 1977. Sampling Techniques. 3rd ed. John Wiley & Sons, New York.
- Coulston, J.W., Moisen, G.G., Wilson, B.T., Finco, M.V., Cohen, W.B., and Brewer, C.K. 2012. Modeling percent tree canopy cover: a pilot study. *Photogramm. Eng. Remote Sens.* 78(7): 715–727. doi:10.14358/PERS.78.7.715.
- De'ath, G. 2007. Boosted trees for ecological modeling and prediction. *Ecology*, 88(1): 243–251. doi:10.1890/0012-9658(2007)88[243:BTFFEMA]2.0.CO;2.
- De'ath, G., and Fabricius, K.E. 2000. Classification and regression trees: a powerful yet simple technique for ecological data analysis. *Ecology*, 81(11): 3178–3192. doi:10.1890/0012-9658(2000)081[3178:CARTAP]2.0.CO;2.
- Elith, J., Leathwick, J.R., and Hastie, T. 2008. A working guide to boosted regression trees. *J. Anim. Ecol.* 77: 802–813. doi:10.1111/j.1365-2656.2008.01390.x.
- Evans, J., and Cushman, S. 2009. Gradient modeling of conifer species using random forests. *Landsc. Ecol.* 24: 673–683. doi:10.1007/s10980-009-9341-0.
- Filippi, A.M., Güneralp, İ., and Randall, J. 2014. Hyperspectral remote sensing of aboveground biomass on a river meander bend using multivariate adaptive regression splines and stochastic gradient boosting. *Remote Sens. Lett.* 5(5): 432–441. doi:10.1080/2150704X.2014.915070.
- Freeman, E., and Frescino, T. 2009. ModelMap: an R package for modeling and map production using Random Forest and Stochastic Gradient Boosting. USDA Forest Service, Rocky Mountain Research Station, 507 25th street, Ogden, Utah, U.S.A. Available at <https://cran.r-project.org/web/packages/ModelMap/vignettes/VModelMap.pdf>.
- Frescino, T.S., and Moisen, G.G. 2012. Comparing alternative tree canopy cover estimates derived from digital aerial photography and field-based assessments. In *Proceedings of Monitoring Across Borders: 2010 Joint Meeting of the Forest Inventory and Analysis (FIA) Symposium and the Southern Mensurationists*. Edited by W. McWilliams and F.A. Roesch. USDA Forest Service, Southern Research Station, Asheville, North Carolina, e-Gen. Tech. Rep. SRS-157. pp. 237–244.
- Friedman, J.H. 2001. Greedy function approximation: a gradient boosting machine. *Ann. Stat.* 29(5): 1189–1232. doi:10.1214/aos/1013203451.
- Friedman, J.H. 2002. Stochastic Gradient Boosting. *Computational Statistics and Data Analysis*, 28: 367–78. doi:10.1016/S0167-9473(01)00065-2.
- Friedman, J., Hastie, T., and Tibshirani, R. 2000. Additive logistic regression: a statistical view of boosting. *Ann. Stat.* 28(2): 337–407. doi:10.1214/aos/1016218223.
- Gislason, P.O., Benediktsson, J.A., and Sveinsson, J.R. 2006. Random Forests for land cover classification. *Pattern Recognit. Lett.* 27(4): 294–300. doi:10.1016/j.patrec.2005.08.011.
- Güneralp, İ., Filippi, A.M., and Randall, J. 2014. Estimation of floodplain aboveground biomass using multispectral remote sensing and nonparametric modeling. *Int. J. Appl. Earth Obs. Geoinf.* 33: 119–126. doi:10.1016/j.jag.2014.05.004.
- Hastie, T., Tibshirani, R., and Friedman, J. 2009. Model assessment and selection. In *The elements of statistical learning*. Springer Series in Statistics, Springer, New York. pp. 219–259.
- Hauke, J., and Kossowski, T. 2011. Comparison of values of Pearson's and Spear-

- man's correlation coefficients on the same sets of data. *Quaestiones Geographicae*, **30**(2): 87–93. doi:10.2478/v10117-011-0021-1.
- Homer, C., Huang, C., Yang, L., Wylie, B., and Coan, M. 2004. Development of a 2001 National Landcover Database for the United States. *Photogramm. Eng. Remote Sens.* **70**(7): 829–840. doi:10.14358/PERS.70.7.829.
- Homer, C., Dewitz, J., Fry, J., Coan, M., Hossain, N., Larson, C., Herold, N., McKerrow, A., VanDriel, J.N., and Wickham, J. 2007. Completion of the 2001 National Land Cover Database for the conterminous United States. *Photogramm. Eng. Remote Sens.* **73**(4): 337–341.
- Jackson, T.A., Moisen, G., Patterson, P.L., and Tipton, J. 2012. Repeatability in photo-interpretation of tree canopy cover and its effect on predictive mapping. In *Monitoring Across Borders: 2010 Joint Meeting of the Forest Inventory and Analysis (FIA) Symposium and the Southern Mensurationists*. Edited by W. McWilliams and F.A. Roesch. USDA Forest Service, Southern Research Station, Asheville, North Carolina, e-Gen. Tech. Rep. SRS-157. pp. 189–192.
- Jennings, S.B., Brown, N.D., and Sheil, D. 1999. Assessing forest canopies and understory illumination: canopy closure, canopy cover and other measures. *Forestry*, **72**(1): 59–73. doi:10.1093/forestry/72.1.59.
- Kellndorfer, J.M., Walker, W., LaPoint, E., Hoppus, M., and Westfall, J. 2006. Modeling height, biomass, and carbon in US forests from FIA, SRTM, and ancillary national scale data sets. In *Proceedings of the IEEE International Symposium on Geoscience and Remote Sensing (IGARSS)*, July 31 – August 4, 2006, Denver, Colorado. pp. 3591–3594.
- Lawrence, R., Bunn, A., Powell, S., and Zambon, M. 2004. Classification of remotely sensed imagery using stochastic gradient boosting as a refinement of classification tree analysis. *Remote Sens. Environ.* **9**(3): 331–336. doi:10.1016/j.rse.2004.01.007.
- Lawrence, R.L., Wood, S.D., and Sheley, R.L. 2006. Mapping invasive plants using hyperspectral imagery and Breiman Cutler classifications (Random Forest). *Remote Sens. Environ.* **100**: 356–362. doi:10.1016/j.rse.2005.10.014.
- Leathwick, J.R., Elith, J., Francis, M.P., Hastie, T., and Taylor, P. 2006. Variation in demersal fish species richness in the oceans surrounding New Zealand: an analysis using boosted regression trees. *Mar. Ecol. Prog. Ser.* **321**: 267–281. doi:10.3354/meps321267.
- Liaw, A., and Wiener, M. 2002. Classification and regression by random forest. *R News*, **2**: 18–22. Available from <http://CRAN.R-project.org/doc/Rnews/>.
- McRoberts, R.E., Holden, G.R., Nelson, M.D., Liknes, G.C., and Gormanson, D.D. 2005. Using satellite imagery as ancillary data for increasing the precision of estimates for the Forest Inventory and Analysis program of the USDA Forest Service. *Can. J. For. Res.* **35**(12): 2968–2980. doi:10.1139/x05-222.
- Moisen, G.G. 2008. Classification and regression trees. In *Encyclopedia of Ecology*. Vol. 1. Edited by S.E. Jørgensen and B.D. Fath. Elsevier. pp. 582–588.
- Moisen, G.G., and Frescino, T.S. 2002. Comparing five modelling techniques for predicting forest characteristics. *Ecol. Model.* **157**: 209–225. doi:10.1016/S0304-3800(02)00197-7.
- Moisen, G.G., Freeman, E.A., Blackard, J.A., Frescino, T.S., Zimmermann, N.E., and Edwards, T.C., Jr. 2006. Predicting tree species presence and basal area in Utah: a comparison of stochastic gradient boosting, generalized additive models, and tree-based methods. *Ecol. Model.* **199**: 176–187. doi:10.1016/j.ecolmodel.2006.05.021.
- Moisen, G.G., Coulston, J.W., Wilson, B.T., Cohen, W.B., and Finco, M.V. 2012. Choosing appropriate subpopulations for modeling tree canopy cover nationwide. In *Proceedings of Monitoring Across Borders: 2010 Joint Meeting of the Forest Inventory and Analysis (FIA) Symposium and the Southern Mensurationists*. Edited by W. McWilliams and F.A. Roesch. USDA Forest Service, Southern Research Station, Asheville, North Carolina, e-Gen. Tech. Rep. SRS-157. pp. 195–200.
- Moore, I.D., Grayson, R.B., and Ladson, A.R. 1991. Digital terrain modelling: review of hydrological, geomorphological, and biological applications. *Hydrol. Processes*, **5**: 3–30. doi:10.1002/hyp.3360050103.
- Nowak, D.J., Crane, D.E., and Stevens, J.C. 2006. Air pollution removal by urban trees and shrubs in the United States. *Urban Forestry and Urban Greening*, **4**(3–4): 115–123. doi:10.1016/j.ufug.2006.01.007.
- Pittman, S.J., Costa, B.M., and Battista, T.A. 2009. Using Lidar bathymetry and boosted regression trees to predict the diversity and abundance of fish and corals. *J. Coastal Res., Spec. Issue*, **53**: 27–38. doi:10.2112/SI53-004.1.
- Powell, S.L., Cohen, W.B., Healey, S.P., Kennedy, R.E., Moisen, G.G., Pierce, K.B., and Ohmann, J.L. 2010. Quantification of live aboveground forest biomass dynamics with Landsat time-series and field inventory data: a comparison of empirical modeling approaches. *Remote Sens. Environ.* **114**(5): 1053–1068. doi:10.1016/j.rse.2009.12.018.
- Prasad, A.M., Iverson, L.R., and Liaw, A. 2006. Newer classification and regression tree techniques: bagging and random forests for ecological prediction. *Ecosystems*, **9**: 181–199. doi:10.1007/s10021-005-0054-1.
- R Development Core Team. 2008. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Available from <http://www.R-project.org>. ISBN 3-900051-07-0.
- Ridgeway, G. 2007. Generalized boosted models: a guide to the gbm package. Available from <https://cran.r-project.org/web/packages/gbm/index.html>.
- Ridgeway, G., et al. 2013. gbm: Generalized boosted regression models. R package version 2.1. Available from <https://cran.r-project.org/web/packages/gbm/index.html>.
- Rollins, M.G., and Frame, C.K. (Editors). 2006. The LANDFIRE Prototype Project: nationally consistent and locally relevant geospatial data for wildland fire management. USDA Forest Service, Rocky Mountain Research Station, Fort Collins, Colorado, General Technical Report RMRS-GTR-175.
- Sankaran, M., Ratnam, J., and Hanan, N. 2008. Woody cover in African savannas: the role of resources, fire and herbivory. *Glob. Ecol. Biogeogr.* **17**: 236–245. doi:10.1111/j.1466-8238.2007.00360.x.
- Tipton, J., Moisen, G., Patterson, P., Jackson, T.A., and Coulston, J. 2012. Sampling intensity and normalizations: exploring cost-driving factors in nationwide mapping of tree canopy cover. In *Proceedings of Monitoring Across Borders: 2010 Joint Meeting of the Forest Inventory and Analysis (FIA) Symposium and the Southern Mensurationists*. Edited by W. McWilliams and F.A. Roesch. USDA Forest Service, Southern Research Station, Asheville, North Carolina, e-Gen. Tech. Rep. SRS-157. pp. 201–208.
- Toney, C., Liknes, G., Lister, A., and Meneguzzo, D. 2012. Assessing alternative measures of tree canopy cover: photo-interpreted NAIP and ground-based estimates. In *Proceedings of Monitoring Across Borders: 2010 Joint Meeting of the Forest Inventory and Analysis (FIA) Symposium and the Southern Mensurationists*. Edited by W. McWilliams and F.A. Roesch. USDA Forest Service, Southern Research Station, Asheville, North Carolina, e-Gen. Tech. Rep. SRS-157. pp. 209–215.
- U.S. Department of Agriculture (USDA). 2009. National Agriculture Imagery Program. USDA Farm Service Agency, Aerial Photography Field Office, Salt Lake City, Utah. Available from <http://www.apfo.usda.gov/FSA/apfoapp?area=home&subject=prog&topic=nai> [accessed 29 March 2012].
- Webb, B.W., and Crisp, D.T. 2006. Afforestation and stream temperature in a temperate maritime environment. *Hydrol. Processes*, **20**(1): 51–66. doi:10.1002/hyp.5898.
- White, D., Kimerling, J., and Overton, S.W. 1992. Cartographic and geometric components of a global sampling design for environmental monitoring. *Cartographic and Geographic Information Systems*, **19**(1): 5–22. doi:10.1559/152304092783786636.