12-2015

Linking remote sensing and various site factors for predicting the spatial distribution of eastern hemlock occurrence and relative basal area in Maine, USA

Kathleen Dunckel
Unity College

Aaron Weiskittel
University of Maine, aaron.weiskittel@maine.edu

Greg Fiske

Steven A. Sader
University of Maine

Erika Latty
Unity College

See next page for additional authors

Submitted to Forest Ecology and Management

Linking remote sensing and various site factors for predicting the spatial distribution of eastern hemlock occurrence and relative basal area in Maine, USA

Kathleen Dunckel 1,2, Aaron Weiskittel 2, Greg Fiske 3, Steven A. Sader 2, Erika Latty 1, Amy Arnett 1

1 Unity College, Center for Natural Resource Management and Protection, 90 Quaker Hill Rd., Unity, ME 04988, United States

2 University of Maine, School of Forest Resources, 201 Nutting Hall, Orono, ME 04469, United States

3 The Woods Hole Research Center, 149 Woods Hole Rd., Falmouth, MA 02540, United States

ABSTRACT

Introduced invasive pests are perhaps the most important and persistent catalyst for changes in forest composition. Infestation and outbreak of the hemlock woolly adelgid (Adelges tsugae; HWA) along the eastern coast of the USA, has led to widespread loss of hemlock (Tsuga canadensis (L.) Carr.), and a shift in tree species composition toward hardwood stands.

Developing an understanding of the geographic distribution of individual species can inform conservation practices that seek to maintain functional capabilities of ecosystems. Modeling is necessary for understanding changes in forest composition, and subsequent changes in biodiversity, and one that can be implemented at the species level. By integrating the use of remote sensing, modeling, and Geographic Information Systems (GIS) coupled with expert knowledge in forest ecology and disturbance, we can advance the methodologies currently available in the literature on predictive modeling.

This paper describes an approach to modeling the spatial distribution of the less common but foundational tree species eastern hemlock throughout the state of Maine (~84,000 km2) at a high resolution. There are currently no published accuracy assessments on predictive models for high resolution continuous distribution of eastern hemlock relative basal area that span the geographic extent covered by our model, which is at the northern limit of the species' range. A two stage mapping approach was used where presence/absence was predicted with an overall accuracy of 85% and the continuous distribution (percent basal area) was predicted with an accuracy of 84%.

Overall, these findings are quite good despite high variability in the training dataset and the general minor component that eastern hemlock represents in the primary forest types in Maine.

Eastern hemlock occurs along the southern half of the state stretching the east-west span with little to no occurrence in the northern regions. Several environmental and site characteristics, particularly average yearly maximum and minimum temperatures, were found to be positively correlated with hemlock occurrence. Eastern hemlock dominated stands appeared predominantly in the southwest corner of the state where HWA monitoring efforts can be focused. Given the importance of climate variables in predicting eastern hemlock, forecasts of future range shifts should be possible using data generated from climate scenarios.

INTRODUCTION

Motivation

We are experiencing a significant loss in biodiversity worldwide, this is considered to be important for a variety of reasons (Randall, 1991; Rolston, 2000), but recent attention has focused on its potential importance for the adequate functioning of the Earth's ecosystems (Schulze and Mooney, 1994; Heywood and Watson, 1995). Forest ecosystems are losing biodiversity through a variety of disturbances that are numerous, including land use, climate change, fire, and wind. In particular, invasive introduced species are disturbance agents to which temperate forests appear to have relatively little resistance (Richardson, 1998, but see Simberloff et al., 2002 for a treatment of tropical forests). As many as 19 introduced insect pests and pathogens are causing changes to forest structure, species composition, and ecosystem function of North American forests and it is anticipated that a warming climate will amplify the effects of these forest pests (Dukes et al., 2009).

In the northeastern United States, mean annual temperatures have increased by 0.8 C over the last century with estimates that they will continue to increase from 2.1 to 5.3 C by 2100 (Campbell et al., 2009). Changes in our climate will precipitate changes in biogeochemical cycling, resulting in potentially dramatic changes in forest composition and productivity (Weiskittel et al., 2011). These climate driven alterations will be coupled with the effects of other disturbances such as forest pathogens and forest management strategies (e.g., harvesting).

Consequently, it is important to understand and forecast both current and future potential species habitat.

In particular, New England forests are currently experiencing a decline in eastern hemlock (Tsuga canadensis (L.) Carr.) due to the hemlock woolly adelgid (Adelges tsugae; HWA), an invasive, aphid-like pest introduced to the United States from Asia (Ellison et al., 2005). HWA can be found in 15 states along the eastern seaboard from Georgia to Maine (Stadler et al., 2005) including several counties in southern and mid-coast Maine (Maine Forest Service, 2014). Albani et al. (2010) predict that HWA will continue to move northward and will be established throughout the eastern hemlock range in Maine in the next 30 years.

Eastern hemlock is a late successional conifer that, because of its deep shade and acidic litter, shapes stand microclimate and influences community and ecosystem characteristics (Eschtruth et al., 2006; Orwig et al., 2002). This strong influence on microclimate affects vegetation organization, successional dynamics, species diversity, and microenvironmental characteristics (Orwig and Foster, 1998). Eastern hemlock dominated forests represent unique characteristics that serve as critical wildlife habitat (Orwig et al., 2002).

Predicting eastern hemlock occurrence is complicated by the fact that it is difficult to distinguish from other conifers by spectral response alone (Doucette et al., 2009) particularly in mixed conifer stands, which is where it tends to occur in Maine. Ancillary GIS data representing

environmental characteristics are often used in conjunction with satellite imagery to define patterns in vegetation cover (Kong et al., 2008). Narayanaraj et al. (2010) found strong relationships between eastern hemlock density (# ha-1), basal area (m2 ha-1) and elevation, distance to streams, and soil moisture. Boyce (2000) also found a strong relationship between the location of eastern hemlock trees and elevation, slope, and NW aspect. These previous analyses highlight the importance of topographic variables in describing the distribution of eastern hemlock, but acknowledge the diverse array of factors that influence it. Given the geographic extent of the study area in this present study, several potential predictors should be evaluated to find the most robust model. A working hypothesis for this analysis was that remote sensing, climatic, and topographic variables would be equally important for predicting both eastern hemlock occurrence and abundance.

Objectives

Developing an understanding of the geographic distribution of individual tree species can inform conservation practices that seek to maintain biodiversity of ecosystems. Mapping eastern hemlock in Maine will be crucial to response efforts by anticipating where HWA infestations will occur. The methodology developed and used here can also be applied to other species level inquiries in northern forests. The primary objectives for this paper were to: (1) predict the occurrence of a less common tree species, eastern hemlock across a large geographic extent that includes the species' northern range, (2) predict percent basal area of eastern hemlock where it occurs, and (3) map the species occurrence and percent basal area at a high resolution.

METHODS

Study area

The state of Maine (~84,000 km2) is located in the northeast corner of the New England region of the U.S. It is bordered by the Canadian provinces of Quebec and New Brunswick, the Atlantic Ocean, and by New Hampshire (see Fig. 1). It falls within mapping zone 66 (42 580 N to 47 280 N and 66 570 W to 71 50 W). Maine is nearly 90% forested and dominated by mixed northern hardwood stands comprised of over 62 tree species. The most prevalent of these species being balsam fir (Abies balsamea L.) Mill.), red maple (Acer rubrum L.), red (Picea rubens Sarg.), white (Picea glauca (Moench) Voss), and black spruce (Picea mariana (Mill.) B.S.P.), sugar maple (Acer saccharum Marsh.), yellow birch (Betula alleghaniensis Britton), and American beech (Fagus grandifolia Ehrh.) (Maine Forest Service 2013). Maine forests are bordered by Boreal Forests in the north and Central Hardwood Forests to the south and are sometimes referred to as ‘‘Acadian Forests”.

Spatial database

A relatively comprehensive spatial database was compiled for the state of Maine comprised of satellite imagery, digital elevation models, and ancillary GIS data. Most of these data can be found and downloaded from the Natural Resources Conservation Services (NRCS) Geospatial

Data Gateway (http://datagateway.nrcs. usda.gov). Over 30 different spatial data layers depicting ecological (e.g., biomass) or environmental (e.g., precipitation) phenomenon were explored to find good predictors for eastern hemlock. Predictor variables were selected by evaluating the coefficient of determination and variable importance plots produced by the random Forest algorithm with a threshold mean decrease in accuracy value of 20%. Preliminary analysis indicated eleven predictors that were most influential in describing eastern hemlock distribution, these were used in the final model and are described in detail below.

Remote Sensing and Google Earth Engine

The use of satellite imagery and remote sensing in ecological and resource management research has been increasing steadily since the 1990s (Fassnacht et al., 2006). Multi-spectral imagery is helpful in discerning land cover types as different wavelengths of electromagnetic energy are reflected differently from different types of land cover. In choosing a remote sensing system some important considerations for researchers are: (1) cost, (2) temporal resolution (frequency of image acquisition), (3) spatial resolution (cell size), and (4) spectral resolution (number of wavebands detected). We chose Landsat-5 thematic mapper (TM), which is, arguably, the most popular mid resolution, passive remote sensor used in natural resource applications prior to the launch of Landsat 8 OLI in February 2013, a similar sensor that has replaced TM. For the study area the size of Maine, Landsat TM offers cost-effective imagery that has a revisit period of 16 days, 30 m pixel resolution (cell size), and is multi-spectral with 7 wavebands detected (3 visible,

1 near infrared, 2 mid-infrared, 1 thermal infrared). An individual Landsat TM scene covers approximately 26,000 km2.

For complete coverage of the state of Maine, nine individual scenes are needed from 3 paths and 4 rows (listed here as Word-wide Reference System path/row: 10/29, 11/27, 11/28, 11/29, 11/30, 12/27, 12/28, 12/29, 12/30). Acquiring Landsat TM imagery for the state of Maine would generally involve hours of searching for scenes and assessing both their cloud content and date/time characteristics. The use of multiple Landsat scenes would also require radiometric normalization and geometrically co-referencing to minimize mosaic seam lines (Cohen et al., 2001). With the advent of the Google Earth Engine much of this pre- processing is already completed and searching the Landsat database can be streamlined. Google's Earth Engine is a cloud-based platform that allows planetary level data storage, mining, and analysis (https://earthengine.google.org). The platform allows unprecedented speed and efficient computing by making use of Google's existing computing infrastructure, it is currently freely available to non-profit and educational institutions, and access is available through both a Javascript and a Python application programming interface (API). Within Earth Engine's data archive exists all publicly available digital Landsat data. These data were provided by the EROS data center in Souix Fall, SD, and georeferenced/corrected by Google using standard image processing routines. Using Earth Engine's processing capabilities, we produced a Landsat TM cloud-free mosaic for the state Maine by compositing the best, cloud-free, pixels through time for the years 2007–2011 during both summer (leaf-on) and winter (leaf-off) seasonal conditions.

Tasseled cap is a spectral vegetation index (SVI) derived from satellite imagery and originally introduced by Kauth and Thomas (1976), that includes brightness, greenness, and wetness (designed to describe water content in soil) (Crist and Cicone, 1984). All three transformations were evaluated in this analysis, only tasseled cap greenness was used in the predictive model.

Soil Survey Geographic (SSURGO) dataset

Soil moisture has been identified as a good predictor of hemlock abundance (Narayanaraj et al., 2010). The SSURGO dataset contains available water storage variables at varying depths (25 cm, 50 cm, 100 cm, 150 cm), which accounts for soil texture and coarse fragment content. All of these depths were assessed as potential predictors, but only 25 cm was used in model calibrations.

Climate

Climate variables including average annual precipitation, average annual maximum temperature, and average annual minimum temperature have been estimated by the Oregon State University's PRISM Climate Group (http://www.prism.oregonstate.edu/). These data represent 30 year normal estimates for 1981–2010 in rasterdata (grid) at a resolution of ~800 m. Estimates of climatic parameters were derived using analytical models that incorporate point data and a digital elevation model (DEM). Although these data are at a much coarser scale than the other eight

predictors (10–30 m; see Table 1), it was decided to make predictions at 30 m resolution reasoning that climate variables do not vary drastically at finer scales.

National biomass and carbon dataset

Scientists at the Woods Hole Research Center (WHRC) developed a 30 m resolution GIS dataset describing forest height, above ground biomass and carbon stock in the conterminous U.S. for the year 2010 (WHRC, 2012). We used above ground biomass in this analysis.

Reference data

A reference database was compiled that includes forest measurements from the Penobscot Experimental Forest (PEF), the Maine Forest Service (MFS), the Hemlock Ecosystem Management Study (HEMS), and US Forest Service, Forest Inventory and Analysis (FIA) program to identify stand composition of geographically defined study plots located across Maine (see Table 2). Each dataset is briefly described below.

USFS Forest Inventory and Analysis (FIA) program

The US Forest Service (USFS) conducts a comprehensive Forest Inventory and Analysis (FIA) program across the US. This program consists of long-term forest monitoring research plots

established through stratified systematic sampling across public and private lands. Each plot consists of four 1/24th acre (0.02 ha) fixed- radius (24.0 ft/7.3 m) subplots. Plot coordinate accuracy is reported to be within 10 m of the actual site Hoppus and Lister, 2005). Although actual FIA plot coordinates are not publicly available, we utilized them here as this study was sponsored with an MOU by the Northern Research Station FIA program. Field data are collected at the plot-level and include forest type, tree species, tree diameter at breast height (dbh), tree height, and condition (for more information on the FIA program and a complete list of data collected please see www.fia.fs.fed.us). The data used in this study are taken from forested plots in Maine (n = 2607) that were measured during an inventory cycle from 2006 to 2010 (see Fig. 1).

Penobscot Experimental Forest (PEF)

The PEF is located in the towns of Bradley and Eddington, Maine. This forest is dominated by mixed northern conifers, including eastern hemlock and is the site of various research endeavors making forest composition well documented. The data used here were taken during surveys from 2000 to 2009 on 807 m2 (0.08 ha) fixed radius plots (n = 502).

Maine Forest Service (MFS)

These data were taken from HWA impact plots from hemlock dominated stands in York, Cumberland, and Lincoln counties. Data were recorded at each of the five sites within three 400m² (0.04 ha) circular plots (n = 15) located on a transect line oriented along the central axis of the stand.

Hemlock Ecosystem Management Study (HEMS)

HEMS is a multi-year study of the ecological and socioeconomic implications of changes in mixed hemlock-hardwood forests being conducted by faculty at Unity College in Unity, Maine. These data consist of 100 m2 (0.01 ha) experimental plots in hemlock dominated stands spread across four sites in Waldo County, Maine (n = 33). For all plot data sources, combined total basal area (m2) and percent eastern hemlock basal area (m2) were calculated for each observation used (n = 3157) (Tables 2 and 3).

Data extraction

Because of the spatial layout of FIA plots (described above), a 3 x 3 pixel neighborhood window around the first subplot was used to calculate and extract values from the predictor spatial data layers in order to capture all four subplots. The center points from the other field reference plots (one plot per point) were used to extract values from the predictor spatial data layers within ArcGIS software v10.1 by ESRI. Spatial joins were used to combine attributes into a single

database for model development. A subset containing 10% (n = 312) of reference points were randomly selected and reserved for map validation (see Fig. 2).

Regression trees

Regression trees are a deviation from the linear model that allow more flexible regression modeling of the response variable by combining predictors in a nonparametric approach. Trees are formed by partitioning an individual variable (predictor) at a node along the range of that variable. A simple average is taken in that partition so that: residual sum of squares (RSS) (partition) = RSS (part1) + RSS (part2) and the partition that has the smallest RSS is chosen for the tree ensemble. These partitions are partitioned again recursively to construct expert trees (Faraway, 2006).

Multivariate tree-based regression models have evolved with improved accuracy in predicting forest structural attributes (Walker et al., 2007). These improvements include processes for bagging (from bootstrap aggregate) and boosting. Bagging involves sampling the dataset with replacement to produce replicate training sets. In this way, all cases are used to construct each tree with more weight placed on cases that are more difficult to predict. Boosting assigns different weights (voting strengths) to trees based on accuracy.

Random forest

A regression tree is best described by the algorithm used to construct it. The random forest algorithm, first introduced by Briemen (2001), draws a bootstrap sample Z of size N from the training data set (with replacement) so that each tree is constructed with a random subset of the data. Out of bag (OOB) predictions of error are calculated from reserved cases (about 1/3rd) from each training data set. The best split among the random subset of predictors is chosen at each node and recursively until the minimum node size is reached. The output is then an ensemble of trees. This process reduces prediction variance of trees, decreases bias (if the trees are sufficiently deep) and increases accuracy through decorrelation. Random forest will not over fit data and can compute variable importance which is an advantage over other modeling techniques (Briemen, 2001). For this analysis, the 'random Forest' package (Liaw and Weiner, 2007) in R v 3.1.3 (R Core Team, 2015).

Model evaluation

Presence/absence

The predictor variables can be organized into five broad categories: climate, satellite imagery, biomass, elevation, and soils (Table 1). Predictor variables were assessed for importance using variable importance plots. For the presence/absence model, the Gini coefficient and mean decrease in accuracy were used. The Gini coefficient is a measure of homogeneity from 0 (homogeneous) to 1 (heterogeneous). Variables that result in nodes with higher purity have a higher decrease in Gini coefficient. The mean decrease in accuracy measures the accuracy of the

model if that variable were to be removed. The resulting binary model was assessed using a Kappa statistic and the Area Under the Curve (AUC). The Kappa statistic (or value) is a metric that measures whether the observed accuracy could be a result of random chance, where a Kappa of 1 indicates perfect agreement and a kappa of 0 indicates agreement equivalent to chance. The AUC is generated from a Receiver Operator Curve (ROC) that assesses model sensitivity and specificity with an AUC of 1 being a perfect model and an AUC of less than 0.5 being a poor model.

Continuous model

The predictor variables for the continuous model remained the same and were evaluated using variable importance plots with measures of percent increase in mean square error in the model (in absence of a given variable) and node purity. The percent increase in mean square error is a permutation (i.e., based on sampling without replacement) calculated during the OOB error phase and considers how the removal of individual predictor variables degrades prediction accuracy (Boulesteix et al., 2012). The more important the variable is in classification, the larger the mean decrease in accuracy. Node purity is based on the criteria used to split the nodes (important predictors are often selected for splitting) so that node purity is calculated by taking the total decrease in node impurities from splitting on each variable, averaged over all trees measured by the residual sum of squares (Boulesteix et al., 2012).

The continuous model was evaluated using coefficient of determination (R2) and root mean square error (RMSE). The average and standard error were calculated on a bootstrap sample of 500 with 100 repetitions where confidence intervals were also obtained. In general, the most parsimonious and robust model was selected for predictions.

Map production

Given the skewed distribution of eastern hemlock basal area in reference plots (see Tables 1 and 2), and the desire to produce an accurate continuous distribution, a two-step map making process was adopted. First, a map depicting presence/absence of eastern hemlock at 30 m resolution was produced in R (ModelMap package) using the categorical model described above. All raster data layers representing the 11 predictor variables were projected to WGS 84 and geographic extents were snapped. Raster data layers depicting climate, soils, and elevation (see Table 1) were resampled to 30 m resolution using bilinear interpolation. This map was then used as a mask where only presence pixels were used to produce a 30 m resolution map of percent basal area eastern hemlock using the continuous model described above (see Fig. 2).

RESULTS

A total number of 3157 reference plots (Fig. 1) were available for analysis with an average basal area eastern hemlock of 10.1% (Tables 2 and 3). Of these reference plots 2815 were used in model calibration and 312 were used in the final map accuracy assessment.

Presence/absence

The predicted probability of occurrence was consistent with the reference data (Fig. 4a and b). The Area Under the Curve (AUC) for this categorical model was 0.91 with 95% confidence interval of 0.90–0.92 (Fig. 4c). The Kappa statistic was 0.65 and 85% of pixels were classified correctly at a 30 m resolution (Fig. 4d).

The 30-year normal average maximum temperature was clearly the most important predictor variable used in the model (see Fig. 3). Hemlock present plots had an average maximum temperature of 12.2 C, whereas hemlock absent plots had an average maximum temperature of 10.6 C. Elevation and winter satellite imagery representing the visible red band were also consistently among the most important predictor variables. The average elevation differed between the two groups by more than 46 m with hemlock present plots having the lower average. The visible red imagery taken during the winter (leaf-off) had a lower average reflectance digital number (i.e., DN, more visible red being absorbed, less being reflected) for hemlock present plots than hemlock absent plots.

The random forest model predicts presence of eastern hemlock with a threshold average maximum temperature of 10.6 C below which eastern hemlock ceased to be predicted. A similar

threshold exists for minimum temperatures of 2.8 C. The probability of the model predicting eastern hemlock occurrence decreased at elevations greater than 91.4 m (see Fig. 7).

Continuous model

Temperature continues to be the most important predictor variable for the hemlock model as well as winter visible red imagery (see Fig. 5). Elevation was also an important predictor for the continuous model. Average maximum and minimum temperature, have positive statistically significant correlations (at the .001 level) with hemlock abundance (Pearson correlation coefficients: 0.43, 0.38 respectively). Elevation and winter red imagery have negative statistically significant correlations (at the .001 level) with hemlock abundance (Pearson correlation coefficients: -0.36, -0.22 respectively; see Fig. 7).

The continuous model explained 57% of the variation in the dependent variable, but there was some variability in model bias. The continuous predictive model appeared to have the necessary variance to produce a map product with acceptable local accuracy. Fig. 6 displays observed versus predicted values of percent basal area for training reference plots (n = 2815). A RMSE of less than 13% was obtained.

Overall map accuracy

The overall accuracy of the final map depicting the continuous distribution of eastern hemlock density (basal area) in Maine was 84% (see Fig. 8). Map accuracy was also tested using reference plots (n = 312). Fig. 9 shows the predicted versus observed values for the continuous distribution of eastern hemlock basal area. The R2 value obtained (0.56) was similar to that of the continuous model calibrated with 2815 reference plots. Of the predicted values for pixels in the final map product, 84% were within ±13% of the observed value and 92% were within ±26% of the observed values.

The two stage mapping approach had a positive impact on the continuous map product by removing the prediction of low level hemlock abundance throughout the entirety of the state. The presence mask reduced the geographic extent of predicted eastern hemlock by 25%.

Predicted distribution

The binary model predicted hemlock occurrence in approximately 20% of pixels that cover the state of Maine at 30 m resolution with a geographic extent that covers 75% of the state. Within those presence pixels, eastern hemlock abundance (% basal area) was predicted with a range of 2–78% with a mean of 18% (see Fig. 8). Eastern hemlock trees occur along the southern half of the state stretching the east-west span with little to no occurrence in the northern regions with a northern extent of 46 520 N. Eastern hemlock dominated stands appeared predominantly in the south-west corner of the state (see Fig. 8). The continuous model predicts the highest abundance

(% basal area) of eastern hemlock in areas with an average yearly maximum temperature P12.8 C and average yearly minimum temperature of P1.7 C (see Fig. 7).

DICUSSION

Overall, our model indicated that eastern hemlock occurrence and abundance could be effectively modeled using remote sensing, climate, and soil variables. In general, the climate and remote sensing variables were the most important, which rejects our null hypothesis that topographic variables would be equally important. In fact, the primary environmental variables (i.e., slope, aspect, and distance to streams) that have been associated with eastern hemlock occurrence in past studies (Narayanaraj et al., 2010; Boyce, 2000) were not found to be important predictors. This may be a function of scale and particularly geographic extent. However, elevation tended to be an important factor even after accounting for climate and soils, which would suggest that topography has some influence on eastern hemlock occurrence.

The natural variation in the physical structure of an area increases with geographic extent. The geographic extent of this study (~84,000 km2) no doubt contributed to the low R2 values obtained by the continuous predictive model. Landscape analyses deal with heterogeneous land areas with interacting ecosystems. In this case, state boundaries are arbitrary to ecosystem interactions and are used here to define the landscape under investigation for practical and political considerations. Field data are generally measured at the stand level and Roberts et al. (2004) found that plant species were the least distinct at the stand scale using remotely sensed

structural measures (e.g., NDVI). Stand level tree species predictions have historically had lower R2 values (see Song et al., 2007; Cohen et al., 2001), which illustrates the necessity that the results of large scale modeling and mapping projects be reviewed with other accuracy metrics (e.g., map product accuracy) and the intended use of the product.

Species level forest cover predictions, and in particular eastern hemlock, have been undertaken in the past with varying results. The most common prediction of eastern hemlock is that of presence/absence with more limited studies for abundance. Two studies used decision tree models to predict eastern hemlock presence in the southeast USA with overall accuracies reported to be 70% (see Clark et al., 2012; Kong et al., 2008). A similar study using fuzzy boundary accuracy assessment reports 70-80% accuracy in hemlock presence classification (see Koch et al., 2005). These values are generally consistent with the findings of this analysis.

For abundance, Pontius et al. (2005) predicted percent basal area eastern hemlock across 2800 km2 using linear regression and Mixture Tuned Matched Filtering (MTMF;used to quantify the hemlock component of each pixel) with an R2 of 0.65 and RMSE of 12%. Scientists from the USFS Northern Research Station created a suite of species specific maps depicting continuous(% basal area) distribution at coarse (250 m) resolution from FIA data that are publicly available for download (see Wilson et al., 2013). Wilson et al. (2012) used MODIS imagery and environmental parameters to predict tree species abundance using a weighting of nearest neighbors (k-nearest neighbor and canonical correspondence). The assessment of eastern hemlock prediction reports an R2 value (at 25 km scale) of 0.72 and RMSE of1.63 ft2/acre. As

the authors point out, these models are less accurate for less frequent tree species and species at the limits of their ranges, which both apply to eastern hemlock in Maine. This can be seen in what is likely over-estimation in the northern limits of the eastern hemlock range in Maine. The map presented in this analysis does not predict any occurrence of eastern hemlock in this area.

CONCLUSION

In this study, the continuous distribution of eastern hemlock basal area was predicted and mapped at high resolution with a high level of accuracy (see Fig. 8). The products and methods used here are cost-effective and accessible to the public.

The two stage mapping approach had a positive impact on the continuous map product by removing the prediction of low level hemlock abundance throughout the entirety of the state. This effectively reduced the geographic extent of predicted eastern hemlock by 25%, which is more consistent with the observed data. We believe that this two-stage approach is useful when mapping a less frequent tree species where masking out absence pixels will help minimize over predictions of occurrence.

A primary limitation of using a nonparametric approach (i.e., random Forest) in modeling a continuous variable (i.e., percent basal area) is that the model cannot extrapolate beyond the observed values. This is evident in Fig. 8 where the predicted values are truncated at 78%. This

is not a major concern for this particular application as land managers will likely be interested in monitoring stands that are hemlock dominated (e.g., P50% eastern hemlock basal area).

The decision to map continuous data (% basal area) as opposed to categorical (classes of %basal area) data was made so as not to impose an artificial class structure (Fassnacht et al., 2006) and more accurately represent the true spatial distribution of eastern hemlock (Cohen et al., 2001). Representing this information with continuous data also enables other projects to use these data and impose their own classes as needed. An example of this is the management of HWA in Maine. Land managers may be particularly interested in monitoring hemlock dominated stands. Those decision makers can impose the limits that they qualify as hemlock dominated and classify the continuous distribution to meet their needs.

Climate variables (i.e., average maximum temperature, average minimum temperature) were consistently among the most important predictors for eastern hemlock. This is an important consideration as we are faced with warming temperatures and redistribution of moisture in a changing climate. As data depicting future climate projections generated by climate scenarios become available, forecasts of future range shifts should be possible with a high degree of accuracy.

ACKNOWLEDGEMENTS

This research was supported by National Science Foundation award EPS-0904155 to Maine EPSCoR at the University of Maine. Actual plot locations were obtained through an MOU with the FIA program USDA, Forest Service Northern Research station. The authors would like to thank Richard McCullough of the FIA program for data support; as well as the anonymous reviewers whose comments greatly improved the clarity and focus of the manuscript.

REFERENCES

Albani, M., Moorcroft, P.R., Ellison, A.M., Orwig, D.A., Foster, D.R., 2010. Predicting the impact of hemlock woolly adelgid on carbon dynamics of eastern United States forests. Can. J. For. Res. 40, 119–133.

Boulesteix, A.-L., Janitza, S., Kruppa, J., König, I.R., 2012. Overview of random forest methodolgy and praticial guidance with empahsis on computational biology and bioinformatics. Wiley Interdiscipl. Rev.: Data Min. Knowl. Discov. 2, 493–507.

Boyce, R.L., 2000. Relationship between environmental factors and hemlock distribution in Mt. Ascutney, Vermont. In: McManus, K.A., K.S. Shields, D.R. Souto (Eds.). Proceedings: Symposium on Sustainable Management of Hemlock Ecosystems in Eastern North America. Gen. Tech. Rep. NE-267, United States Department of Agriculture, Forest Service, Northeastern Experiment Station, Newtown Square, Pennsylvania, pp. 113–121.

Briemen, L., 2001. Random forests. Mach. Learn. 45, 5–23.

Campbell, J.L., Rustad, L.E., Boyer, E.W., Christopher, S.F., Driscoll, C.T., Fernandez, I.J., Ollinger, S.V., 2009. Consequences of climate change for biogeochemical cycling in forests of northeastern North America this article is one of a selection of papers from NE Forests 2100: a synthesis of climate change impacts on forests of the northeastern US and eastern Canada. Can. J. For. Res. 39 (2), 264–284.

Clark, J.T., Fei, S., Liang, L., Rieske, L.K., 2012. Mapping eastern hemlock: comparing classification techniques to evaluate susceptibility of a fragmented and valued resource to an exotic invader, the hemlock woolly adelgid. For. Ecol. Manage.

266, 216–222.

Cohen, W.B., Maiersperger, T.K., Spies, T.A., Oetter, D.R., 2001. Modelling forest cover attributes as continuous variables in a regional context with Thematic Mapper data. Int. J. Remote Sens. 22 (12), 2279–2310.

Crist, E., Cicone, R.C., 1984. A physically-based transformation of thematic mapper data–the TM tasseled cap. IEEE Trans. Geosci. Remote Sens. 22 (3), 256–263.

Doucette, J.S., Stiteler, W.M., Quackenbush, L.J., Walton, J.T., 2009. A rule-based approach for predicting the eastern hemlock component of forests in the northeastern United States. Can. J. For. Res. 39, 1453–1464.

Dukes, J.S., Pontius, J., Orwig, D., Garnas, J.R., Rodgers, V.L., Brazee, N., Cooke, B., Theoharides, K.A., Stange, E.E., Harrington, R., Ehrenfeld, J., Gurevitch, J., Lerdau, M., Stinson, K., Wick, R., Ayres, M., 2009. Responses of insect pests, pathogens, and invasive plant species to climate change in the forests of northeastern North America: what can we predict? Can. J. For. Res. 39, 231–248.

Ellison, A.M., Bank, M.S., Clinton, B.D., Colburn, E.A., Elliott, K., Ford, C.R., Foster, D.R., Kloeppel, B.D., Knoepp, J.D., Lovett, G.M., Mohan, J., Orwig, D.A., Rodenhouse, N. L., Sobczak, W.V., Stinson, K.A., Stone, J.K., Swan, C.M., Thompson, J., von Holle, B., Webster, J.R., 2005. Loss of foundation species: consequences for the structure and dynamics of forested ecosystems. Front. Ecol. Environ. 9, 479–

486.

Eschtruth, A.K., Cleavitt, N.l., Battles, J.J., Evans, R.A., Fahey, T.J., 2006. Vegetation dynamics in declining eastern hemlock stands: 9 years of forest response to hemlock woolly adelgid infestation. Can. J. For. Res. 36, 1435–1450.

Faraway, J.J., 2006. Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models. Chapman & Hall/CRC, Boca Raton, FL, 301 pp..

Fassnacht, K.S., Cohen, W.B., Spies, T.A., 2006. Key issues in making and using satellite-based maps in ecology: a primer. For. Ecol. Manage. 222 (2006),167–181.Heywood, V.H., Watson, R.T., 1995. Global Biodiversity Assessment. Cambridge University Press, New York, NY.

Hoppus, M., Lister, A., 2005. The status of accurately locating FIA plots using GPS. In: McRoberts, R., Reams, G., Van Deusen, P., McWillianls, W., (Eds.), Proc. of the 7th Annual Forest Inventory and Analysis Symposium, Portland, Maine, October 3–6, 2005. US For. Serv. Gen. Tech. Rep. WO-77, Washington, DC, pp. 179–184.

Kauth, R.J.,Thomas, G.S., 1976. The Tasseled Capagraphic description of the spectral-temporal development of agricultural crops as seen by Landsat. In: Proceedings of the Symposium on Machine Processing of Remotely Sensed Data. Purdue University, West Lafayette, Indiana, pp. 4B41–4B51.

Koch, F.H., Cheshire, H.M., Devine, H.A., 2005. Mapping hemlocks via tree-based classification of satellite imagery and environmental data. In: Third Symposium on Hemlock Woolly Adelgid in the Eastern United States, 104–114, B. Onken, R. Reardon, comps. FHTET-2005-1. U.S. Department of Agriculture, Forest Service, Forest Health Technology Enterprise Team, Morgantown, WV.

Kong, N., Fei, S., Rieske-Kinney, L., Obrycki, J., 2008. Mapping hemlock forests in Harlan County, Kentucky. In: Bettinger, P., Merry, K., Fei, S., Drake, J., Nibbelink, N., Hepinstall, J. (Eds.), Proceedings of the 6th Southern Forestry and Natural Resources GIS Conference. University of Georgia, Athens, GA.

Liaw, A., Weiner, M., 2007. RandomForest (R software for random forest). Fortran original (L. Breiman and A. Cutler), R port (A. Liaw, M. Wiener) Version 4.5-19 and 4.5-25. <http://cran.r-project.org/web/packages/randomForest/index. html>. Maine Forest Service, 2014. <http://www.maine.gov/doc/mfs/HemlockWoollyAdelgid.htm> (accessed 05.19.15).

Narayanaraj, G., Bolstad, P.V., Elliott, K.J., Vose, J.M., 2010. Terrain and landform influence on Tsuga canadensis (L.) Carriere (eastern hemlock) distribution in the southern Appalachian mountains. Castanea 75 (1), 1–18.

Orwig, D.A., Foster, D.R., 1998. Forest response to the introduced hemlock woolly adelgid in southern New England, USA. J. Torrey Bot. Soc. 125, 59–72.

Orwig, D.A., Foster, D.R., Mausel, D.L., 2002. Landscape patterns of hemlock decline in New England due to the introduced hemlock woolly adelgid. J. Biogeogr. 29, 1475–1487.

Pontius, J., Hallett, R., Martin, M., 2005. Using AVIRIS to assess hemlock abundance and early decline in the Catskills, New York. Remote Sens. Environ. 97, 163–173.

R Core Team, 2015. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0,<http://www.R project.org>.

Randall, A., 1991. The value of biodiversity. Environ. Econ. 20 (2), 64–68. Richardson, D.M., 1998. Forestry trees as invasive aliens. Conserv. Biol. 12, 18–26. Roberts, Dar A., Ustin, S., Ogunjemiyo, S., Greenberg, J., Dobrowski, S., Chen, J.,

Hinckley, T., 2004. Spectral and structural measures of northwest forest vegetation at leaf to landscape scales. Ecosystems 7, 545–562.

Rolston, H., 2000. The land ethic at the turn of the millennium. Biodivers. Conserv. 9,1045–1058.

Schulze, E., Mooney, H. (Eds.), 1994. Biodiversity and ecosystem function. Springer- Verlag.

Simberloff, Relva, D.M.A., Nuñez, M., 2002. Gringos en el bosque: introduced tree invasion in a native Nothofagus/Austrocedrus forest. Biol. Invasions 4, 35–53.

Song, C., Schroeder, T.A., Cohen, W.B., 2007. Predicting temperate conifer forest successional stage distributions with mulitemporal Landsat Thematic Mapper imagery. Remote Sens. Environ. 106, 228–237.

Stadler, B., Muller, T., Orwig, D., Cobb, R., 2005. Hemlock woolly adelgid in New England forests: canopy impacts transforming ecosystems processes and landscapes. Ecosystems 8, 233–247.

Walker, W.S., Kellndorfer, J.M., LaPoint, E., Hoppus, M., Westall, J., 2007. An empirical in SAR-optical fusion approach to mapping vegetation canopy height. Remote Sens. Environ. 109, 482–499.

Weiskittel, A.R., Crookston, N.L., Radtke, P.J., 2011. Linking climate, gross primary productivity, and site index across forests of the western United States. Can. J. For. Res. 41 (8), 1710–1721. http://dx.doi.org/10.1139/x11-086.

Wilson, B.T., Lister, A.J., Riemann, R.I., Griffith, D.M., 2013. Live tree species basal area of the contiguous United States (2000–2009). USDA Forest Service, Northern Research Station, Newtown Square, PA, http://dx.doi.org/10.2737/ RDS-2013-0013.

Wilson, B.T., Lister, A.J., Riemann, R.I., 2012. A nearest-neighbor imputation approach to mapping tree species over large areas using forest inventory plots and moderate resolution raster data. For. Ecol. Manage. 271, 182–198.

Woods Hole Research Center, 2012. <http://www.whrc.org/mapping/nbcd/index. html> (accessed 05.19.15).

Fig. 1. Shows the geographic extent of the study area and all reference plots (N = 3157) used in model calibrations and accuracy assessments.

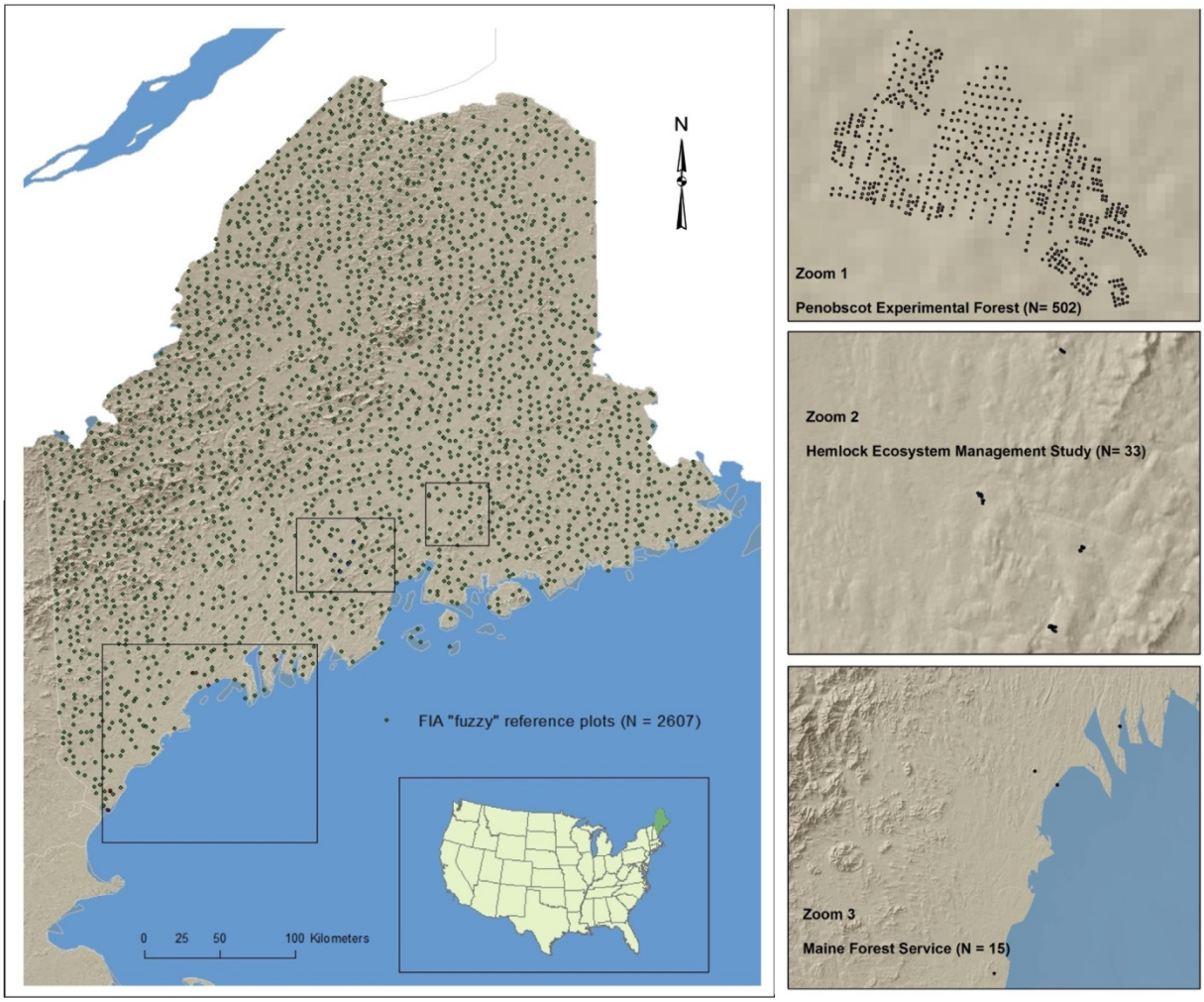

Table 1

Lists the independent variables that proved most important in explaining percent basal area eastern hemlock.

	Predictor	Abbrev.	Resolution (m)	Description	Source
Climate	Maximum temperature	tempmax	800	The 30-yr normal average maximum temperature (1961–1990)	PRISM
	Minimum temperature	tempmin	800	The 30-yr normal average minimum temperature (1961–1990)	PRISM
	Precipitation	precip	800	The 30-yr normal yearly precipitation (1961–1990)	PRISM
Satellite imagery	Visible red	wred	30	Winter (leaf-off) visible red band	Landsat TM
	Greenness	tcgreen	30	Summer (leaf-on) tasseled cap ''greenness"	Landsat TM
	Mid-IR	wmir	30	Winter (leaf-off) mid-IR band	Landsat TM
	Near-IR	snir	30	Summer (leaf-on) near-IR band	Landsat TM
	Thermal-IR	stir	30	Summer (leaf-on) thermal-IR band	Landsat TM
Biomass	Biomass	biomass	30	Above ground biomass	WHRC
Elevation	Digital elevation model	elevation	10	Elevation	USGS
Soils	Available water storage	aws25	10	Available water storage at 25 cm	SSURGO

Table 2

Various field data sources/sites and a summary of the number of plots where eastern hemlock were present, absent, and dominant. We used measurements including tree species and diameter at breast height (dbh) to calculate percent eastern hemlock basal area per plot.

Source	Hemlock-Dominated[2]	Hemlock Present	Hemlock absent	Total
Forest inventory and analysis	57	656	1951	2607
Hemlock ecosystem management study	24	33	0	33
Penobscot experimental forest	120	434	68	502
Maine Forest Service	12	15	0	15
Total	247	1138	2019	3157

[2]P50% eastern hemlock basal area.

Table 3

Source	Mean	Min	Max	SD
Forest inventory and analysis	5.3	0.0	100.0	13.0
Hemlock ecosystem management study	68.1	16.8	100.0	25.1
Penobscot experimental forest	31.0	0.0	94.3	24.9
Maine forest service	63.1	11.8	93.2	20.8
Overall	10.1	0.0	100.0	19.4

Fig. 2. A conceptual model illustrating the general workflow undertaken in the two stage mapping approach.

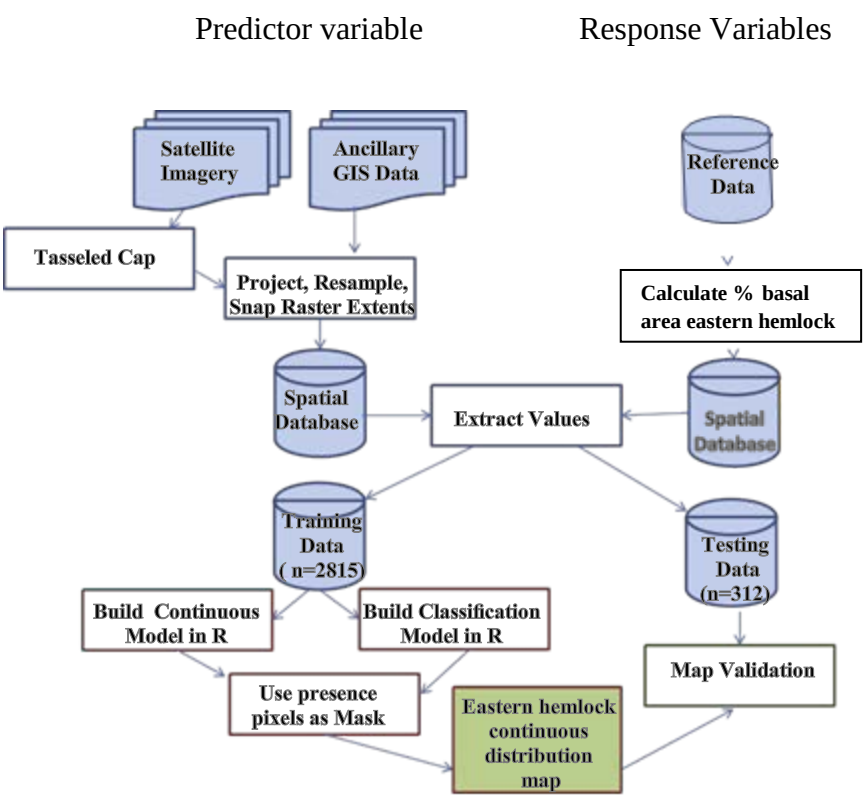

Fig. 3. Graphs of variable importance in the binary model for presence/absence of eastern hemlock in Maine. (a) Mean decrease in the accuracy of the model if that variable were to be removed. (b) Mean decrease in the Gini coefficient. Variables that result in nodes with higher purity have a higher decrease in Gini coefficient. The variables are described in Table 2.

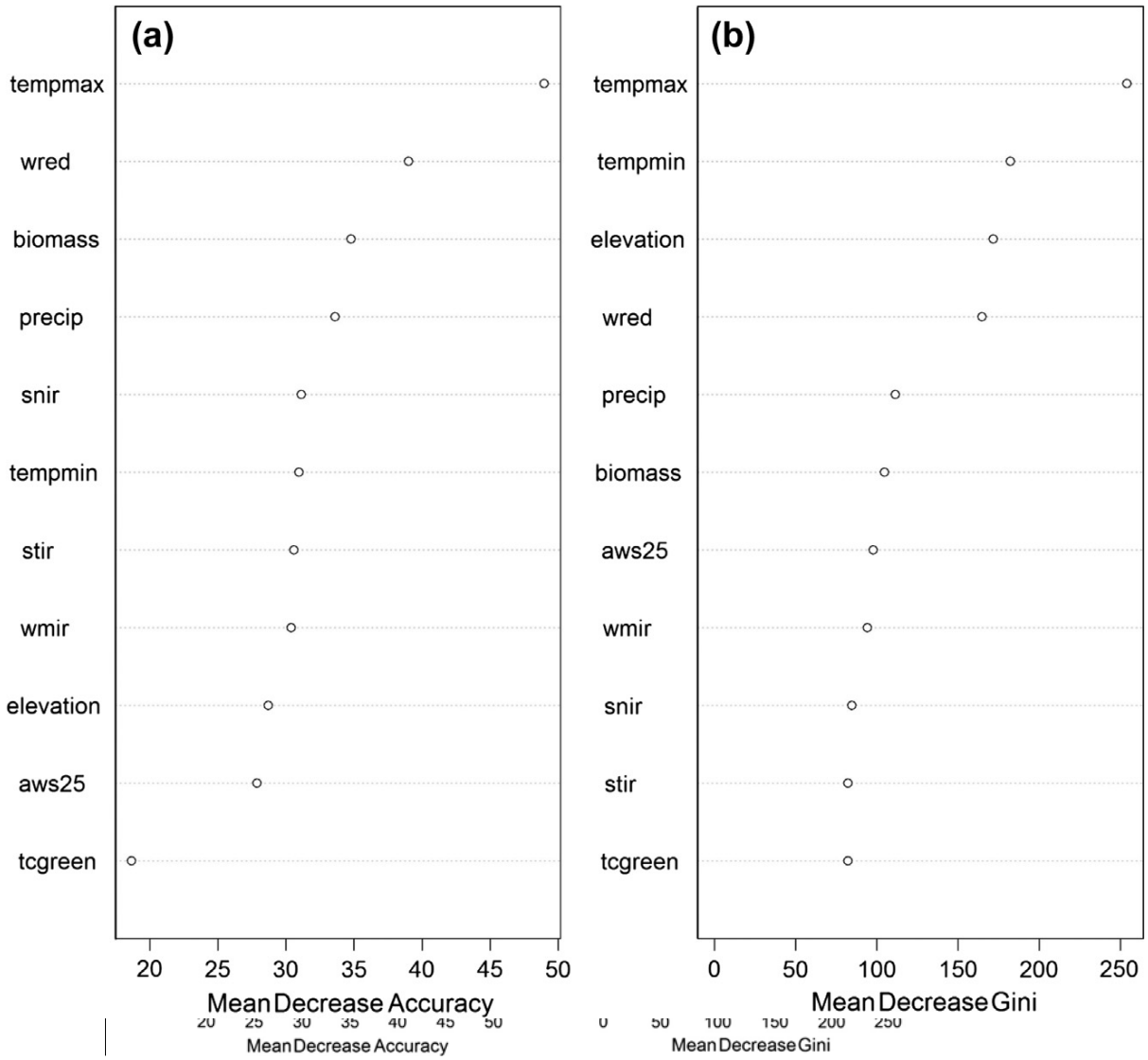

Fig. 4. Accuracy assessment measures for the binary presence/absence model of eastern hemlock in Maine. (a) A presence/absence histogram of observed values as a function of predicted probability. (b) Shows observed vs predicted values in terms of bins, where the y-axis is a ratio of observed # plots that have hemlock present/total # in bin, and the x-axis is the predicted probability of occurrence. The numbers above each point give the total # of plots in that bin. (c) Receiver Operator Characteristic (ROC) curve and the associated area under the curve (AUC), a threshold independent measure of model quality. (d) Shows error (sensitivity, specificity, Kappa) as a function of threshold, where sensitivity and specificity cross at a high value indicating quality model. The Kappa statistic stays relatively high over a range of threshold values.

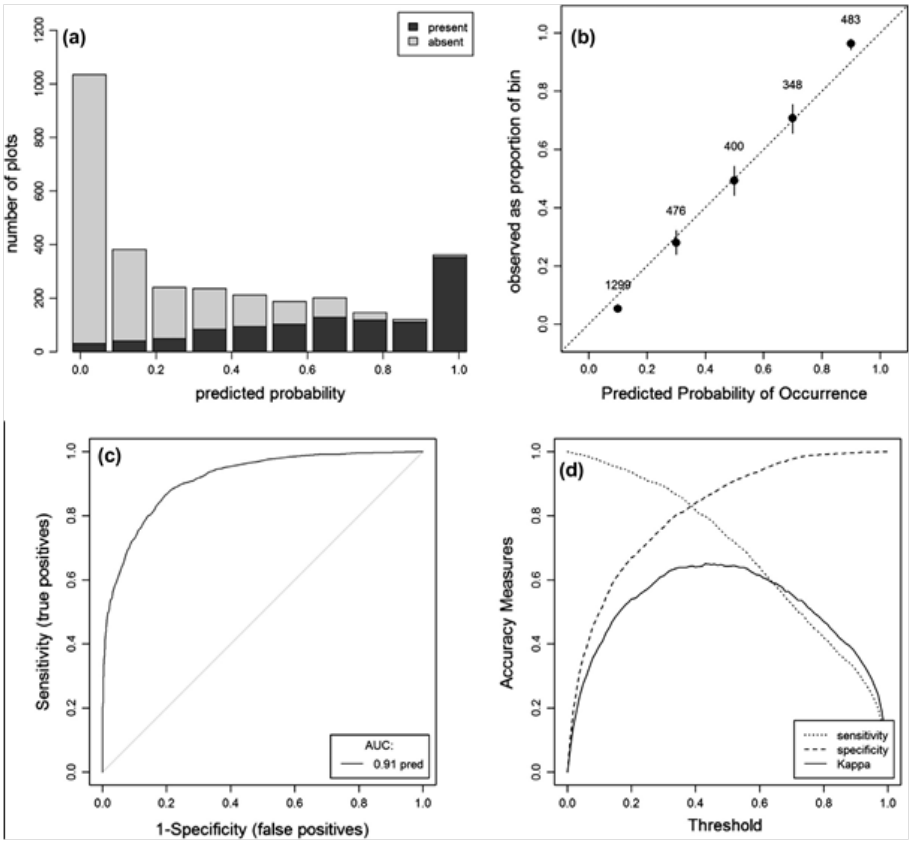

Fig. 5. Variable importance plots for predicted hemlock abundance (% basal area). (a) Percent increase in mean squared error in the model in absence of a given variable. (b) Shows the importance of the first four variables over the remaining eight. The variables are described in Table 1

]

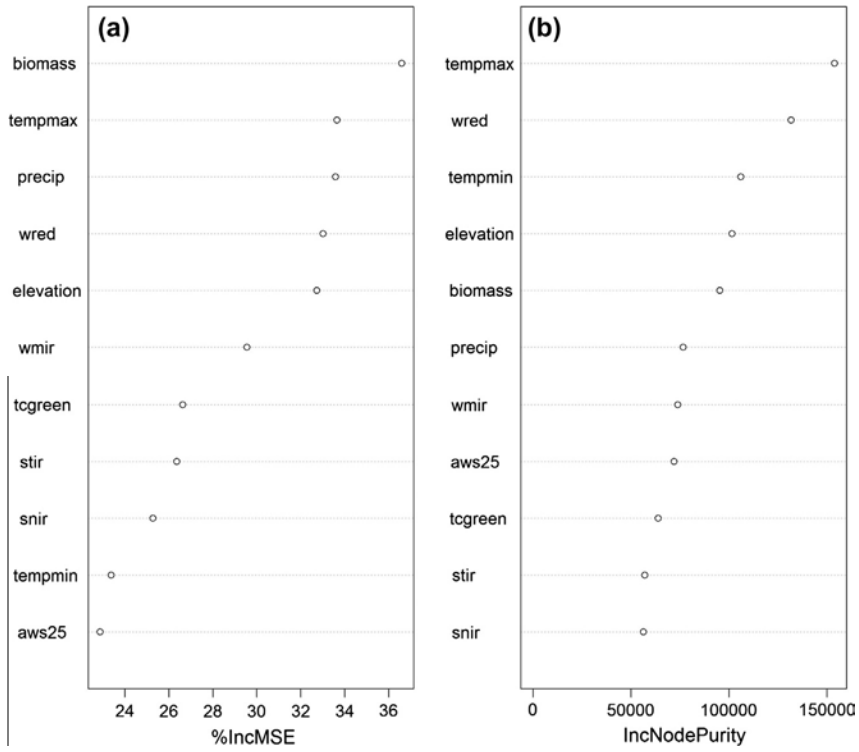

Fig. 6. Observed versus predicted values of hemlock abundance (% basal area) for training reference plots (n = 2815).

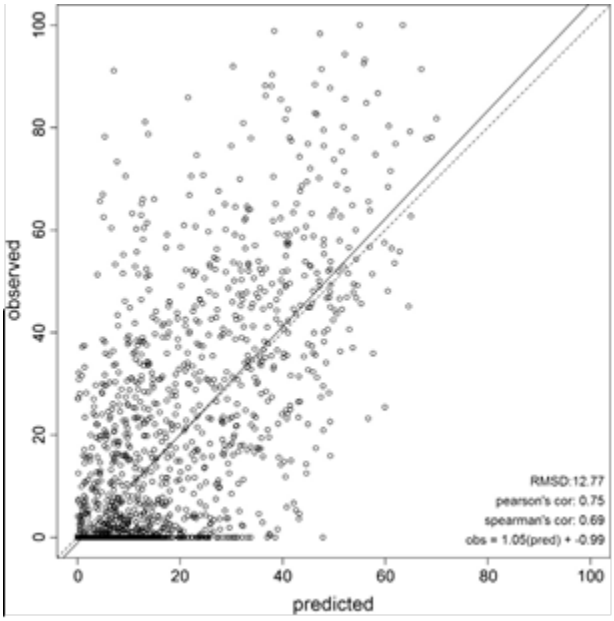

Fig. 7. Shows the spatial relationships between the random Forest model predictions of eastern hemlock(% basal area) and the four most important predictor variables

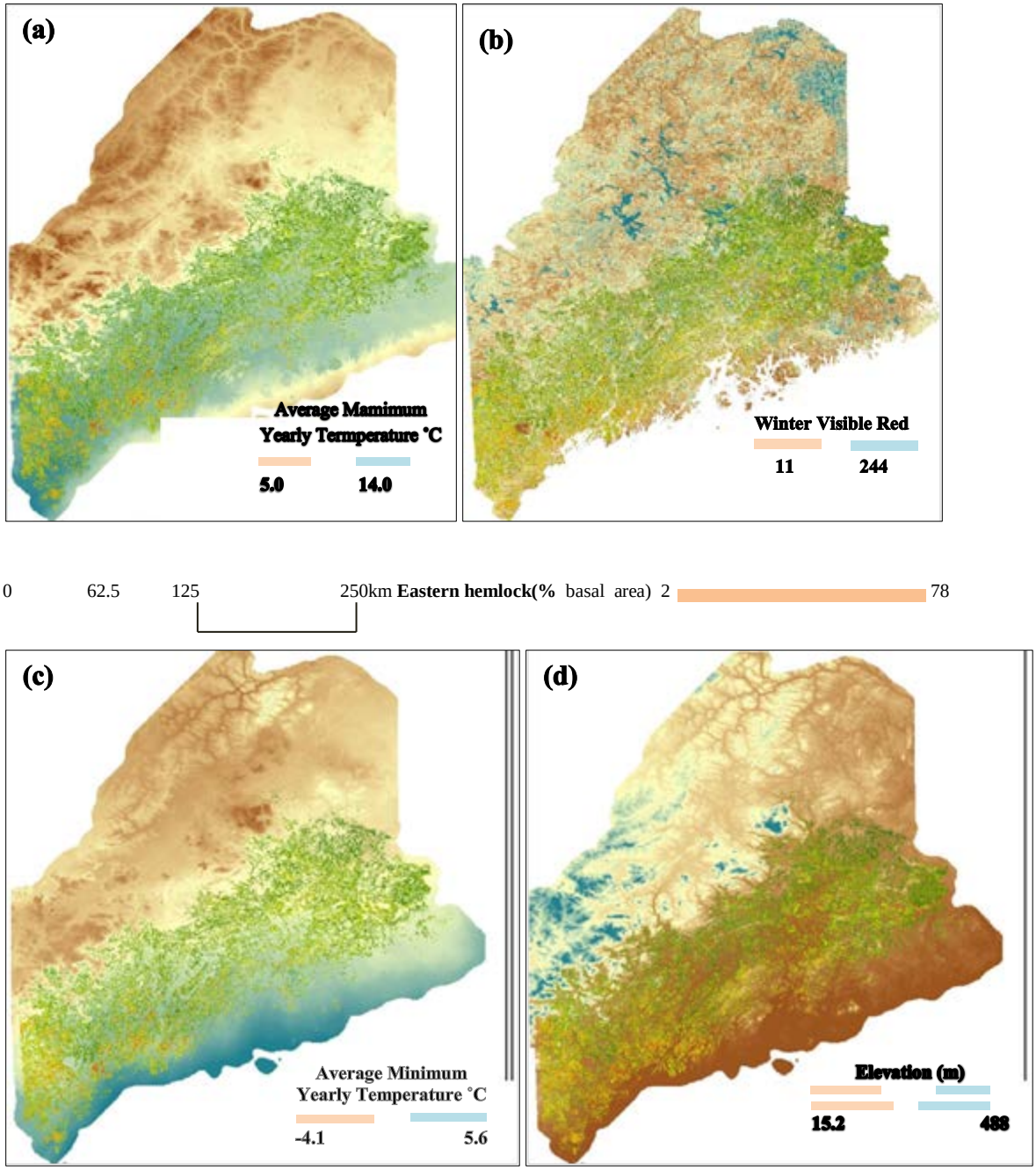

Fig. 8. The estimated distribution of eastern hemlock (% basal area) at 30 m resolution throughout the state of Maine, located in the northeastern corner of the U.S. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

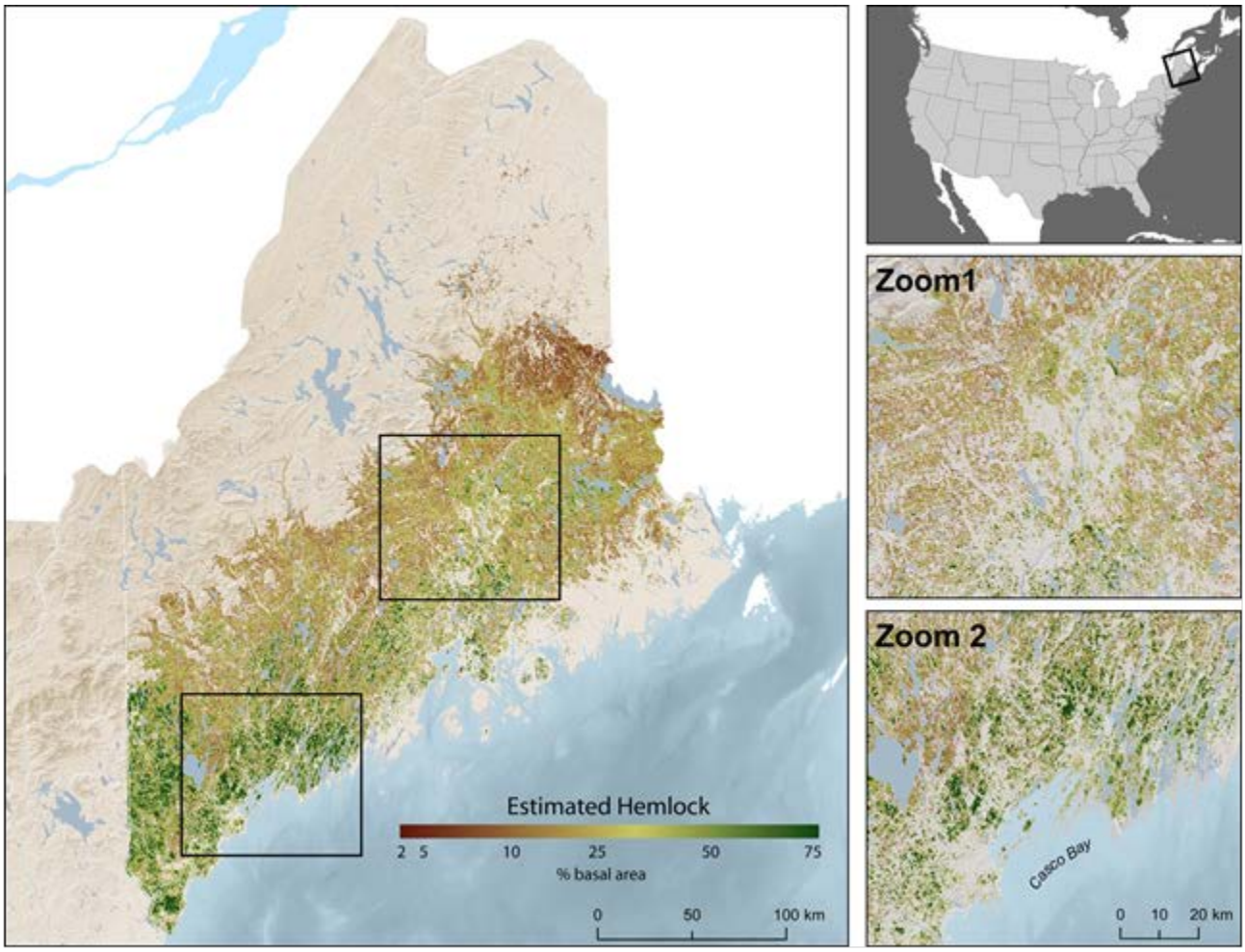

Fig. 9. Final map product accuracy assessment showing observed vs predicted values of hemlock abundance (% basal area) for test plots (n = 312).