Incorporating Climate Sensitivity for Eastern United States' Tree Species into the Forest Vegetation Simulator

Huiquan Jiang

Dissertation submitted to the faculty of the Virginia Polytechnic Institute and State University in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in

Forestry

Philip J. Radtke, Chair

Harold E. Burkhart

Stephen P. Prisley

Marion R. Reynolds, Jr.

July 14, 2015

Blacksburg, Virginia

Keywords: Climate change; Random Forests; Site index; Mortality; Diameter increment; Climate and soils; Bootstrap

Incorporating Climate Sensitivity for Eastern United States' Tree Species into the Forest Vegetation Simulator

Huiquan Jiang

Abstract

Detecting climate-induced effects in forest ecosystems become increasingly important as more evidence of greenhouse-gas-related climate change were founded. The Forest Vegetation Simulator (FVS) is an important growth and yield model used to support management and planning on public forest lands over the southern United States, however its prediction accuracy was challenged due to its climate- insensitive nature. The goal of this study was to develop species-specific prediction models for eastern U.S. forest tree species with climate and soil properties as predictors in order to incorporate the effects of climate and soils-based variables on forest growth and yield into FVS-Sn. Development of climate- sensitive models for site index, individual-tree mortality and diameter increment were addressed separately, which were all developed using Random Forests on the basis of USDA Forest Service Forest Inventory and Analysis program linked to contemporary climate data and soil properties mapped in the USDA Soil Survey Geographic SSURGO database. Results showed climate was a stronger driver of site index than soils. When soils and climate were used together, site index predictions for species grouped as conifers or hardwoods were almost as precise as species-specific models for many of the most common eastern forest tree species. Model comparison was conducted to pursue the most suitable individual-tree mortality prediction model for 20 most important species among Logistic Regression, Random Forests, and Artificial Neural Networks. Results showed that

Random Forests with all indicators involved generally performed well, especially sound for species with medium and high mortality. At a chosen threshold, it frequently achieved the equally highest value of sensitivity and specificity among chosen candidates. To evaluate the prediction ability of Random Forests model on individual-tree diameter increment, Multiple Linear Regression model was built as baseline on each of most common 20 species eastern U.S. area. Comparison results showed that Random Forests gained advantages in model validation and future projection under climate change. Using the developed climate-sensitive models, multiple maps were produced to illustrate how forest tree growth, yield, and mortality of individual tree may change in the eastern U.S. over the 21st century under several climate change scenarios.

Acknowledgments

First and foremost, I would like to express my sincere gratitude to my advisor Professor Philip J. Radtke for his patient guidance and mentorship during my PhD period at Virginia Tech. His extensive knowledge, precise and logical research attitude became a beacon, which is always lighting the way in my research work. His continuous support and encouragement are the essential driving force for me to overcome all the difficulties before arriving at the destination.

I would like to thank my committee members: Dr. Harold E. Burkhart, Dr. Stephen Prisley, and Dr. Marion R. Reynolds for useful guidance and suggestions on my research work, which greatly helped me move faster in the research. Great appreciation is given to Dr. Aaron R. Weiskittel, Dr. John W. Coulston and Dr. Patrick J. Guertin for their insightful advice on Chapter 2, which significantly improved my work for peer-review publication.

I would like to thank the US Army Corps Engineering Research Laboratory for financial support towards my graduate program. The funding sources are the important factors in my PhD research work, which is greatly appreciated. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the funding agencies. I also deeply thank the technical support start from the branch department of USDA Forest Service including Forest Service Management Center, Moscow Forestry Sciences Lab, Northern Research Station & FIA and Southern Research Station & FIA. I am thankful to the Department of Forest Resources and Environmental Conservation (FREC) at Virginia Tech for the administrative support and cooperation.

I greatly appreciate the support my dear friends gave me during my time in Blacksburg. Thanks to my past and present fellow graduate students Melissa Shockey, Clara DeYoung, Trevor Savile, Charles Obuya Sabatia, Mickey Allen, Nabin Gyawali and Ram Thapa for their encouragement and friendship. I would express my sincere appreciation to my parents, parents in law and other family members. Their unconditional support and endless pure love is my essential spiritual support, providing the courage and faith to overcome all the difficulties. In the meantime, I am especially indebted to my husband, Wei Jing, who always stands by me and shares his most

valuable time with me over the past ten years. Our six- month old daughter, Adela Jing, is the fountain of inspiration and hope for our life in the past, now and the future.

Table of Contents

List of Tables

List of Figures

Chapter 1 Introduction

Mounting evidence over greenhouse-gas related global warming has raised concerns about the effects that regional or continental-scale climate change will have on North American forest ecosystems (Iverson et al. 2004). Following the observed mean global temperature increase of about 0.8 - 1.0° C over the past half-century, climate models predict possible additional mean temperature increases between 2.5° - 6.6° C over the next century. Changes of this magnitude are likely to have noticeable effects on the forest ecosystems of North America (IPCC 2007; Iverson et al. 2004). Likely impacts of climate change on forests include potential shifts in species geographic distributions (Hamann and Wang 2006; Iverson and Prasad 2002), species-specific changes in growth or mortality rates (Andalo et al. 2005; Andreu et al. 2007; Battles et al. 2008; Rehfeldt et al. 1999), changes in site productivity (Nigh 2006; Weiskittel et al. 2011), and other outcomes, such as increased frequency or intensities of fire, insect or disease outbreaks, and other disturbance events (Lenihan et al. 2003; Woods et al. 2010).

Forest ecosystem predictive models in the past were mainly grouped in to three categories: statistical or traditional-based empirical models; process or mechanistic models; and hybrid models, which combined elements of the other two.

1.1 Traditional-based empirical models

Traditional-based forest empirical models include yield tables, stand models and individual tree models. In most traditional-based empirical models, the response variable could be basal area ha^{-1} or mean-top-height, which is expressed as a regression function of variables such as tree size, tree density, age, and site quality, then might be modified by the effect of competition (Dale et al. 1985). Model building typically requires large collections of inventory data or field observations for testing and quantifying the nature of growth and yield relationships. Mathematical functional relationships between predictors and response are required when formulating empirical models, and most work in this area assumes the functional forms chosen are correct (Rykiel 1996). An advantage of the functional approach is that interpretation of model relationships and relative rates of change are readily derived (Taylor et al. 2009). Empirical models are easily

implemented for use in forest management or ecological analyses, and can achieve good efficiency and accuracy in short-term forecasting (Peng and Wen 2006).Nevertheless, a fundamental disadvantage of traditional empirical forest growth and yield models is that they are not typically able to directly account for variation in environmental factors (Kimmins 1990). In this sense they are insensitive to climate change since they ignore underlying physiological processes such as photosynthesis and respiration, or feedbacks in carbon, nutrient, and hydrological cycles that can be strongly influenced by climate change.

Several computer-based forest dynamics simulation models have been developed that successfully rely on traditional empirical frameworks, such as PROGNOSIS (Stage 1973), STEMS (Belcher et al. 1983), TWIGS (Miner et al. 1988), PTAEDA (Burkhart et al. 1987) and FVS (Teck et al. 1996).

1.2 Process models

Mechanistic or process-based forest ecosystem models take a somewhat different approach to modeling relationships between observed model inputs and predicted outputs. Their framework relies on mathematical functions that characterize underlying ecological and physiological mechanisms of plant and ecosystem functioning, rather than describing empirical relationships among forest inventory variables (Woodward 1987). Some of the processes commonly accounted for in these models include photosynthesis and respiration (carbon allocation), evaporation, transpiration and water uptake, nutrient use and cycling, and many others. Typically, the underlying processes are described using one or more mathematical functions based on knowledge of the mechanisms of plant physiology (Landsberg 1986). Model parameters are often estimated from statistical analyses of research data that closely examine the underlying relationships. Examples include the plant respiration response to elevated temperature or the photosynthetic responses to changing light intensity or moisture deficit (Aber et al. 1996; Sullivan et al. 1996).

Mechanistic models are often designed to account for climate and edaphic factors using functional forms that explain causal effects of environment on vegetation (Constable and Friend 2000). Their use by forest managers and practitioners has been limited, mainly by the models'

data-intensive requirements and challenges in parameterizing such models for use over a range of spatio-temporal scales and for different species. Typical forest inventory databases contain measurements of tree, stand, and site attributes, which may be useful for developing empirical growth and yield models; however, mechanistic models generally require detailed physiological and environmental measurements that are expensive and relatively uncommon. Mechanistic models often operate at finer levels of resolutions in time and space than most empirical growth and yield models currently in use (Korzukhin et al. 1996). Some forest modelers have avoided working with such paradigms because they were not seen as pratical for forest mangement; moreover, they do not necessarily lead to improved accuracy or applicability when compared to empirical forest-management models that operate at coarser resolutions (Mäkelä et al. 2000; Mohren and Burkhart 1994). As the complexity – number of variables, functions, and functional interactions – of physiological models increases, the more likely they are to produce biased results due to the propagation of errors between sub-models and the iteration of predicted intermediate variables subject to errors (Mason et al. 2011). Some studies have concluded that process-based models are not well suited for simulating the variety of silvicultural prescriptions needed for forest management and planning (Sands et al. 2000). Process models will serve primarily only in forest research and policy evaluation, and that empirical models will continue to serve as primary tool for forest management (Korzukhin et al. 1996).

A number of process models have been designed to address climate change or related effects in forest ecosystems, including Biomass (McMurtrie and Landsberg 1992), ACIDIC (Kareinen et al. 1998), BGC (Running and Coughlan 1988), SECRETS (Deckmyn et al. 2004),TREEDYN3 (Bossel 1996), TEMFES (Nikolov and Fox 1994) and TRIPLEX1.0 (Xiaolu et al. 2005).

1.3 Hybrid models

Hybrid models aim to adopt the usefulness of both empirical and mechanistic modeling approaches by combining elements of both. The usual rationale for this approach is to achieve the accuracy and linkages to inventory-based measurements of empirical models along with the ability to account for ecophysiological measurements and parameters as in process-based models (Johnsen et al. 2001; Mäkelä et al. 2000). A primary motivating factor for the rising interest in

hybrid models in past decades is the need to account for climate-change effects on forest vegetation in management-based growth and yield models (Gustafson et al. 2000; Landsberg et al. 2001).While no uniform definition has been adopted for what comprises a hybrid model, and no standard method exists for how to formulate hybrid models, several approaches have been somewhat widely used.

One approach to hybrid modeling involves the use of inventory-based information as a means to constrain or condition the outputs of process-based models (Waring and McDowell 2002). The constraining approach can be accomplished using observations from field studies, or using predicitons from well-validated empirical models (Radtke et al. 2002; Radtke and Robinson 2006). A related approach involves joining predictions from two models – one mechanistic and one empirical – in a composite-type hybrid model (Kelsey et al. 2003; Waterworth et al. 2007). To accomplish this result, researchers have formulated growth indices from mechanistic model outputs that were subsequently used to modify empirical model predictions (Henning and Burk 2004; Snowdon et al. 1998). In other work to join both types of models, the outputs of a process-based model have been used as inputs for an empirical model, or *vice versa* (Baldwin et al. 2001). Similarly, variables typically associated with mechanistic models can be used in place of conventional inputs in empirical models, as was done by Mason et al. (2011) who substituted a cumulative solar radiation index in place of time in a growth and yield modeling system.

Other hybrid models combine functional forms and equations characteristic of both mechanistic processes and empirical models (Sievänen et al. 1988). One example, PipeQual, is a model that utilizes a process-based approach to the development of stem form and crown structure, in connection with empirical submodels of branch numbers, locations, and inclinations (Makela 1997). The 3-PG model of Landsberg and Waring (1997) calculates total carbon fixed from photosynthetically active radiation based on both physiological processes and empirical constants and relationships. For example, gross primary production was estimated from an empirical constant they referred to as canopy quantum efficiency. The Forest-BGC model was calibrated to a detailed set of physiological measurements from a small number of trees and was then shown to predict tree growth well in a larger, independent validation study. A subsequent adaptation known as TREE-BGC employed a competition algorithm and allometric relationships

to the internal model structure to enable the generation of tree-level outputs such as dbh and total height (Korol et al. 1991).

Statistical modeling approaches such as Multiple Linear Regression (MLR), Generalized Additive Models (GAM), and others have been used to formulate hybrid models, as was done by Robards (2009), who developed MLR models for six commercially important conifer species in California that included climatic, topographic and location variables in addition to conventional mensurational predictors. In a more sophisticated approach Huang et al. (2011) combined a partially linear model with smoothing splines in order to predict growing-stock volume based on several forest inventory and climate-derived variables. Bravo-Oviedo et al. (2010) developed a climate-based dominant height projection model for maritime pine (*P. pinaster* Aiton) augmenting initial height and age predictors with temperature, seasonal precipitation, drought length, and lithology as drivers.

Hybrid models that account for both growth and yield modeling concepts and the effects of environmental variables have become widespread in the literaure. Waterworth et al. (2007) enhanced the FullCAM system to allow its application in plantations by calibrating a generalized growth model. Girardin et al. (2008) employed different methods – empirical, process-based, and hybrid modeling – to estimate tree growth responses to changing climate in boreal central Canada. Battaglia et al. (1999) used the process-based model PROMOD to predict site index, which was then used as an input to an empirical yield model to predict the time course of volume and height growth in short-rotation plantations. From their common use in research and increasing use in practical applications, hybrid models represent an important step in the development of forest ecosystem modeling.

To account for the climate changes on forest ecosystem, forest managers and policymakers require modeling tools capable of informing plans or practices designed to promote long-term forest sustainability. Some existing models may account for climate variables in their predictions by applying ecophysiological principles, such as mass and energy balances related to plant photosynthesis, assimilation, respiration, and transpiration such as LANDIS (He et al. 1999), FVS-BGC (Kelsey et al. 2003). Although important for their ability to advance the science of plant ecophysiology, such process or mechanistic models may not operate at functional or

temporal scales suitable for forest management (Korzukhin et al. 1996). To be useful, a management-oriented model must be able to accurately predict current conditions, and to make future predictions for inventory-based attributes such as tree size or mortality rates by species. These attributes are known to be affected by forest management activities ranging from planting to thinning or harvesting trees. They are also affected by growing conditions, including the soil media in which forests grow, as well as any long-term changes in the climate at a particular growing location.

The Forest Vegetation Simulator (FVS) is a growth and yield prediction system capable of predicting current and future forest inventory attributes while accounting for a wide range of management activities' impacts on growth and yield. Until recently FVS had no mechanism for accounting for any effects of climate or soils-related variables on its predictions. Recent developments have led to the augmentation known as Climate-FVS, which accounts for temperature and precipitation-related variables in making predictions of forest tree growth, yield, and mortality (Crookston et al. 2010). Climate-FVS was developed for Western U.S. species and climate regimes, with model logic suited to western conditions. As such Climate-FVS has until now been implemented only for western FVS variants (US Forest Service 2013). To date no such augmentation has been developed for eastern variants.

The Southern Variant (FVS-Sn) is an important growth and yield model used to support management and planning on public forest lands in the U.S. Forest Service's National Forest System over the southern United States. With National Forests in thirteen southern states covering over 10 million hectares and subject to strict federal requirements for management and planning, FVS-Sn is an important modeling tool. To date FVS-Sn has no capability to account for the impacts of changing climate on southern forests. While the approach used to incorporate climate impacts into Climate-FVS for western variants may be generally relevant to forests growing in the U.S. South, its western focus creates some inconsistencies for use in eastern settings, including differences in species, spatial and temporal modeling scales, management activities and their intensity, availability of soils and forest inventory data, and the ranges of climate and terrain that affect western versus eastern forests. Rather than duplicating the Climate-FVS structure and logic for application to eastern forests, a systematic study is needed

to evaluate, adopt, modify, and develop new approaches for incorporating climate and soils effects on the growth and yield of eastern forests.

The purpose of this work was to incorporate the effects of climate and soils-based variables on forest growth and yield into FVS-Sn. In order to accomplish this any modifications or augmentations must match the variables, scale, and structure of FVS-Sn in its current implementation. The model must be driven primarily by inventory-based attributes including tree species, numbers, and sizes. Added inputs should incur little additional cost to model users. In addition, predictions from a climate-sensitive FVS-Sn should be consistent with the model in its present form, as well as being compatible with the FVS modeling system in general. To achieve these goals the following objectives were pursued in detail:

- Development of a set of models to predict site index for important eastern species. The main predictors for site index models included a range of temperature and precipitation variables, along with variables related to soil properties. The climate and soils predictors were based on existing public data sets available for the eastern U.S.

- Development of species-specific individual-tree mortality models that can be applied over projection intervals appropriate for eastern species. In addition to tree and stand-level predictors, climate and soils variables were incorporated in the mortality models.

- Development of species-specific individual-tree diameter growth models that included traditional tree and stand variables as predictors, along with climate and soils.

The goals pursued here made use of regression tools well-suited for the task of modeling complex response-surfaces that may include numerous interactions or nonlinear relationships. Following the lead of the developers of Climate-FVS, classification and regression-tree-based ensemble prediction methods were examined thoroughly for their utility here. The resulting models were designed for incorporation with the FVS system using similar technology as that used by Climate-FVS in its western implementation. Following model development, an exploration were made into the spatio-temporal relationships between climate and individual tree species under future projected climate scenarios using the Southern Variant Climate-FVS.

Results were intended to demonstrate the use of Climate-FVS in applications for forest management and planning under future climate change scenarios.

1.4 References

Aber JD, Reich PB, Goulden ML, 1996. Extrapolating Leaf CO_2 Exchange to the Canopy: A Generalized Model of Forest Photosynthesis Compared with Measurements by Eddy Correlation. Oecologia 106(2):257-265.

Andalo C, Beaulieu J, Bousquet J, 2005. The impact of climate change on growth of local white spruce populations in Québec, Canada. Forest Ecology and Management 205(1–3):169-182.

Andreu L, GutiÉRrez E, Macias M, Ribas M, Bosch O, Camarero JJ, 2007. Climate increases regional tree-growth variability in Iberian pine forests. Global Change Biology 13(4):804-815.

Baldwin VC, Jr., Burkhart HE, Westfall JA, Peterson KD, 2001. Linking growth and yield and process models to estimate impact of environmental changes on growth of loblolly pine. 47(1):77-82.

Battaglia M, Sands PJ, Candy SG, 1999. Hybrid growth model to predict height and volume growth in young Eucalyptus globulus plantations. Forest Ecology and Management 120(1–3):193-201.

Battles J, Robards T, Das A, Waring K, Gilless JK, Biging G, Schurr F, 2008. Climate change impacts on forest growth and tree mortality: a data-driven modeling study in the mixed-conifer forest of the Sierra Nevada, California. Climatic Change 87(1):193-213.

Belcher DW, Holdaway MR, Brand GJ, North Central Forest Experiment S. 1983. A description of STEMS : the stand and tree evaluation and modeling system U.S. Dept. of Agriculture, Forest Service, North Central Forest Experiment Station, [Saint Paul, Minn.].

Bossel H, 1996. treedyn3 forest simulation model. Ecological modelling 90(3):187-227.

Bravo-Oviedo A, Gallardo-Andrés C, del Río M, Montero G, 2010. Regional changes of Pinus pinaster site index in Spain using a climate-based dominant height model. Canadian Journal of Forest Research 40(10):2036-2048.

Burkhart HE, Farrar RL, Amateis RL, Daniels RF. 1987. Simulation of individual tree growth and stand development in Loblolly Pine plantations on cutover, site-prepared areas Publication Number FWS-1-87. School of Forestry and Wildlife Resources, Virginia Polytechnic Institute and State University.

Constable JVH, Friend AL, 2000. Suitability of process-based tree growth models for addressing tree response to climate change. Environmental Pollution 110(1):47-59.

Crookston NL, Rehfeldt GE, Dixon GE, Weiskittel AR, 2010. Addressing climate change in the forest vegetation simulator to assess impacts on landscape forest dynamics. Forest Ecology and Management 260(7):1198-1211.

Dale VH, Doyle TW, Shugart HH, 1985. A comparison of tree growth models. Ecological modelling 29(1–4):145-169.

Deckmyn G, Laureysens I, Garcia J, Muys B, Ceulemans R, 2004. Poplar growth and yield in short rotation coppice: model simulations using the process model SECRETS. Biomass & Bioenergy 26(3):221-227.

Girardin MP, Raulier F, Bernier PY, Tardif JC, 2008. Response of tree growth to a changing climate in boreal central Canada: A comparison of empirical, process-based, and hybrid modelling approaches. Ecological modelling 213(2):209-228.

Gustafson EJ, Shifley SR, Mladenoff DJ, Nimerfro KK, He HS, 2000. Spatial simulation of forest succession and timber harvesting using LANDIS. Canadian Journal of Forest Research 30(1):32-43.

Hamann A, Wang T, 2006. Potential effects of climate change on ecosystem and tree species distribution in British Columbia. Ecology 87(11):2773-2786.

He HS, Mladenoff DJ, Boeder J, 1999. An object-oriented forest landscape model and its representation of tree species. Ecological Modelling 119(1):1-19.

Henning JG, Burk TE, 2004. Improving growth and yield estimates with a process model derived growth index. Can. J. For. Res.-Rev. Can. Rech. For. 34(6):1274-1282.

Huang JIN, Abt BOB, Kindermann G, Ghosh S, 2011. Empirical analysis of climate change impact on loblolly pine plantations in the southern United States. Natural Resource Modeling 24(4):445-476.

IPCC. 2007. Climate change 2007: mitigation. contribution of working group III to the fourth assessment report of the Intergovernmental Panel on Climate Change Cambridge University Press, Cambridge.

Iverson LR, Prasad AM, 2002. Potential redistribution of tree species habitat under five climate change scenarios in the eastern US. Forest Ecology and Management 155(1–3):205-222.

Iverson LR, Schwartz MW, Prasad AM, 2004. How fast and far might tree species migrate in the eastern United States due to climate change? Global ecology and biogeography letters 13(3):209-219.

Johnsen K, Samuelson L, Teskey R, McNulty S, Fox T, 2001. Process models as tools in forestry research and management. Forest Science 47(1):2-8.

Kareinen T, Ilvesniemi H, Nissinen A. 1998. Analysis of forest soil chemistry and hydrology with a dynamic model ACIDIC Finnish Society of Forest Science [u.a.], Helsinki.

Kelsey SM, Dean WC, Andrew JM, Eric LS, 2003. FVSBGC: A hybrid of the physiological model STAND-BGC and the forest vegetation simulator. Canadian Journal of Forest Research 33(3):466.

Kimmins JP, 1990. Modelling the Sustainability of Forest Production and Yield for a Changing and Uncertain Future. The Forestry Chronicle 66(3):271-280.

Korol RL, Running SW, Milner KS, Hunt Jr ER, 1991. Testing a mechanistic carbon balance model against observed tree growth. Canadian Journal of Forest Research 21(7):1098-1105.

Korzukhin MD, Ter-Mikaelian MT, Wagner RG, 1996. Process versus empirical models: which approach for forest ecosystem management? Canadian Journal of Forest Research 26(5):879-887.

Landsberg JJ. 1986. Physiological ecology of forest production Academic Press, London; Orlando.

Landsberg JJ, Johnsen KH, Albaugh TJ, Allen HL, McKeand SE, 2001. Applying 3-PG, a simple process-based model designed to produce practical results, to data from loblolly pine experiments. Forest Science 47(1):43.

Landsberg JJ, Waring RH, 1997. A generalised model of forest productivity using simplified concepts of radiation-use efficiency, carbon balance and partitioning. Forest Ecology and Management 95(3):209-228.

Lenihan JM, Drapek R, Bachelet D, Neilson RP, 2003. Climate change effects on vegetation distribution, carbon, and fire in California. Ecological Applications 13(6):1667-1681.

Makela A, 1997. A carbon balance model of growth and self-pruning in trees based on structural relationships. Forest Science 43(1):7-24.

Mäkelä A, Landsberg J, Ek AR, Burk TE, Ter-Mikaelian M, Ågren GI, Oliver CD, Puttonen P, 2000. Process-based models for forest ecosystem management: current state of the art and challenges for practical implementation. Tree Physiology 20(5-6):289-298.

Mason EG, Methol R, Cochrane H, 2011. Hybrid mensurational and physiological modelling of growth and yield of Pinus radiata D.Don. using potentially useable radiation sums. Forestry 84(2):99-108.

McMurtrie RE, Landsberg JJ, 1992. Using a simulation model to evaluate the effects of water and nutrients onthe growth and carbon partitioning of Pinus radiata. Forest Ecology and Management 52(1–4):243-260.

Miner CL, Walters NR, Belli ML, North Central Forest Experiment S. 1988. A guide to the TWIGS program for the North Central United States U.S. Dept. of Agriculture, Forest Service, North Central Forest Experiment Station, [Saint Paul, Minn.].

Mohren GMJ, Burkhart HE, 1994. Contrasts between biologically-based process models and management-oriented growth and yield models-preface. Forest Ecology and Management 69(1-3):1-5.

Nigh G, 2006. Impact of climate, moisture regime, and nutrient regime on the productivity of Douglas-fir in coastal British Columbia, Canada. Climatic Change 76(3-4):321-337.

Nikolov NT, Fox DG, 1994. A coupled carbon-water-energy-vegetation model to assess responses of temperate forest ecosystems to changes in climate and atmospheric CO2. Part I. Model concept. Environmental Pollution 83(1–2):251-262.

Peng C, Wen X, 2006. Forest simulation models. In: Computer Applications in Sustainable Forest Management--Shao G, Reynolds K, eds.: Springer Netherlands. 101-125.

Radtke PJ, Burk TE, Bolstad PV, 2002. Bayesian melding of a forest ecosystem model with correlated inputs. Forest Science 48:701-711.

Radtke PJ, Robinson AP, 2006. A Bayesian strategy for combining predictions from empirical and process-based models. Ecological Modelling 190(3-4):287-298.

Rehfeldt GE, Ying CC, Spittlehouse DL, Hamilton DA, Jr., 1999. Genetic responses to climate in Pinus contorta: Niche Breadth, Climate Change, and Reforestation. Ecological Monographs 69(3):375-407.

Robards TA, 2009. Empirical forest growth model evaluations and development of climate-sensitive hybrid models. Ph.D. thesis, United States - California: University of California, Berkeley.

Running SW, Coughlan JC, 1988. A general model of forest ecosystem processes for regional applications I. Hydrologic balance, canopy gas exchange and primary production processes. Ecological Modelling 42(2):125-154.

Rykiel EJ, 1996. Testing ecological models: the meaning of validation. Ecological Modelling 90:229-244.

Sands PJ, Battaglia M, Mummery D, 2000. Application of process-based models to forest management: experience with PROMOD, a simple plantation productivity model. Tree Physiology 20(5-6):383-392.

Sievänen R, Burk TE, Ek AR, 1988. Construction of a stand growth model utilizing photosynthesis and respiration relationships in individual trees. Canadian Journal of Forest Research 18(8):1027-1035.

Snowdon P, Benson M, Woollons R, 1998. Incorporation of climatic indices into models of growth of Pinus radiata in a spacing experiment. New Forests 16(2):101-123.

Stage AR. 1973. Prognosis Model for Stand Development Res. Pa INT-137. U.S. Department of Agriculture,Forest Service, Intermountain Forest and Range Experiment Station, Ogden, UT. 32.

Sullivan NH, Bolstad PV, Vose JM, 1996. Estimates of net photosynthetic parameters for twelve tree species in mature forests of the southern Appalachians. Tree Physiology 16(4):397-406.

Taylor AR, Chen HYH, VanDamme L, 2009. A Review of Forest Succession Models and Their Suitability for Forest Management Planning. Forest Science 55(1):23-36.

Teck R, Moeur M, Eav B, 1996. Forecasting Ecosystems with the Forest Vegetation Simulator. Journal of Forestry 94(12):7-10.

US Forest Service, 2013. Climate-FVS http://www.fs.fed.us/fmsc/fvs/whatis/climate-fvs.shtml (accessed April 4, 2014).

Waring RH, McDowell N, 2002. Use of a physiological process model with forestry yield tables to set limits on annual carbon balances. Tree Physiology 22(2-3):179-188.

Waterworth RM, Richards GP, Brack CL, Evans DMW, 2007. A generalised hybrid process-empirical model for predicting plantation forest growth. Forest Ecology and Management 238(1–3):231-243.

Weiskittel AR, Crookston NL, Radtke PJ, 2011. Linking climate, gross primary productivity, and site index across forests of the western United States. Canadian Journal of Forest Research 41(8):1710-1721.

Woods AJ, Heppner D, Kope HH, Burleigh J, 2010. Forest health and climate change: A British Columbia perspective. Forestry chronicle 86(4):412-422.

Woodward FI. 1987. Climate and plant distribution Cambridge University Press, Cambridge [Cambridgeshire]; New York.

Xiaolu Z, Changhui P, Qing-Lai D, Jiaxin C, Sue P, 2005. Predicting forest growth and yield in northeastern Ontario using the process-based model of TRIPLEX1.0. Canadian Journal of Forest Research 35(9):2268-2280.

Chapter 2 Climate and Soils-based Models of Site Productivity in Eastern U.S. Tree Species

Abstract

As concerns rise over potential effects of greenhouse-gas-related climate change on terrestrial ecosystems, forest managers require growth and yield modeling capabilities responsive to changing climate conditions. Our goal was to develop prediction models of site index for eastern US forest tree species with climate and soil properties as predictors, for use in predicting potential responses of forest productivity to climate change. Species-specific site index data from the USDA Forest Service Forest Inventory and Analysis program were linked to contemporary climate data and soil properties mapped in the USDA Soil Survey Geographic SSURGO database. Random Forests regression-tree-based ensemble prediction models of site index were constructed based on 37 climate-related and 15 soils attributes. In addition to species-specific site index, aggregate models were developed for species grouped into two broad categories for conifer (softwood) and broadleaved (hardwood) species groups. Species-specific models based on climate and soils predictors explained the most variation in site index of any models tested (R^2 = 62.5%, RMSE 3.2 m). Comparable results were found when grouping species into conifer and hardwood groups (R^2 = 63.9%, RMSE 4.6 m for conifers; R^2 = 35.9%, RMSE = 4.2 m for hardwoods). Model predictions based on multiple global circulation models and Intergovernmental Panel on Climate Change development scenarios were tested for statistical significance using bootstrap resampling methods. Results showed significant increases over the 21st Century in mean site index for conifers ranging between +0.5 and +2.4 m. Over the same time period, mean hardwood site index showed decreases of as much as -1.7 m for the scenarios tested. The results demonstrate the utility of using climate and soils data in predicting site index across a large geographic region, and the potential of climate change to alter forest productivity in the Eastern United States.

Key words: bootstrap; climate change, climate envelope models; Random Forests; regression trees; site index

Species nomenclature: http://apps.fs.fed.us/fiadb-downloads/REF_SPECIES.CSV

2.1 Introduction

Greenhouse-gas-related global warming has led to changing patterns of precipitation and temperature over much of North America in the past half-century (IPCC 2007), with the potential for significant impacts on forests and other terrestrial ecosystems (Yaussy et al. 2013). Monserud et al. (2008) and Weiskittel et al. (2011) suggest that the potential for site productivity to be altered under climate change depends on location and species, but previous work has primarily focused on forests in western North America. As the demands on forests grow it is increasingly important to develop climate-sensitive modeling tools to assist forest managers and policy makers in making cost-effective decisions in forest resource management (Crookston et al. 2010).

Site index is based on the mean height of dominant and co-dominant trees and the age of the stand. Commonly used as a forest site quality measure, the concept of site index assumes forests are even-aged and mono-specific (Skovsgaard and Vanclay 2008). In practice, many forests in the eastern US are composed of multi-aged, mixed-species cohorts. Consequently, limitations of using site index as a measure of site productivity are often encountered in the eastern US and elsewhere. As a result, other methods have been explored including predicting observed site index from climate, topography and soil attributes (Sharma et al. 2012).

Predicting site index from site attributes has been examined at a number of geographic areas within the United States (US), Canada, and elsewhere (Beaulieu et al. 2011; Sharma et al. 2012; Weiskittel et al. 2011). Previous studies evaluated different modeling techniques, identified the most influential ecological variables, quantified the effects of scale and scaling on predictive modeling, and provided spatial maps of site index. A few studies have been conducted to assess the potential effect of climate change on spatial patterns of site index over time (Monserud et al. 2008; Weiskittel et al. 2011). To our knowledge, there have been no similar studies conducted across the entire eastern US.

A variety of statistical approaches have been used to predict potential productivity of forest stands, ranging from multiple linear regression to artificial neural networks (Aertsen et al. 2010). Nonparametric techniques such as regression-tree-based methods are among the most flexible and robust to modeling challenges such as correlated predictors, nonlinear or non-monotonic relationships, and variable interactions, along with the ability to include either categorical or continuous variables (Prasad et al. 2006; Wang et al. 2005). In addition, nonparametric techniques focus on reducing predictive error instead of relying on *a priori* formulations of interaction effects, so that they can be more useful in exploring hidden structures in ecological data sets than traditional statistical methods (Prasad et al. 2006).

Simulating changes in vegetation characteristics with respect to environmental variables can be extremely complex, posing significant challenges to traditional parametric regression analysis (Prasad et al. 2006). Random forests (RF; Breiman 2001) has recently gained increasing attention for its ability to account for nonlinear model forms involving many predictors and complex interactions among variables, especially in large data sets (Prasad et al. 2006). RF is a nonparametric ensemble classification and regression tool, which constructs collections of regression trees – when the response variable is continuous – with each tree based on a different sample of observations from a training data set selected using bootstrap sampling. In addition to the bootstrap selection of sample observations, bootstrapping is also used to choose what predictors are available for use at each node in regression trees. Bootstrapping provides a way of independently testing regression model results with "out of bag" (OOB) information, namely the observations and predictors excluded from a particular bootstrap sample used in an analysis step. The fitted model in RF is a collection of regression trees trained to in-bag and tested on OOB observations. No closed form solution is readily available, but the RF object can easily be used to make new predictions. When used as an ensemble, the bootstrapped regression trees deliver predictions that are robust to noise and potential nonlinear relationships between predictors and a response variable. In addition, the method is typically unaffected by predictors subject to between-variable correlation or complex interactions. Finally, RF is known to be relatively unaffected by problems of overfitting that often arise in regression models fitted to a large numbers of predictors (Cutler et al. 2007). A number of applications have used RF models to predict species habitat and abundance (Iverson et al. 2008; Joyce and Rehfeldt 2013; Rehfeldt

et al. 2006). Weiskittel et al. (2011) used RF for predicting site index in the western US, but to date it has not been used for modeling site index in the eastern US.

Our goal was to develop models of site index for eastern US conifer and hardwood tree species using RF with climate and soils variables as predictors. These models are intended for use in identifying the potential responses of eastern forests to climate change over a broad geographic range in eastern North America. To pursue this goal, two specific questions were investigated: 1) the degree to which prediction efficacy is reduced in modeling site index when only broad species grouping (conifers vs. hardwoods) is considered, compared to predictions from species-specific models; and 2) the degree to which using soils and climate predictors of site index together improves model accuracy compared to using climate or soils predictors alone. We also set out to demonstrate how the models developed can be used to predict how site index may change in the eastern U.S. over the 21st century under several climate change scenarios. To make long-term projections more robust, a bootstrap-based technique was developed to assess the statistical significance of predicted site index changes over the next century.

2.2 Materials and Methods

2.2.1 Study area

The study area was defined to span the geographic ranges of most forest tree species growing in the eastern US, which we defined as those east of the Great Plains physiographic province (Fenneman 1946). Data were compiled from 37 states that included areas east of 100° W longitude, comprising states of North and South Dakota, Nebraska, Kansas, Oklahoma, Texas, and all states further east.

2.2.2 Vegetation data

Vegetation data were compiled through the publicly-available online database of the USDA Forest Service, Forest Inventory and Analysis (FIA) national program. FIA uses a repeated measure panel design (Bechtold and Patterson 2005). Typically in the eastern US, states are surveyed using a 5-panel or 7-panel design. Under the 5-panel design 20% of the plots are observed each year. Under the 7-panel design 16% of the plots are observed each year. We

selected the most recent complete set of panels for each state (last panel observed in 2010 or 2011) for analysis. Only plots measured under the national annual inventory design (Gillespie 1999), excluding plantations and with no observable silvicultural treatments were used in the study. We excluded plantations and plots with silvicultural treatments because we wanted to avoid forests influenced by management activities designed to improve productivity, or those where genetically improved planting stocks are grown.

The primary variable of interest was site index, a measure of site productivity defined as the average total height (m) that dominant and co-dominant trees attain at a specified base age in even-aged stands (Skovsgaard and Vanclay 2008). Site index recorded in the FIA database for each field plot is based on dominant and co-dominant "site trees" measured for total height and age at breast height (1.37 m or 4.5 feet), with age determined from increment core extraction and field examination of tree rings. Roughly 45% of all FIA field plots had site tree measurements recorded, with the site tree measurements generally well-represented over the forested areas of the eastern U.S. Only the site trees assigned a base age of 50 years, encompassing 98.5% of the FIA site index observations, were used here in model development. Exclusion of plots with other base ages, e.g., 25 years, facilitated comparisons among species and the grouping of data into two broad species groups, conifers and hardwoods (Table 2.1). Site index was then calculated separately from site trees, either by species or species group, and averaged on each plot.

Since FIA plots are not limited to single-species, even-aged forests, we performed some exploratory analyses to determine how much this concern might affect plot records of site index obtained from the FIA field measurements of dominant tree heights and ages. Estimates of annual MODIS-derived gross primary productivity (GPP) on a 1-km grid generated from the MOD17 algorithm (Running et al. 2004) were obtained from the Numerical Terradynamic Simulation Group repository at http://www.ntsg.umt.edu/modis/. Site index values from hardwood and conifer site species were compared to GPP estimates for FIA field plots using Spearman's rank correlation coefficients and inspected graphically. To check the prevalence of mixed-species or uneven-aged forest conditions in the site tree data, we counted the numbers of site-species represented on each plot and the ranges of site-tree ages on each plot.

2.2.3 Climate data

For model development, climate data were compiled for a contemporary (observed) period of time. For model application, climate data were compiled for a future time period. Contemporary climate data were obtained from climate station records spanning three decades (1961 to 1990), chosen to overlap with the development of present-day intermediate-aged and mature forests in the eastern US (Crookston 2012). Future climate conditions were projected from widely-available general circulation model (GCM) predictions, characterizing decades having midpoint years 2030, 2060, and 2090, to match time-periods over which many eastern forests will be subjected to changing climate conditions resulting from elevated worldwide greenhouse gas emissions (IPCC 2007).

A suite of climate variables (Table 2.2) was compiled for both contemporary and future periods based on averaged monthly values for maximum, mean, and minimum daily temperatures, monthly total precipitation, and several derived annual climate variables potentially related to tree growth, such as growing season precipitation (Rehfeldt et al. 2006). Spatial resolution was increased from 2.8° to .0083° (about 1 km) using thin plate splines to downscale predictions of future climate made with coarse-resolution GCMs (Rehfeldt et al. 2006).

Future climate data were based on three different GCMs and several development scenarios, or storylines, formulated by the Intergovernmental Panel on Climate Change (IPCC; Table 2.3). The choice of GCMs was based on their availability as downscaled future climate data (Moscow Forestry Sciences Laboratory 2013), in the same format and having the same variables as were used in developing the site index models from contemporary conditions, and following recent work modeling forest dynamics in response to climate change that adopted the same GCMs (Crookston et al. 2010; Iverson et al. 2011). The four storylines, A1B (global rapid economic growth with balanced energy fossil fuel and non-fossil-fuel technology), A2 (regionally oriented economic development), B1 (global environmental sustainability), and B2 (local environmental sustainability) represent different patterns in demographic, social, economic, technological and environmental development worldwide. Scenarios B1 and B2 are more environment-oriented when compared with A1B and A2. Scenarios examined here were chosen to cover a range of

cumulative CO_2 emissions over the 21st century, with the A2 scenario having the highest atmospheric CO_2 accumulation, followed by A1B, B2, and B1 in descending order (IPCC 2000).

Both contemporary and future climate data were downloaded from the USDA Forest Service Rocky Mountain Research Station online climate data repository (Moscow Forestry Sciences Laboratory 2013). The repository was queried for contemporary and future climate data based on the published geographic coordinates and elevation of FIA field plots. Small differences in actual plot coordinates and published values were deemed unimportant for climate variables in the eastern US based on their strong spatial autocorrelation over distances ≤ 1 km (Wang et al. 2011).

2.2.4 Soils data

Soil types and their associated attributes were compiled from the USDA Soil Survey Geographic (SSURGO) online database, linked to FIA field plots based on their actual geographic coordinates. Available soil survey data for the eastern US were downloaded from the USDA NRCS Geospatial Data Gateway server (USDA Natural Resources Conservation Service 2013). At the time of downloading, soils data were unavailable for some parts of several states in the eastern U.S., most notably Alabama, Georgia, Maine, Minnesota, Mississippi, New Hampshire, and New York. Most void areas were small – the size of one or two counties – except for a larger area in northwest Maine and several relatively large tracts of public lands in northern Minnesota. Various soil properties related to the growth or survival of forest tree species based on the findings of Iverson et al. (2008) were included as predictors in model development (Table 2.4).

2.2.5 Model development

Random Forests regression was employed for developing climate-sensitive, regression-tree-based ensemble models for predicting site index. The RF algorithm builds a set of independent regression trees; each of which is constructed using bootstrapped observations known as "in bag" samples from a training data set. The remaining OOB observations are used to calculate mean square error (MSE) for the related regression tree, and the square root of MSE (RMSE). MSE is a measure of model accuracy expressed as $\frac{\sum_{i=1}^{i=n_{OOB}}(y_i-\hat{y}_i)^2}{n_{OOB}}$, *w*here n_{OOB} is the number of OOB

observations for each regression tree, y_i the ith OOB observed value (for site index in this case) and $\hat{y}_i$ the prediction of y_i. Because only the OOB observations are used in calculating the RMSE of a regression tree, y_i and $\hat{y}_i$ are independent. The ensemble RMSE, one of the RF predictive performance measures, results from averaging the RMSE over all regression trees. The adjusted coefficient of multiple determination (R^2), which adjusts for the number of predictors used in model fitting, was also used here as a measure of model goodness-of-fit (Kutner 2005).

2.2.6 Model comparisons

In comparing species-specific site index models to those having species grouped into conifers or hardwoods, climate variables and soil properties (Table 2.4, Table 2.2) were included as predictors. Species-specific models were fitted by including "species" as a categorical predictor in training a RF model to observed site index. Group-specific models were fitted using a binary categorical predictor to specify conifers or hardwoods. Predictions from competing models were compared by the examination of simple correlation coefficients, with reasoning that strong correlations between model-predicted site index for individual species versus species groups indicate similar performance of the competing models.

Site index prediction RMSE from OOB regression trees was also calculated as a measure of model accuracy to compare species versus group-specific RF models. Efron (1983) showed that the average fraction of observations included in any bootstrap sample is 0.632. Applied to RF, this property implies that the number of regression trees for which a single observation is OOB will be, on average, (1 - 0.632)×n_{tree}, where n_{tree} is the number of regression trees in the ensemble RF. For a collection of n FIA field plots, a total of n×n_{tree} regression tree predictions are possible; however, only about (1 - 0.632) n×n_{tree} of them are OOB predictions, i.e. predictions that are independent of observed values. The collection of OOB regression tree predictions of site index was thus compared to observed site index values in calculating the corresponding RMSE.

Reverse null hypotheses (Reynolds 1984) were formulated to determine the minimum negligible detectable difference (ϵ^*) that could be specified to arrive at a conclusion of "equivalence" of

competing models (Radtke and Robinson 2006). The reverse tests used the following OOB difference statistic

$$D_i = \widehat{SIGRP_i} - \widehat{SISPP_i} \quad [2.1]$$

where

$\widehat{SIGRP_i}$ = group-specific (conifer or hardwood) OOB prediction of site index for plot *i*

$\widehat{SISPP_i}$ = species-specific OOB prediction of site index for plot *i*

The mean OOB predicted difference across all *n* plots where a species occurred was then used to formulate a two-tailed reverse t-test to find ϵ^*.

$$P\left(t < \left|\frac{|\overline{D}| - |\epsilon^*|}{\sigma_{\overline{D}}}\right|\right) = \alpha \quad [2.2]$$

where

$\overline{D} = \frac{1}{n}\sum_{i=1}^{n} D_i$

$\sigma_{\overline{D}}$ = estimated standard error of $\overline{D}$

α = .05

Reverse tests were performed for the ten most frequently observed conifers and ten most frequently observed hardwood species, along with a comparison for all species grouped together.

The R package "randomForest" (Liaw and Wiener 2002; R Core Team 2012) was used for model development, with the specified number of regression trees used in each ensemble RF (n_{tree} = 300) reduced from the default value of 500 after verifying that OOB error consistently reached minima when n_{tree} was much smaller than 500. Field observations of site index were used as the response variable, with 15 soils and 37 climate variables used as predictors. Even though some FIA plots had no site tree information recorded, a large number of observations were available for both conifers (n = 21,531) and hardwoods (n = 31,776). Because some forested areas of the eastern US are not covered by the SSURGO database soil maps, an alternative set of RF models, one for hardwoods and one for conifers, was developed using climate variables alone as predictors. For comparison, site index models based on just the 15 soils variables were developed to test how soil properties alone performed in predicting site index.

Permutation accuracy importance was used to assess which variables were most meaningful in predicting site index. The "randomForest" package reports this measure as "%incMSE," the percentage increase in OOB MSE computed for non-permuted versus randomly-permuted predictors (Genuer et al. 2010; Grömping 2009). This importance measure is regularly used in studies that examine variable-importance using RF. Details of the %incMSE measure are given in Liaw and Wiener (2002). The measure of variable importance known as node impurity was not used here because of its bias in disproportionately selecting variables having larger numbers of values (Breiman 1984). The conditional permutation accuracy importance measure of Strobl et al. (2008) was not used here despite its favorable properties in reliably ranking variable importance in RF, because its computational requirements were prohibitive with the large data set used here.

2.2.7 Site index prediction

Contemporary predictions of site index for each FIA plot were obtained using the RF out-of-bag prediction function. Out-of-bag predictions are based only on the RF regression trees that were trained using bootstrap samples that excluded the plot for which a prediction was made. This step removes the dependency between contemporary site index predictions and the values observed on a particular field plot. Assuming the dependency would be minimal for predictions using

future climate data as inputs, all RF predictions – both in-bag and OOB – were used in predicting future site index. Future site index predictions were obtained using RF models with future climate conditions in 2030, 2060, and 2090 specified as inputs. To simplify reporting, only results from 2090 are presented here. Soil properties, as defined by the SSURGO data, were assumed to remain unchanged at a particular plot location from contemporary to future conditions, so SSURGO data were used as inputs without alteration between contemporary and future scenarios. Where no soils data were available, we used RF regression models based only on climate variables to predict future site index. Elsewhere, both climate and soils were used. For each IPCC development scenario, projections based on inputs from one or more GCMs were averaged (Table 2.3) to account for differences between different GCM predictions run under the same scenario.

The bootstrap-generated collection of regression trees in RF generally provides a *de facto* method of testing hypotheses or generating prediction intervals for point estimates obtained from the ensemble regression model. Such was not entirely the case here because, while the RF models provided point and interval estimates of either contemporary site index or future site index, the estimates of interest here involved the change in site index (ΔSI) between predicted 2090 site index (SI_{2090}) and contemporary ($SI_{contemp}$) values, i.e. $\Delta SI = SI_{2090} - SI_{contemp}$.

To facilitate hypothesis testing, we replicated the training data B = 120 times using bootstrap sampling, then trained a different RF to each of the B training data sets. The resulting models were used to generate B predictions of both SI_{2090} and $SI_{contemp}$ for each FIA plot, which were then subtracted to obtain B predictions of ΔSI for each GCM and emissions scenario. Thus, bootstrapped confidence intervals (Efron 1994) were placed around ΔSI estimates to determine if predicted change was 'significant' or not ($\alpha = .05$). Based on a null hypothesis H_0: $\Delta SI = 0$, only plots having 95% bootstrap confidence intervals that did not include zero were deemed to provide significant model evidence of changing site index over the 21st century. Frequency distributions and thematic maps of ΔSI under various emissions scenarios were examined and summarized to assess potential impacts of climate change on productivity in eastern forests.

2.3 Results

2.3.1 Site Index Measurements

Field-obtained site index estimates for FIA plots were grouped as conifers and hardwoods for comparison to MODIS GPP. To account for the nonlinear nature of the relationship between site index and GPP, Spearman's rank correlation was used to measure the strength of any relationship between GPP and site index. Rank correlation coefficient values of 0.612 (Figure 2.1A) and 0.255 (Figure 2.1B) were calculated, respectively, for conifer and hardwood species. Both values provided strong evidence ($p < .001$) to reject a null hypothesis of no correlation between site index and satellite-derived GPP. In conifers, 93% of the site index measurements involved only one species of site tree on a plot. In hardwoods the value was slightly lower at 83% of plots having a single species of site tree observed. The ranges of site tree ages on plots were ≤ 5 years in 86% of conifer plots and 73% of hardwood plots.

2.3.2 Group versus species-specific models

Total variation in observed site index across the eastern US was larger for conifers than for hardwoods, with standard deviations of 7.5 m and 5.2 m, respectively. The species-specific RF model with climate and soils as predictors explained 62.5% of overall variation in site index across the eastern United States. The conifer and hardwood group model explained only 50.6% of the overall variation. Prediction errors for the grouped conifer and hardwood model were generally higher than the species-specific model, with an overall average difference in prediction RMSE across species of 0.56 m (Table 2.5). Results for all but two species, northern white cedar and yellow-poplar, gave RMSE values within 1.0 m of the corresponding conifer and hardwood group model predictions. In all but two species, quaking aspen aspen and yellow poplar, correlation coefficients (r) between $\widehat{SIGRP}$ and $\widehat{SISPP}$ were ≥ 0.75. Predicted differences $\overline{D}$ (%) between grouped vs. species-specific model results were < 10% of observed species mean site index in 17 of the 20 species we studied, with northern white cedar, quaking aspen, and yellow poplar having the largest inconsistencies between $\widehat{SIGRP}$ and $\widehat{SISPP}$.

Hypothesized negligible differences of 1 m between group and species-specific models would result in a conclusion of "equivalence" in half of the species tested (Table 2.5). Prediction differences across all species; however, resulted in a very small ϵ^*. The smaller ϵ^* is, the more

convincing evidence is for equivalence of predictions between $\widehat{SIGRP}$ and $\widehat{SISPP}$. Results showed that the same three species having low correlations between $\widehat{SIGRP}$ and $\widehat{SISPP}$ had ϵ^* > 2 m. Due to their larger ϵ^* values, equivalence tests for these species would only be rejected if relatively liberal definitions of "equivalence" were specified.

2.3.3 Climate and Soils Predictors

Regression models based on both climate and soils predictors had lower RMSE than models based on either climate or soils alone (Table 2.6). Adding soils predictors to models already using climate variables resulted in a reduction of only 0.1 m in RMSE for either conifer or hardwood models. Using soils alone resulted in lower prediction accuracy than models having climate variables as predictors (Table 2.6). Climate and soils information explained a larger percentage of the total variation in conifer site index than in hardwoods. The fact that RMSE values for conifer and hardwood models were nearly the same (e.g., Table 2.6; models 1-2) resulted from the greater total variation in conifer site index than hardwoods across the study region.

2.3.4 Model Predictors

Model fit statistics improved notably when the number of predictors was increased from 2 to 4. Adding additional predictors beyond the first 4 resulted in smaller improvements. Beyond six predictors little additional improvement was noted (Figure 2.2). Both soils and climate variables were among the most important predictors of site index in both conifers and hardwoods, according to the permutation accuracy importance measures (Models 1 & 2; Table 2.7). Variables that measure within-year variation in precipitation and temperature, e.g., PRATIO and TDIFF, were found to be among the most important climate predictors. Climate variables that measure gross annual temperature and precipitation, e.g., DD5 and MAP, were noted among the top predictors as well, particularly in multiplicative interaction terms involving seasonal variables, AMI and PRATIOXDD5 (Table 2.7).

Relationships between various predictors and site index along with interactions between soils and climate variables were examined using marginal plots of observed conifer site index versus

PRATIOXDD5 for high (> 7) versus low (≤ 7) pH values (Figure 2.3 a-b). Data were stratified into high versus low pH since it was one of the most important soil variables in site index models 1 and 2 (Table 2.7). The shapes of trends between climate variables and site index were generally not much different for plots having low pH values compared to those with higher pH, although far fewer plots exceeded pH values of 7 than those below 7 (Figure 2.3, left vs. right panels). Despite trend shapes being similar, differences in the magnitudes of the overall trends were noted for sites having high vs. low soil pH (cf. Figure 2.3 c vs. d and e vs. f).

To illustrate an approach for interpreting model relationships and identifying potential causal factors with regression-tree based modeling (McKenney and Pedlar 2003), an example regression tree was trained for loblolly pine (*Pinus taeda* L.) site index using soils and climate variables as predictors (Figure 2.4). The primary split in the loblolly pine model was based on the minimum temperature (°C) in the coldest month (MMIN; see Table 2.2) where cold winters (MMIN < -0.65) were generally associated with low-ranking site index predictions (Figure 2.4). Following the primary split, additional splits were made based on two nearly identical measures of summertime high temperatures, namely MMAX and MTWM. The second split of the left branch in the regression tree occurred for MMAX ≥ 33.05, thus placing those sites with both warm summer high temperatures and cold winter low temperatures as the second lowest of any group in the regression tree terminal nodes, with a predicted site index of only 24.61 m. The highest predicted site index resulted from a combination of factors, including sites that experienced relatively mild winters, a narrow range of summer high temperatures (26.65 ≤ MTWM < 27.75), nearly flat terrain (SLOPE < 1.95%), and above average annual moisture index (AMI ≥ 3.77).

Marginal plots and rank correlation coefficients were generated using predicted conifer ΔSI from 1990 to 2090 along with climate variable changes predicted under IPCC development scenario A2 (Figure 2.5). The strongest correlations involving ΔSI and climate variables – nine of which are shown in Figure 2.5 (see Table 2.2 for definitions) – include the following model predictors: SMI and AMI, which quantify the ratio of growing-degree-days to precipitation; SDAY, FDAY, and FFP which characterize the length of the frost-free season; and GSDD5, DD5, MAT, and MMAX, which describe mean and accumulated growing-season temperatures as well as summer

high temperatures. Changes in any of these variables that coincided with predicted increases in either the moisture indices, the length of the summer frost-free period, or increasing ambient temperatures, especially during the growing season, all corresponded to overall decreasing site index.

Only the top four most important predictors are listed in Table 2.7, but subsequent results were generated from models that included all predictors. Although Figure 2.2 indicated that a relatively small number of predictors would sufficiently explain a significant amount of the variation of site index, all climate and soil variables were retained in final models to ensure maximum predictive accuracy (Genuer et al. 2010).

2.3.5 Mapping site index

Site index models generally showed a pattern of increase geographically from north to south for conifers and from north and west to southeast for hardwoods when predicted from Models 1-2 using soils data and the contemporary climate conditions (Figure 2.6). Model predictions for conifers resulted in site index values below 20 m for most states north of 40° N latitude, with the most extreme low values noted in Maine and the Midwest. In hardwoods the lowest site index predictions occurred in Texas and other states west of the Mississippi River (Figure 2.6). Hardwood site index values > 20 m were typically predicted only south of 45° N latitude and in southern states east of eastern Texas.

Changes in site index predicted under various IPCC emissions scenarios showed the potential for either increasing or decreasing forest productivity by the end of the 21st Century, depending on the scenario considered and species group (Figure 2.7, Figure 2.8). Predicted site index changes (ΔSI) exhibiting significant evidence of change based on bootstrap results, i.e. where evidence was strong to reject Ho: $\Delta SI = 0$, were plotted in the maps and frequency graphs shown (Figure 2.7, Figure 2.8). The overall mean ΔSI in conifers increased in scenarios A1B ($\overline{\Delta SI}$ = 1.7 m), A2 ($\overline{\Delta SI}$ = 0.5 m), and B1 ($\overline{\Delta SI}$ = 2.4 m). While small in absolute terms, changes represented on a percentage basis were larger, representing 8.1%, 2.3%, and 11.4% increases in site index, respectively for scenarios A1B, A2, and B1. Site index decreased overall under the B2 scenario by 1.0 m, or 4.6% of the average contemporary site index across the Eastern United States.

Histograms of significant ΔSI predictions over the 21st Century showed only a small number of the predictions corresponded to absolute changes of site index > 10 m, with most of the differences within ± 5 m of contemporary (Figure 2.7, Figure 2.8). Over the entire study area, average hardwood ΔSI decreased for scenarios A1B ($\overline{\Delta SI}$ = 0.2 m), A2 ($\overline{\Delta SI}$ = 1.7 m), and B2 ($\overline{\Delta SI}$ = 1.1 m), while a mean increase ($\overline{\Delta SI}$ = 0.8 m) in predictions was noted for scenario B1 (Figure 2.8).

Spatial patterns of ΔSI over time varied between conifers and hardwoods and depending on the climate scenario examined. Areas of significantly decreasing site index in conifers included southern-tier states from Texas to South Carolina (Figure 2.7). Areas of potential increasing conifer site index included New England, western Lakes States, and the lower Midwest (Figure 2.7). Several regions showed conflicting results between scenarios, with either significant increases or decreases in conifer site index occurring in the Mid-Atlantic States and Ohio River Valley, depending on the climate scenario studied. Areas of significantly decreasing site index in hardwoods were noted across a broader area of the South and in central states including Indiana, Ohio, and Kentucky (Figure 2.8). Significant increases in hardwood ΔSI were consistently predicted in New England, central Texas, and northern Michigan. Hardwoods site index changes in some states, e.g., Missouri, Pennsylvania, and Wisconsin, were either positive or negative, depending on which IPCC scenario was examined.

Relative values of ΔSI meeting the bootstrap criterion for significant change were mapped to show the potential for productivity increases or decreases as a percent of contemporary site index (Figure 2.9). Relative results showed decreases in conifer site index as large as 60% under the A2 scenario in some areas including Wisconsin and Minnesota. In hardwoods, relative decreases as large as 40% were predicted widely across the South from eastern Texas to central Florida (Figure 2.9). For conifers, increases > 80% were noted in areas of Missouri and Kansas. Hardwood site index increases > 80% were likewise observed in north-central Texas.

2.4 Discussion

As forest management changes to address concerns over greenhouse-gas-related warming, new models will be needed that incorporate climate variables as predictors (Nothdurft et al. 2012).

Forest site productivity models that incorporate climate and soils like those developed here can operate at multiple scales and account for factors relevant to forest management such as species or species group. These models are complementary to approaches that incorporate remote sensing data or predictions from process-based models with regional forest inventories (Kwon and Larsen 2013).

A concern with the use of FIA plots for modeling site index is that no controls are established to ensure that plots are installed only in even-aged, single-species forests or stands. Yet, the concept of site index assumes forests are even-aged and mono-specific (Skovsgaard and Vanclay 2008). The selection protocol for site trees on FIA field plots states "only trees that have remained in a dominant or co-dominant crown position throughout their entire life span" should be selected (USDA Forest Service 2012), a criterion aimed at meeting one of the underlying assumptions of site index determination. In addition, trees that are damaged or show signs of past suppression in tree rings are not selected. When no suitable site trees are available, as judged by field crews, the measurements for site tree heights and ages are left blank for the plot. Cursory analyses showed that large majorities of plots had just one species of site tree selected, 83% and 93% for hardwoods and conifers, respectively. Large majorities of plots also showed relatively low variation in the ranges of site-tree ages. While there's no doubt that a significant proportion of FIA plots are established in multi-aged or mixed species stands, the proportion of plots having site trees that reflect multi-aged or mixed-species compositions is relatively small.

Despite the concern over failing to meet all assumptions of site index measurement or modeling, the strength of correlation between FIA-observed site index and satellite-derived GPP indicated that FIA site index is at worst weakly correlated with GPP in eastern hardwoods, or at best strongly correlated with GPP in eastern conifers (Weiskittel et al. 2011). Numerous studies have shown strong positive relationships between site index measured on different species over broad regions or functional groups such as conifers and hardwoods (Carmean and Hahn 1983; Doolittle 1958; Nigh 1995). Previous work showed direct volume growth based measurements of net primary productivity (NPP) from FIA plots were weakly to moderately correlated with MODIS-based NPP (Kwon and Larsen 2013). National inventory data are frequently used in

developing site index or productivity models despite the limitations of such data (Nothdurft et al. 2012; Sharma et al. 2012).

We divided species into two broad groups, conifers and hardwoods to avoid the uncertainty associated with potential shifts in species geographic distributions that may result from climate change (Iverson and McKenzie 2013; Joyce and Rehfeldt 2013). An important assumption for this approach was that tree species within groups shared the same responses to climate and soils variables. In reality, the impacts of climate and soils vary by species (Iverson et al. 2008; Rehfeldt et al. 2006). As a result, model accuracy of group-specific models was lower than for species-specific models. While group-specific models did not perform as well as species-specific models, equivalency thresholds of 2 m or less were noted for 85% of the species studied (Table 2.5). An interpretation of this result is that, if a mean difference between model predictions of up to 2 m can be considered unimportant, hypotheses tests will result in conclusions of "equivalence" for such species (Reynolds 1984). Reporting ϵ^* (e.g., Table 2.5) has an important distinction over tests where a minimum important effect size (MIES) is specified *a priori*. Rather than concluding two models are "equivalent" for a specific MIES, knowing ϵ^* provides a way to determine the smallest MIES that could be specified and result in a conclusion of equivalence (e.g., Radtke and Robinson 2006).

The eastern species-group site index predictive models in this study were developed linking plot-based vegetation measurements with climate and soils information. Overall, our models were consistent with previous findings, with adjusted R^2 values between 35.9% and 63.2% and RMSE ranging from 4.2 to 4.5 m, roughly 20% of the mean SI value across the eastern US (Table 2.6). Brown and Loewenstein (1978) used multiple regression to find that soil and topographic variables, along with stand age, explained 70% of the variation in site index in mixed coniferous forests of northern Idaho. Using several modeling techniques Aertsen et al. (2011) developed models that explained between 40% and 80% of variation in site index for three tree species in the Flanders region of Belgium. The multiple regression model of Nigh (2006) explained 55% of variation in inland Douglas-fir (*Pseudotsuga menziesii* [Mirb] Franco var. glauca) site index in British Columbia based on climate variables. It's important to note that regression fit statistics such as those reported in studies referenced above are typically based on "in-bag" observations,

where the same observations used in fitting models were also used in calculating prediction errors. Such in-bag estimates of standard errors – what Efron (1983) referred to as "optimistic" – usually underestimate standard errors. In contrast, the error rates reported here were based entirely on comparisons of model predictions with out-of-bag observations.

In a study from the western United States, Weiskittel et al. (2011) developed a RF model of site index for 20 conifer species grouped together that explained 78% of overall variation. The OOB RMSE of their model (4.6 m) was consistent with the models developed here. The hardwood models developed here had a slightly higher accuracy than conifer models as indicated by RMSE, despite their explaining a lower portion of total variation in site index (Table 2.6). This result follows from the fact that total variation in conifer site index was larger than the total variation in hardwood site index across the eastern United States.

Limitations of climate-envelope models like those developed here have been discussed in some detail, especially as they relate to predicting habitat or geographic distributions of tree species (Iverson et al. 2011; McKenney et al. 2007). The problem regarding changes in species geographic distributions over time is an example. One important assumption made here is that biotic interactions' effects on site index are uniform spatially and temporally (Jeschke and Strayer 2008). Another is that the interaction of genetics and climate remain constant over space and time. We also assumed that soil properties remain static over the time scales for which the models will be used, which may be unreasonable over long time scales if soil forming processes are accelerated by warming and increased precipitation (Akin 1991, Ch. 7).

Results showed that climate and soils together can produce slightly more accurate predictions than when climate alone is used to predict site index in forests across the eastern United States. It follows that, in areas where SSURGO data sets are unavailable, or where relative simplicity is desired, climate-only models can be used in place of the climate-and-soils models without much loss of precision. Models that used only climate variables as predictors were somewhat more accurate than those that used only soils variables (Table 2.6). Because the published geographic coordinates for FIA plots – purposefully shifted from the true plot locations to preserve landowner privacy (McRoberts et al. 2005) – were used in merging vegetation data with climate variables, we expected some additional measure of error. Evidently this error was small given the

outcomes of climate-only model fits. This result is sensible given the prominent role elevation plays in predicting climate attributes across space given that elevations between published and actual plot locations are highly correlated over small (< 1 km) distances (Rehfeldt 2006; Wang et al. 2011). With soils, we aimed to attain a higher degree of predictive power by using fine resolution (1:24,000) SSURGO soils data rather than the coarser-scale (1:250,000) STATSGO data (Iverson et al. 2008; USDA Natural Resources Conservation Service 1994). In contrast to the climate variables, precise plot coordinates were used to merge FIA vegetation data to SSURGO soils map units, also in an attempt to increase the efficacy of soils data in modeling site index. Despite the comparatively high spatial resolution of SSURGO vs. STATSGO data, considerable variation in soil components within map units and the spatial uncertainty of map-unit polygon boundaries likely contributed to the inability of soils alone in the work conducted here to explain variations in site index better, once climate variables were accounted for (Gatzke et al. 2011). Incorporating soils data into forest productivity models is a challenging problem due to limited scales and resolutions of mapped soils variables. (Aertsen et al. 2012b; Carmean 1975). Further, because general soil survey maps were primarily developed for agricultural and land use suitability applications, their utility in providing inputs to regional-scale ecosystem simulations or forest productivity models has been shown to be less than efficacious (Grigal 2009).

Although RF predictions are known to be relatively unaffected when correlations exist between predictors, using permutation accuracy as a measure of variable importance in such cases has been shown to be unreliable. Permutation accuracy can also be affected by the scale of predictor variables, or for categorical predictors having large numbers of classes (Nicodemus et al. 2010; Strobl et al. 2007). The conditional variable importance measure proposed by Strobl et al. (2008) is preferable for use in ranking the importance of predictors in RF models (Sabatia and Burkhart 2014); however, its use here was impractical due to the computational requirements of its implementation in the R package "*party"* (Strobl et al. 2009). The variable importance scores and ranks reported in Table 2.7 should be viewed with this consideration in mind. The models developed here should not be seen as tools designed to answer specific questions about which climate or soils variables are most strongly related to site index in eastern U.S. forests (Prasad et

al. 2006). With further computational advances in variable importance measures, future work might aim to pursue such questions more thoroughly.

When assessing differences between future and contemporary model-predicted site index, it was necessary to generate multiple RF models by bootstrapping to obtain prediction intervals for ΔSI. The intervals were useful for determining predictions where H_o: ΔSI = 0 was rejected so that only "significant" changes would be represented on maps (Figure 2.7, Figure 2.8). For emphasis it's relevant to restate here that the ΔSI values shown in Figure 2.7 and Figure 2.8 are all ensemble means whose 95% bootstrap confidence intervals did not include zero. It follows that any of the values shown that lie close to zero in the histograms necessarily have variances sufficiently small to reject a null hypothesis of ΔSI = 0. Results in Figure 2.5 showed that even some of the largest predicted site index changes – both increases and decreases – were found to be not statistically different from zero. This underscores the utility of the bootstrap interval estimates in the work. In other research, variance estimates were available from parametric models used to predict site index or from the training data directly, but unavailable for predictions made under novel conditions (Latta et al. 2009; Nothdurft et al. 2012).

Several studies have shown how greenhouse-gas related climate change may affect site index under a range of CO_2 emissions scenarios, with some species experiencing increases or decreases, and considerable regional variation (Bravo-Oviedo et al. 2010; Monserud et al. 2008; Nothdurft et al. 2012). Regional differences in site index responses to changing climate were noted here, along with differences between hardwoods and conifers. Few studies have examined potential site index changes in the eastern US so comparisons with other studies are somewhat limited. Prediction of potential habitats for eastern US tree species showed general movement of species habitats from southwest to northeast (Iverson et al. 2008; Joyce and Rehfeldt 2013). Despite the fact that we chose only a subset of IPCC climate scenarios and in some cases averaged predictions from two or three GCMs, future site index predictions can be made for any future climate scenarios, so long as they provide values for the model inputs. Future research may seek to combine predictions of habitat suitability with potential site index under various climate change scenarios to determine how well geographic shifts in species range match areas of sustained or increased productivity. In such work the species-specific site index modeling

approach developed here may be a better option than the relatively simple conifer versus hardwood approach emphasized in some of the results presented.

Of the climate-related variables tested for predicting conifer site index, permutation accuracy importance was greatest for PRATIOxDD5, which is highest in climates with relatively warm, moist growing seasons and gradually increases from north to south in contemporary records. Climate change is generally expected to increase PRATIOxDD5 somewhat evenly across the eastern United States (maps not shown). The marginal relationship of PRATIOxDD5 and conifer site index shows an optimal range, i.e. site index is greatest in a range of PRATIOxDD5 between 2000 and 4000, and lower outside that range (Figure 2.3 a-b). This variable may explain predicted increases in site index in areas such as New England that are below the optimal range in the contemporary record, as well as decreases in areas of the Deep South that are already near the upper end of the optimal range in the contemporary record. In contrast, some areas, such as southwestern Wisconsin that are below optimum PRATIOxDD5 in the contemporary record, are predicted to move into the optimal range with climate change; yet, our results show a significant probability that conifer site index will generally decrease there. Such results point to the complex relationships of site index to other climate and soils variables, perhaps with nonlinear or interacting effects. PRATIOxDD5 was also the most important predictor of site index in western conifers in the model developed by Weiskittel et al. (2011) and is consistent with previous work involving other western conifer species (Farr and Harris 1979; Nigh 2006). By comparison, Yeh and Wensel (2000) found that site index was most strongly related to summer temperatures and winter precipitation in conifer species in Northern California, an area where snowpack plays a relatively large role in soil water budgets compared to many areas of the East. The negative relationships noted between ΔSI and increasing temperature or moisture indices (Figure 2.5) were consistent with the findings of Sabatia and Burkhart (2014), who studied site index gradients in relation to climate and soils in planted loblolly pine forests.

Site index of conifers was also related to soil variables including pH, NO10, and TAWC (Table 2.7), which have been noted as meaningful predictors of forest productivity and height growth in past studies (Aertsen et al. 2012a; Uzoh 2001). For example, Gale et al. (1991) identified a range of soil pH between 5 – 7 as being optimal for root growth in white spruce (*Picea glauca*

(Moench) Voss) plantations, results that were consistent by inspection of site index vs. soil pH in the FIA data (results not shown here). Higher values of NO10 indicate a larger effective soil depth, and deep soils are typically considered to be more productive than shallow soils (Louw and Scholes 2006; Rhoton and Lindbo 1997). Soil water holding capacity is also widely known to be a meaningful predictor in models of forest productivity (Coops et al. 2011; Ercanli et al. 2008), as was noted here for the variable TAWC.

The example regression tree shown in Figure 2.4 is useful for illustrative purposes; however, it should not be taken to represent the full range of predictions possible from the models developed here. While it represents just one regression tree pertaining to one species, the Random Forests models are comprised of hundreds of regression trees and are capable of predicting site index for many species or species groups (Table 2.5). As such, it is particularly difficult to assign ecological interpretations from Random Forests models (Prasad et al. 2006; Sabatia and Burkhart 2014). The regression tree in Figure 2.4 shows that a single variable alone – MMIN in this case – despite being involved in the primary split in the tree, may not be able to distinguish between lowest and highest site index values in the training data. To separate FIA loblolly pine site index observations into groups having means differing by as little as 1 m, the regression tree used seven predictors in various combinations. Those combinations can be examined in Figure 2.4 to extract some information about the ecological and biological factors associated with high or low site productivity; however, due to the many variables involved and possible interactions among them, coupled with the observational nature of the data and the complexity of the Random Forests ensemble, such interpretations are unlikely to be applicable across the full range of predictions possible.

Interpreting the site index changes predicted here for the 21st century is not straightforward given the geographic variation and differences between conifer and hardwood responses to the emissions scenarios tested. Methods related to nearest-neighbor imputation (cf. Crookston and Finley 2008; Packalén et al. 2012) for identifying pairs or groups of similar observations may provide ways to better interpret how vegetation characteristics in the contemporary climate–soils envelope are related to future conditions under various climate change scenarios. By pursuing such approaches, it should be possible to answer questions such as “what contemporary

conditions at one point on the map are most similar to future conditions predicted (under different climate scenarios) at a different map location?" Despite our not being able to answer such questions from the results presented here, some consistent patterns were observed. One is that more FIA plots showed significant positive ΔSI and fewer showed significant negative ΔSI under scenario B1 than any other scenarios tested. Scenarios A2 and B2 consistently showed the opposite effect, giving rise to more FIA plots having negative future predictions of ΔSI and fewer having positive ΔSI predictions compared to the other scenarios, a result that matches the scenarios' effects on basal area predictions in several western species reported by Crookston et al. (2010). More detailed examination of model results is warranted to gain better understanding of causal relationships or to provide meaningful ecological interpretations.

2.5 Summary

Identifying the relationships between site index and climate or soil properties is an ongoing challenge. This study developed a suite of site index models based on both climate and soils that can be used to evaluate the impact of climate change on site productivity in eastern forest species. Site index predictions for species grouped as conifers or hardwoods were nearly as precise as species-specific models for many of the most common eastern forest tree species. Soil properties were somewhat less useful in predicting site index across eastern US forests as were climate variables; however, soils and climate used together provided slightly more predictive power than either climate or soils alone. Multiple maps were produced to illustrate contemporary patterns of site productivity across the region, and the potential for significant site index change with greenhouse-gas related global warming. A bootstrap procedure was implemented to determine if differences between contemporary and future SI predictions were statistically significant. Results like these may serve policymakers and forest managers in their attempts to plan for climate change impacts on regional forest ecosystem services management in coming decades. Spatial patterns of predicted site index change may also serve in identifying areas of concern for long-term forest management, or to focus efforts at monitoring potential changes in forest productivity. At present, our ability to interpret spatial patterns for meaningful interpretations is limited and requires further investigation.

2.6 References

Aertsen W, Kint V, De Vos B, Deckers J, Van Orshoven J, Muys B, 2012a. Predicting forest site productivity in temperate lowland from forest floor, soil and litterfall characteristics using boosted regression trees. Plant Soil 354(1-2):157-172.

Aertsen W, Kint V, Muys B, Van Orshoven J, 2012b. Effects of scale and scaling in predictive modelling of forest site productivity. Environmental Modelling & Software 31:19-27.

Aertsen W, Kint V, Van Orshoven J, Muys B, 2011. Evaluation of modelling techniques for forest site productivity prediction in contrasting ecoregions using stochastic multicriteria acceptability analysis (SMAA). Environmental Modelling & Software 26(7):929-937.

Aertsen W, Kint V, van Orshoven J, Özkan K, Muys B, 2010. Comparison and ranking of different modelling techniques for prediction of site index in Mediterranean mountain forests. Ecological Modelling 221:1119-1130.

Akin WE. 1991. Global patterns: climate, vegetation and soils. University of Oklahom Press, Norman, OK.

Beaulieu J, Raulier F, Pregent G, Bousquet J, 2011. Predicting site index from climatic, edaphic, and stand structural properties for seven plantation-grown conifer species in Quebec. Canadian Journal of Forest Research 41(4):682-693.

Bechtold WA, Patterson PL, 2005. The enhanced forest inventory and analysis program - national sampling design and estimation procedures Asheville, NC: USDA Forest Service, Southern Research Station. 85.

Bravo-Oviedo A, Gallardo-Andrés C, del Río M, Montero G, 2010. Regional changes of Pinus pinaster site index in Spain using a climate-based dominant height model. Canadian Journal of Forest Research 40(10):2036-2048.

Breiman L. 1984. Classification and regression trees Wadsworth International Group, Belmont, Calif.

Breiman L, 2001. Random Forests. Machine Learning 45(1):5-32.

Brown HG, Loewenstein H, 1978. Predicting site productivity of mixed conifer stands in northern Idaho from soil and topographic variables. Soil Sci. Soc. Am. J. 42(6):967-971.

Carmean WH, 1975. Forest site quality evaluation in the United States. In: Advances in Agronomy--Brady NC, ed. New York: Academic Press. 209-269.

Carmean WH, Hahn JT, 1983. Site comparisons for upland oaks and yellow poplar in the Central States. Journal of Forestry 81(11):736-739.

Coops NC, Gaulton R, Waring RH, 2011. Mapping site indices for five Pacific Northwest conifers using a physiologically based model. Applied Vegetation Science 14(2):268-276.

Crookston N, 2012. Details on spatial extents, temporal information and data elements http://forest.moscowfsl.wsu.edu/climate/details.php (accessed July 9, 2013,).

Crookston NL, Finley AO, 2008. YaImpute: an R Package for kNN imputation. J. Stat. Softw. 23(10):16.

Crookston NL, Rehfeldt GE, Dixon GE, Weiskittel AR, 2010. Addressing climate change in the forest vegetation simulator to assess impacts on landscape forest dynamics. Forest Ecology and Management 260(7):1198-1211.

Cutler DR, Edwards TC, Jr., Beard KH, Cutler A, Hess KT, Gibson J, Lawler JJ, 2007. Random forests for classification in ecology. Ecology 88(11):2783-2792.

Doolittle WT, 1958. Site index comparisons for several forest species in the Southern Appalachians. Soil Science Society of America Journal 22(5):455-458.

Efron B, 1983. Estimating the error rate of a prediction rule: improvement on cross-validation. Journal of the American Statistical Association 78(382):316-331.

Efron BTR. 1994. An introduction to the bootstrap Chapman & Hall, New York.

Ercanli I, Gunlu A, Altun L, Baskent EZ, 2008. Relationship between site index of oriental spruce Picea orientalis (L.) Link and ecological variables in Macka, Turkey. Scandinavian Journal of Forest Research 23(4):319-329.

Farr WA, Harris AS, 1979. Site index of Sitka spruce along the Pacific Coast relataed to latitude and temperatures. Forest Science 25(1):145-153.

Fenneman NM, and Johnson, D.W., 1946. Physiographic divisions of the conterminous United States Washington, D.C.: U.S. Geological Survey (USGS). Special map series, scale 1:7,000,000.

Gale MR, Grigal DF, Harding RB, 1991. Soil productivity index: predictions of site quality for white spruce plantations. Soil Sci. Soc. Am. J. 55(6):1701-1708.

Gatzke SE, Beaudette DE, Ficklin DL, Luo Y, O'Geen AT, Zhang M, 2011. Aggregation strategies for SSURGO data: effects on SWAT soil inputs and hydrologic outputs. Soil Sci. Soc. Am. J. 75(5):1908-1921.

Genuer R, Poggi J-M, and Tuleau-Malot C, 2010. Variable selection using random forests. Pattern Recognition Letters 31(14):2225-2236.

Gillespie AJR, 1999. Rationale for a national annual forest inventory program. Journal of Forestry 97(12):16-20.

Grigal DF, 2009. A soil-based aspen productivity index for Minnesota. Forest Ecology and Management 257(6):1465-1473.

Grömping U, 2009. Variable importance assessment in regression: linear regression versus Random Forest. The American Statistician 63(4):308-319.

IPCC, 2000. Special report on emissions scenarios: Summary for policymakers--Nakićenović N, Swart R, eds.: Intergovernmental Panel on Climate Change. 20.

IPCC. 2007. Climate change 2007: mitigation. contribution of working group III to the fourth assessment report of the Intergovernmental Panel on Climate Change Cambridge University Press, Cambridge.

Iverson LR, McKenzie D, 2013. Tree-species range shifts in a changing climate: detecting, modeling, assisting. Landsc. Ecol. 28(5):879-889.

Iverson LR, Prasad AM, Matthews SN, Peters M, 2008. Estimating potential habitat for 134 eastern US tree species under six climate scenarios. Forest Ecology and Management 254(3):390-406.

Iverson LR, Prasad AM, Matthews SN, Peters MP, 2011. Lessons learned while integrating habitat, dispersal, disturbance, and life-history traits into species habitat models under climate change. Ecosystems 14(6):1005-1020.

Jeschke JM, Strayer DL, 2008. Usefulness of bioclimatic models for studying climate change and invasive species. In: Year in Ecology and Conservation Biology 2008--Ostfeld RS, Schlesinger WH, eds. Boston: Blackwell Scientific. 1-24.

Joyce DG, Rehfeldt GE, 2013. Climatic niche, ecological genetics, and impact of climate change on eastern white pine (Pinus strobus L.): Guidelines for land managers. Forest Ecology and Management 295:173-192.

Kutner MH. 2005. Applied linear statistical models McGraw-Hill Irwin, Boston. 227-229.

Kwon Y, Larsen CPS, 2013. An assessment of the optimal scale for monitoring of MODIS and FIA NPP across the eastern USA. Environmental Monitoring and Assessment 185(9):7263-7277.

Latta G, Temesgen H, Barrett TM, 2009. Mapping and imputing potential productivity of Pacific Northwest forests using climate variables. Canadian Journal of Forest Research 39(6):1197-1207.

Liaw A, Wiener M, 2002. Classification and regression by randomForest. R News 2(3):18-22.

Louw JH, Scholes MC, 2006. Site index functions using site descriptors for Pinus patula plantations in South Africa. Forest Ecology and Management 225(1–3):94-103.

McKenney DW, Pedlar JH, 2003. Spatial models of site index based on climate and soil properties for two boreal tree species in Ontario, Canada. Forest Ecology and Management 175(1-3):497-507.

McKenney DW, Pedlar JH, Lawrence K, Campbell C, Hutchinson MF, 2007. Potential impacts of climate change on the distribution of North American trees. BioScience 57:939-948.

McRoberts RE, Holden GR, Nelson MD, Liknes GC, Moser WK, Lister AJ, King SL, LaPoint EB, Coulston JW, Smith WB, Reams GA, 2005. Estimating and circumventing the effects of perturbing and swapping inventory plot locations. Journal of Forestry 103(6):275-279.

Monserud RA, Yang Y, Huang S, Tchebakova N, 2008. Potential change in lodgepole pine site index and distribution under climatic change in Alberta. Canadian Journal of Forest Research 38(2):343-352.

Moscow Forestry Sciences Laboratory, 2013. Custom climate data requests http://forest.moscowfsl.wsu.edu/climate/customData/index.php (accessed July 24, 2013

Nicodemus KK, Malley JD, Strobl C, Ziegler A, 2010. The behaviour of random forest permutation-based variable importance measures under predictor correlation. BMC Bioinformatics 11:110.

Nigh G, 2006. Impact of climate, moisture regime, and nutrient regime on the productivity of Douglas-fir in coastal British Columbia, Canada. Climatic Change 76(3-4):321-337.

Nigh GD, 1995. The geometric mean regression line - a method for developing site index conversion equations for species in mixed stands. Forest Science 41(1):84-98.

Nothdurft A, Wolf T, Ringeler A, Bohner J, Saborowski J, 2012. Spatio-temporal prediction of site index based on forest inventories and climate change scenarios. Forest Ecology and Management 279:97-111.

Packalén P, Temesgen H, Maltamo M, 2012. Variable selection strategies for nearest neighbor imputation methods used in remote sensing based forest inventory. Canadian Journal of Remote Sensing 38(5):557-569.

Prasad AM, Iverson LR, Liaw A, 2006. Newer classification and regression tree techniques: bagging and Random Forests for ecological prediction. Ecosystems 9(2):181-199.

R Core Team, 2012. R: A language and environment for statistical computing Vienna, Austria: R Foundation for Statistical Computing.

Radtke PJ, Robinson AP, 2006. A Bayesian strategy for combining predictions from empirical and process-based models. Ecological Modelling 190(3-4):287-298.

Rehfeldt GE, 2006. A spline model of climate for the Western United States. Fort Collins, CO: USDA Forest Service, Rocky Mountain Research Station. 21.

Rehfeldt GE, Crookston NL, Warwell MV, Evans JS, 2006. Empirical analyses of plant‐climate relationships for the western United States. International Journal of Plant Sciences 167(6):1123-1150.

Reynolds MR, 1984. Estimating the error in model predictions. Forest Science 30(2):454-469.

Rhoton FE, Lindbo DL, 1997. A soil depth approach to soil quality assessment. Journal of Soil and Water Conservation 52(1):66-72.

Running SW, Nemani RR, Heinsch FA, Zhao MS, Reeves M, Hashimoto H, 2004. A continuous satellite-derived measure of global terrestrial primary production. Bioscience 54(6):547-560.

Sabatia CO, Burkhart HE, 2014. Predicting site index of plantation loblolly pine from biophysical variables. Forest Ecology and Management 326:142-156.

Sharma R, Ram PS, Andreas B, Tron E, 2012. Site index prediction from site and climate variables for Norway spruce and Scots pine in Norway. Scandinavian Journal of Forest Research 27(7):619.

Skovsgaard JP, Vanclay JK, 2008. Forest site productivity: a review of the evolution of dendrometric concepts for even-aged stands. Forestry 81(1):13-31.

Strobl C, Boulesteix A-L, Kneib T, Augustin T, Zeileis A, 2008. Conditional variable importance for random forests. Bmc Bioinformatics 9:307.

Strobl C, Boulesteix A-L, Zeileis A, Hothorn T, 2007. Bias in random forest variable importance measures: Illustrations, sources and a solution. Bmc Bioinformatics 8:25.

Strobl C, Hothorn T, Zeileis A, 2009. Party on! The R Journal 1(2):14-17.

USDA Forest Service, 2012. Forest Inventory and Analysis national core field guide Volume 1: field data collection procedures for phase 2 plots, Version 6.0 Arlington, VA: USDA Forest Service. 427.

USDA Natural Resources Conservation Service, 1994. State soil geographic (STATSGO) data base data use information Fort Worth, Texas: U.S. Department of Agriculture, Soil Conservation Service. 113.

USDA Natural Resources Conservation Service, 2013. Geospatial data gateway http://datagateway.nrcs.usda.gov/ (accessed July, 24, 2013).

Uzoh FCC, 2001. A height increment equation for young ponderosa pine plantations using precipitation and soil factors. Forest Ecology and Management 142(1–3):193-203.

Wang HJ, Prisley SP, Radtke PJ, 2011. Errors in terrain-based model predictions caused by altered forest inventory plot locations in the Southern Appalachian Mountains, USA. Mathematical and Computational Forestry & Natural-Resource Sciences 3(2):114-123.

Wang Y, Raulier F, Ung C-H, 2005. Evaluation of spatial predictions of site index obtained by parametric and nonparametric methods—A case study of lodgepole pine productivity. Forest Ecology and Management 214(1–3):201-211.

Weiskittel AR, Crookston NL, Radtke PJ, 2011. Linking climate, gross primary productivity, and site index across forests of the western United States. Canadian Journal of Forest Research 41(8):1710-1721.

Yaussy DA, Iverson LR, Matthews SN, 2013. Competition and climate affects US hardwood-forest tree mortality. Forest Science 59(4):416-430.

Yeh H-Y, Wensel LC, 2000. The relationship between tree diameter growth and climate for coniferous species in northern California. Canadian Journal of Forest Research 30(9):1463-1471.

Table 2.1 Conifer and hardwood species† and their frequencies in SITETREE plots

Conifers	Freq	Hardwoods	Freq	Hardwoods	Freq
Loblolly pine	9127	white oak	4699	shagbark hickory	488
Shortleaf pine	2151	yellow-poplar	3752	scarlet oak	449
E. white pine	1752	quaking aspen	3558	bitternut hickory	371
Balsam fir	1692	northern red oak	3245	northern pin oak	351
Slash pine	1554	red maple	2592	silver maple	307
Red pine	1511	sugar maple	2541	yellow birch	266
Black spruce	1419	white ash	2165	balsam poplar	266
E. tamarack	1194	black oak	1785	cherrybark oak	263
N.white-cedar	884	green ash	1591	black walnut	239
Virginia pine	810	sweetgum	1304	pin oak	220
Jack pine	777	black ash	1239	slippery elm	149
White spruce	764	chestnut oak	1017	shingle oak	142
Longleaf pine	680	paper birch	900	American beech	131
Red spruce	659	Am. basswood	876	black locust	123
Eastern hemlock	439	American elm	860	pignut hickory	79
Eastern redcedar	352	post oak	834	e. cottonwood	55
Scotch pine	126	bigtooth aspen	718	chinkapin oak	49
Pond pine	85	hackberry	586	mockernut hickory	48
Sand pine	64	bur oak	575	black hickory	28
Atl. white-cedar	15	black cherry	561	pecan	12
Larch spp.	8	water oak	505	Shellbark hickory	11
Douglas-fir	1	southern red oak	493	rock elm	6

†Species nomenclature reference at http://apps.fs.fed.us/fiadb-downloads/REF_SPECIES.CSV

Table 2.2 Contemporary climate variables† used as predictors in regression analyses.

Acronym	Definition	Max	Min	Mean	SD
MAT	Mean annual temperature (°C)	24.9	0.5	11.9	5.5
MTCM	Mean temperature in the coldest month	20.4	-18.1	-1.3	8.0
MMIN	Minimum temperature in the coldest month	16.7	-24.4	-7.2	7.8
MTWM	Mean temperature in the warmest month	30.5	13.3	23.6	3.4
MMAX	Maximum temperature in the warmest month	38.4	17.7	30.0	3.3
MAP	Mean annual precipitation (mm)	2216	223	1094	272.7
GSP	Growing season precipitation, April-September	1106	161	604	106.3
TDIFF	Summer-winter temperature differential, (MTWM – MTCM)	37.6	8.2	24.9	5.2
DD5	Degree-days > 5°C	7217	849	3302	1284
DD0	Degree-days < 0°C	1900	0	437.0	521.1
MMINDD0	Minimum degree-days <0°C	2958	0	926.2	819.1
SDAY	Julian date of the last spring freeze	175	0	109.1	28.8
FDAY	Julian date of the first autumn freeze	365	241	292	23.4
FFP	Length of frost-free period (days)	365	63	182	52.9
GSDD5	Degree-days >5°C accumulating within the frost-free period	7217	632	2813	1187
D100	Julian date the sum of degree-days >5°C reaches 100	158	7.0	81.8	39.1
AMI	Annual moisture index, DD5/MAP	22.8	0.5	3.1	1.5
SMI	Summer moisture index, GSDD5/GSP	27.1	0.8	4.7	2.1
SMRPB	Summer precipitation balance	3.5	0.5	1.1	0.2
SMRSPRPB	Summer/Spring precipitation balance	5.5	0.4	1.1	0.4
PRATIO	Ratio of summer precipitation to total precipitation, GSP/MAP	0.8	0.4	0.6	0.1

Note: Interactions used in the analyses are MAP x DD5, MAP x MTCM, GSP x MTCM, GSP x DD5, DD5 x MTCM, MAP x TDIFF, GSP x TDIFF, MTCM/MAP, MTCM/GSP, DD5/GSP, AMI x MTCM, SMI x MTCM, TDIFF/MAP, TDIFF/GSP, PRATIO x MTCM, and PRATIO x DD5.

†Temperature-related variables defined in units of °C and precipitation values in mm.

Table 2.3 GCMs and scenarios used in data downloaded from Moscow FSL, with precipitation and temperature summaries[†] for each scenario

Abbreviation	Storyline	GCM group name	Mean ± SD	
			MAP[†]	MAT[†]
CGCM3_A1B	A1B	Canadian Center for Climate Modeling and Analysis	1158.1±280.3	15.5±5.0
CGCM3_A2	A2	Canadian Center for Climate Modeling and Analysis	1115.4±258.5	17.02±5.25
GFDLCM21_A2	A2	Geophysical Fluid Dynamics Laboratory		
HADCM3_A2	A2	Hadley Center/World Data Center		
CGCM3_B1	B1	Canadian Center for Climate Modeling and Analysis	1162.32±300.83	14.56±5.19
GFDLCM21_B1	B1	Geophysical Fluid Dynamics Laboratory		
HADCM3_B2	B2	Hadley Center/World Data Center	1171.46±296.16	15.38±5.49

[†]MAP (Mean annual precipitation, mm) and MAT (Mean annual temperature, ℃) calculated from the average of one or more GCMs predictions for each climate scenario in the 2090s.

Table 2.4 Soils variables used as predictors in model development.

Abbreviation	Soil Properties	Max	Min	Mean	SD
SBD	Soil bulk density (g/cm^3) - weighted average of components	2.2	0.3	1.6	0.1
SILT	Percent silt (0.002 to 0.05 mm) - weighted average of components	90.0	0.0	32.4	18.5
SAND	Percent sand (0.05mm to 2.0 mm) - weighted average of components	99.0	0.4	46.1	26.4
CLAY	Percent clay (< 0.002 mm) - weighted average of components	80.8	0.0	21.5	13.7
KFFACT	Soil erodibility factor - weighted average of components	0.6	0.0	0.3	0.1
NO10	Percent soil passing #10 sieve (coarse) - weighted average of components	100.0	19.1	85.8	15.2
NO200	Percent soil passing #200 sieve (fine) - weighted average of components	100.0	1.8	51.7	23.9
OM	Organic matter content (% by weight) - weighted average of components	98.0	0.0	1.2	2.0
pH	Soil pH- weighted average of components	9.3	2.8	5.6	0.9
TAWC	Total available water capacity - weighted average of components (0 to 100 cm)	0.5	0.0	0.1	0.0
KSAT	Saturated hydraulic conductivity - weighted average of components (0 to 100 cm)	615.6	0.1	26.6	33.6
SLOPE-dominant	Slope Gradient (%) - dominant component	90.0	0.0	12.0	14.2
SLOPE-weighted	Slope gradient (%) - weighted average of components	90.6	0.0	11.8	14.1
Drclassdcd	Dominant drainage class - weighted average of component	na	na	na	na
Ponding Frequency	Ponding frequency % occurrence	na	na	na	na

Table 2.5. Comparisons of species-specific and species-group (conifers or hardwoods) site index models.

	OOB RMSE (m)							
Species	Species	Group	r	obs.	$\overline{D}$ (m)	$\overline{SI}$ (m)	$\overline{D}$ (%)	ϵ^*(m)
Loblolly pine	4.63	4.77	0.86	8980	0.77	27.6	2.8	0.8
Shortleaf pine	3.31	4.22	0.86	2133	1.56	20.5	7.6	1.66
Eastern white pine	4.78	5.19	0.9	1573	0.96	19.0	5.1	1.04
Balsam fir	3.04	3.3	0.76	1393	0.21	13.9	1.5	0.29
Red pine	3.56	4.37	0.8	1312	1.45	18.6	7.8	1.53
Slash pine	3.62	3.94	0.88	1396	0.99	24.5	4.1	1.07
Black spruce	2.48	3.23	0.77	746	0.61	11.3	5.4	0.73
Tamarack (native)	3.65	3.96	0.83	596	0.48	14.0	3.4	0.62
N. white-cedar	2.4	4.53	0.79	644	2.7	8.9	30.4	2.81
Virginia pine	4.38	4.39	0.8	780	0.21	22.2	1.0	0.33
White oak	3.22	4.16	0.87	4597	1.66	19.1	8.7	1.71
Yellow-poplar	4.48	5.99	0.62	3687	3.31	28.2	11.7	3.37
Quaking aspen	3.09	4.06	0.7	2986	2.1	20.3	10.3	2.15
Northern red oak	3.45	3.66	0.81	3115	0.27	20.2	1.3	0.33
Sugar maple	2.83	3.45	0.79	2414	1.23	26.2	4.7	1.29
Red maple	3.54	3.92	0.76	2409	1.24	18.0	6.9	1.31
Black oak	3.28	3.66	0.86	1754	0.22	20.3	1.1	0.31
White ash	3.65	4.01	0.75	2131	0.93	18.0	5.2	1
Green ash	3.9	4.15	0.9	1463	0.87	21.2	4.1	0.94
Sweetgum	4.36	4.44	0.83	1288	0.78	26.2	3.0	0.87
All species	3.83	4.39	0.9	60955	0.02	20.9	0.1	0.03

$\overline{D}$ = mean difference between species-specific and grouped site index model predictions Eq. [2.1]

$\overline{SI}$ = mean observed site index by species
ϵ^* = minimum negligible detectable difference Eq. [2.2]

Table 2.6. Random Forests regression site index model comparisons.

	Climate & soils		Climate alone		Soils	
	Conifer	Hardwood	Conifer	Hardwood	Conifer	Hardwood
Model #	1	2	3	4	5	6
RMSE (m)	4.5	4.2	4.6	4.3	4.9	4.5
Adj. R^2 (%)	63.2	35.9	63.9	31.9	57.0	26.0

Note: Observed mean site index of conifers and hardwoods in SITETREE table are respectively 22.0 m and 20.7 m

Table 2.7 Random Forests predictor variables and their importance scores (%IncMSE) in models of conifer and hardwood site index (see Table 2.6 for definitions of models 1-4).

Model 1		Model 2		Model 3		Model 4	
variable	%IncMSE	variable	%IncMSE	variable	%IncMSE	variable	%IncMSE
pH	33.6	PRATIO	36.8	SMRSPRPB	29.8	GSPxTDIFF	40.9
NO10	31.2	pH	36.6	PRATIO	28.0	PRATIO	39.9
PRATIOxDD5	29.4	TAWC	35.1	GSPxTDIFF	27.5	MAPxTDIFF	35.2
TAWC	29.3	GSPxTDIFF	34.5	AMI	26.8	SMRSPRPB	34.9

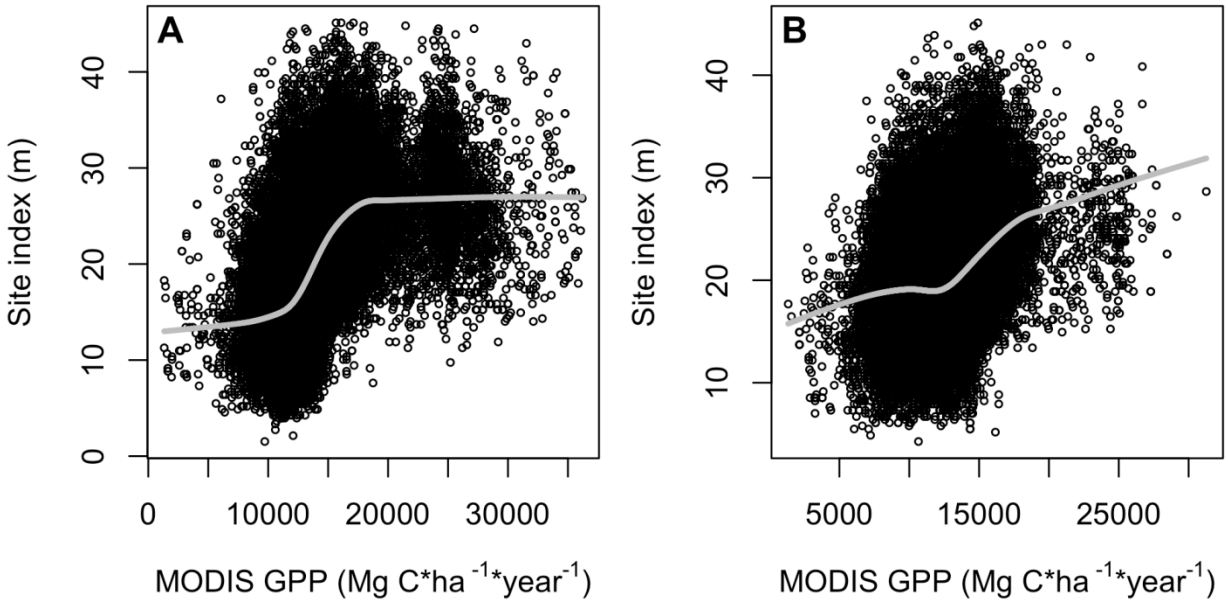


Figure 2.1. Field-plot site index versus MODIS gross primary productivity (GPP) for FIA conifer (A) and hardwood (B) site species (lowess smoother superimposed).

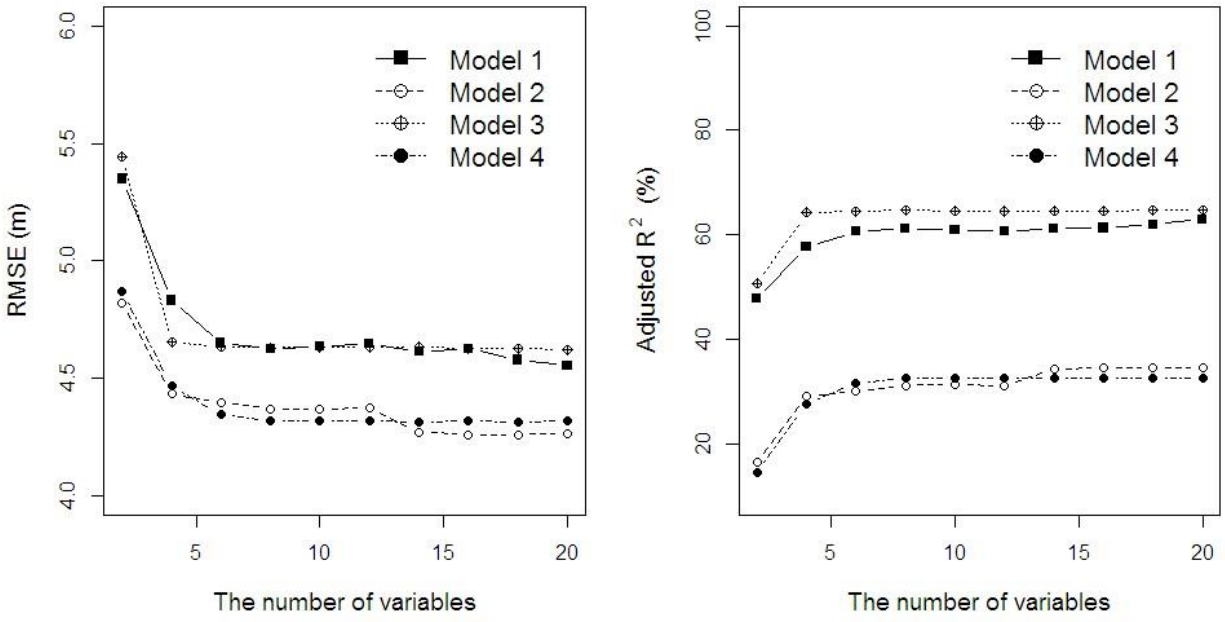


Figure 2.2. Random Forests regression fit statistics (RMSE and adjusted R^2) improvement with additional number of predictors for tested models 1-4 (Table 2.6).

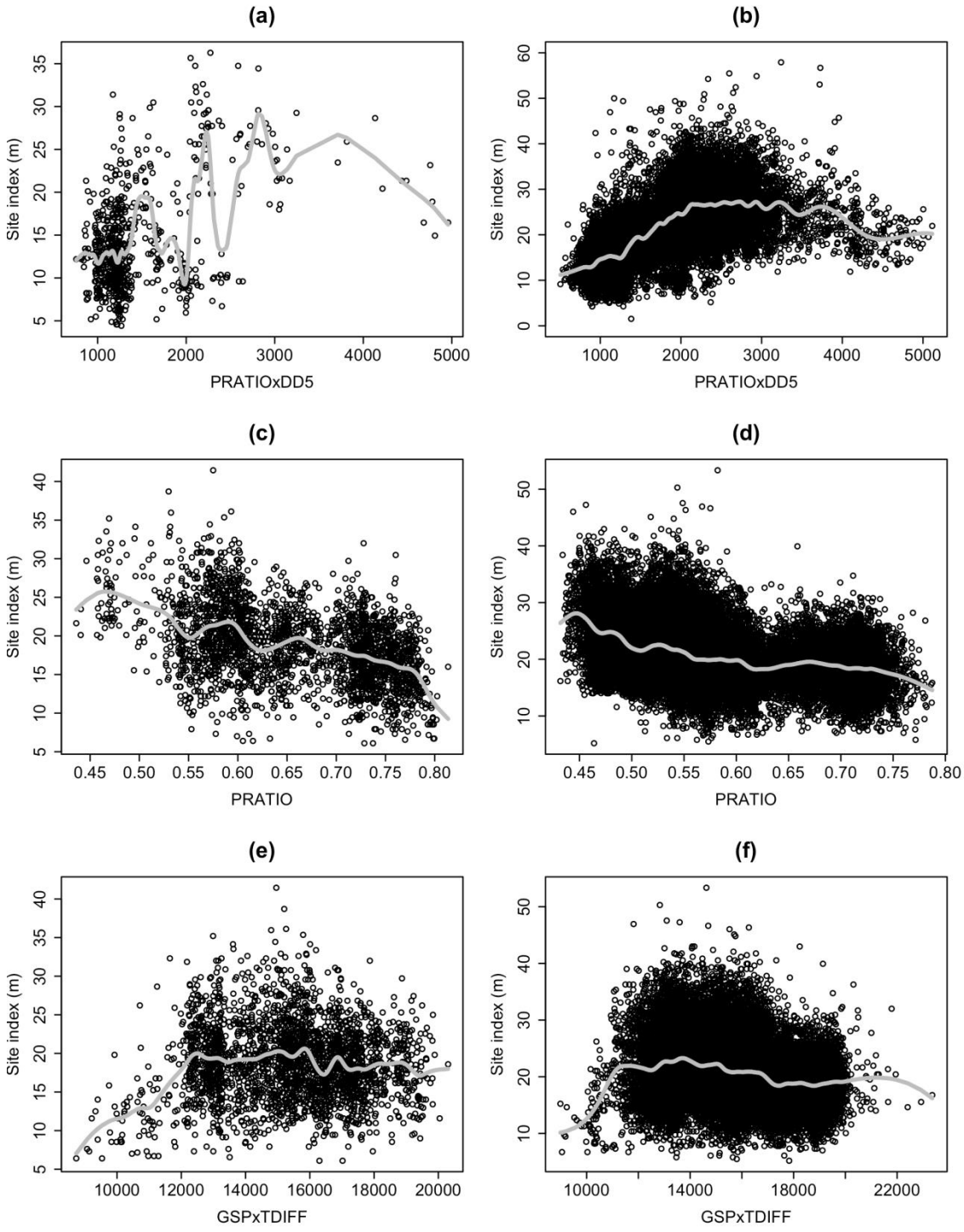


Figure 2.3 The relationship between observed site index and contemporary climate information over eastern US (a-b are for conifers. c-f are for hardwoods. Left picture is for soil pH > 7, and right picture is for soil pH ≤ 7).

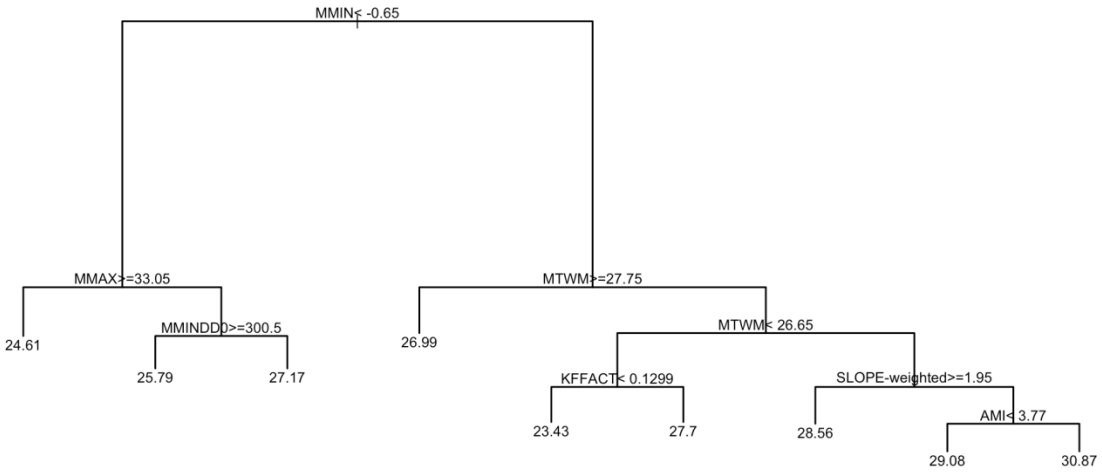


Figure 2.4. Regression tree for FIA loblolly pine (*P. taeda* L.) site index (m) based on climate and soils-related predictors from contemporary conditions.

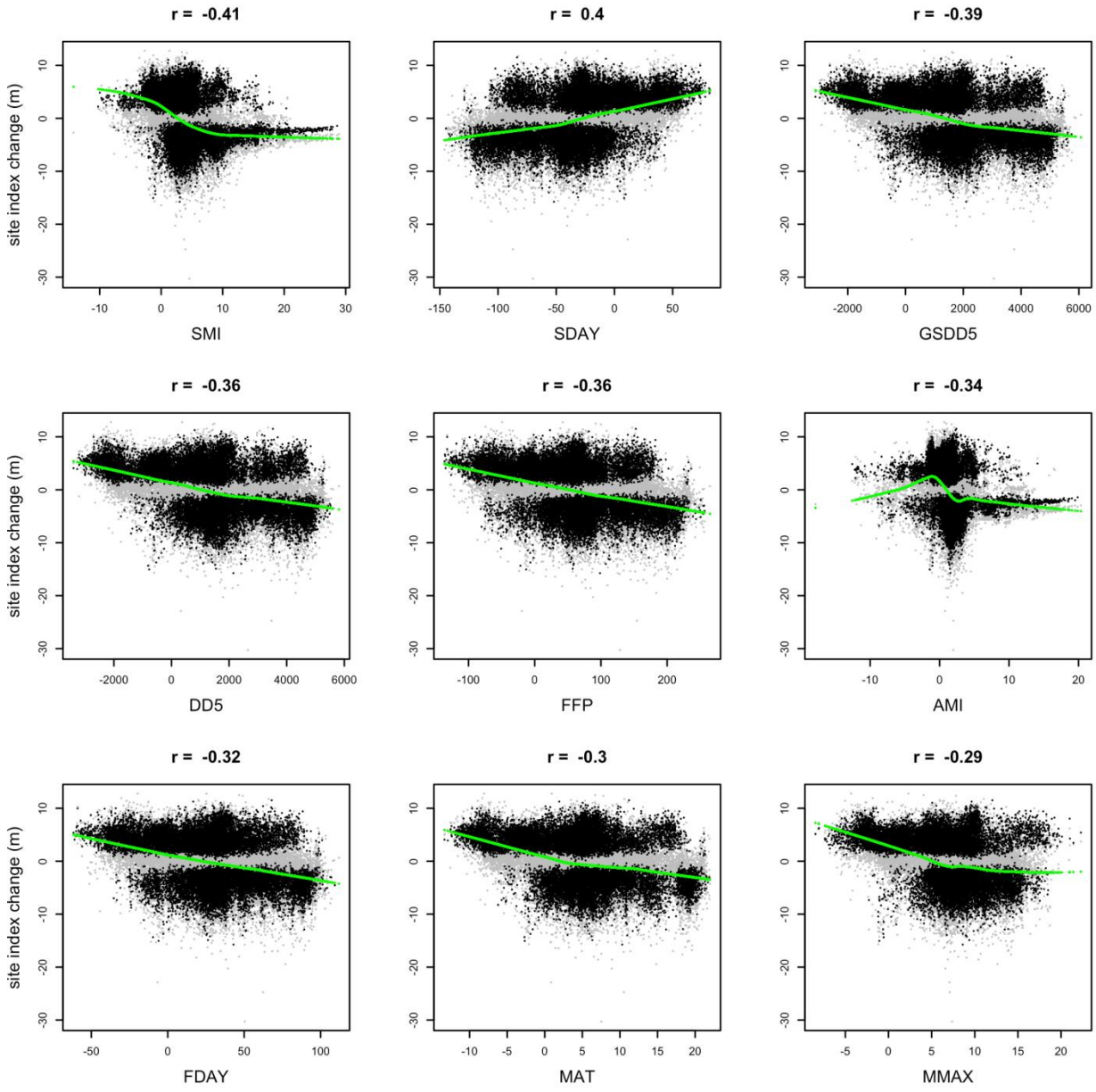


Figure 2.5. Marginal correlation plots for predicted conifer ΔSI versus change in climate variables from 1990 – 2090 for the A2 development scenario. Gray data points show all ΔSI predictions, and black indicates only those significantly different from zero ($\alpha = .05$). Lowess smoothers are superimposed.

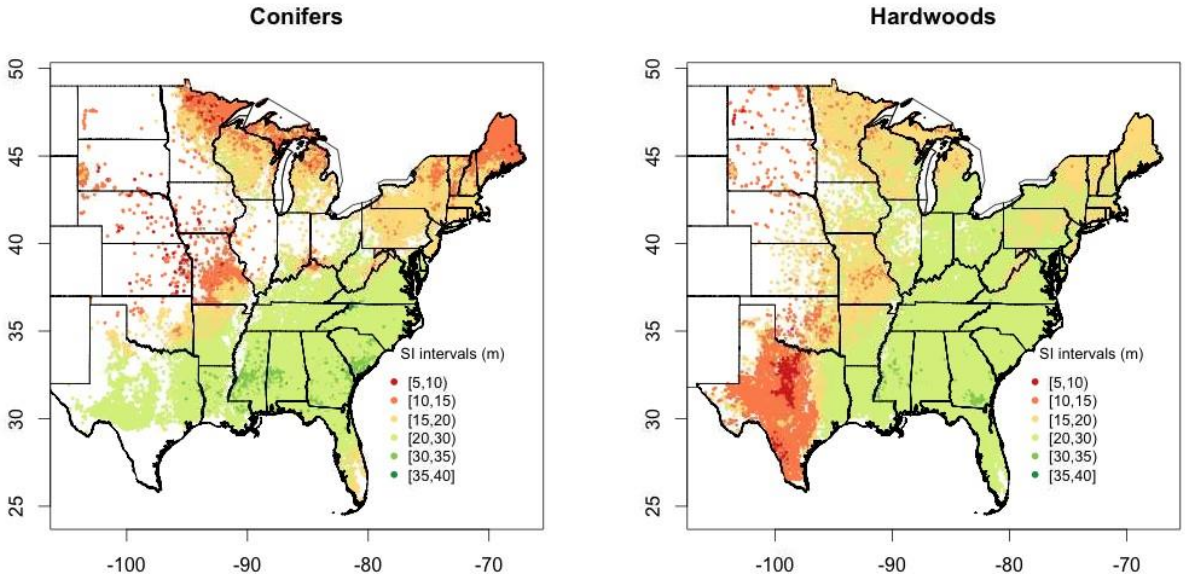
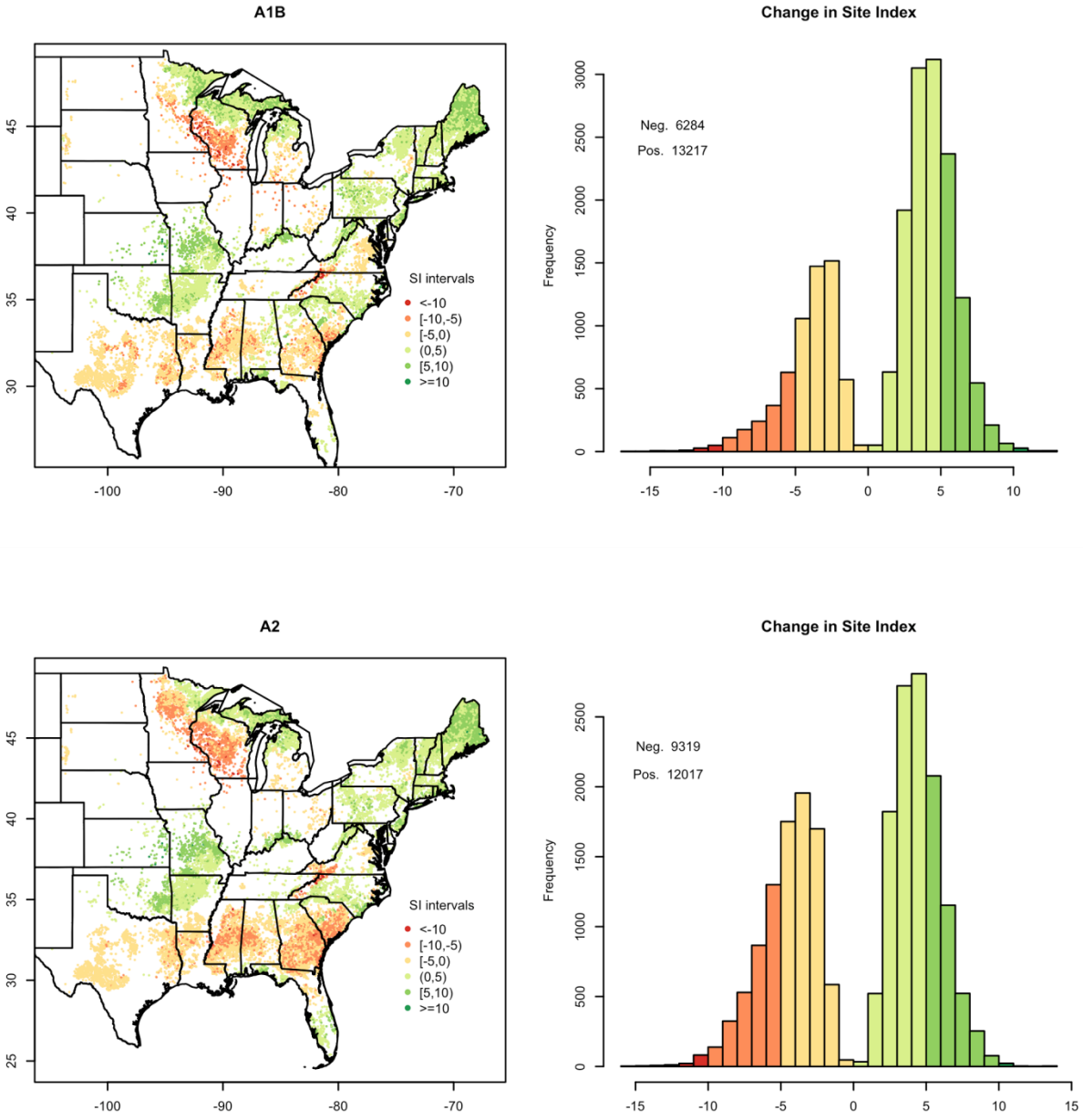


Figure 2.6 Random Forests predictions of site index based on soils and contemporary climate normals (1961 – 1990) for conifers (Table 2.6; model 1) and hardwoods (Table 2.6; model 2) in the eastern United States.



(Figure 2.7, continued on next page)

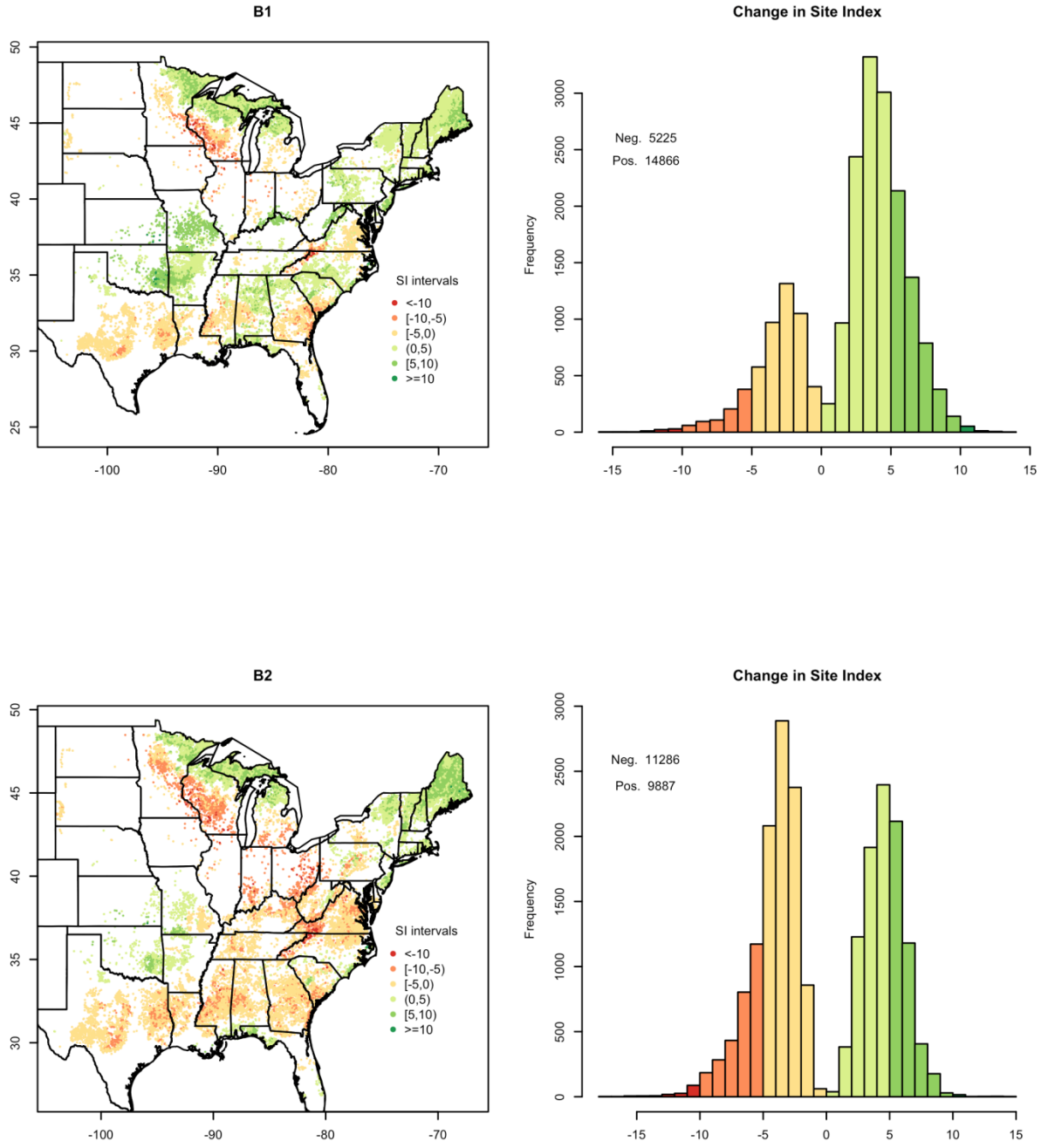
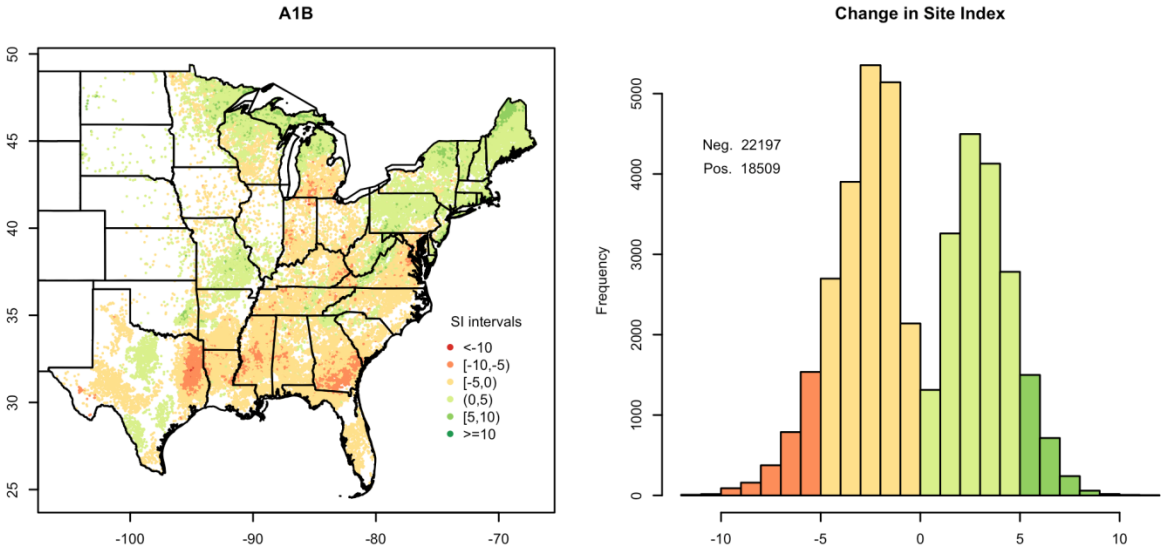
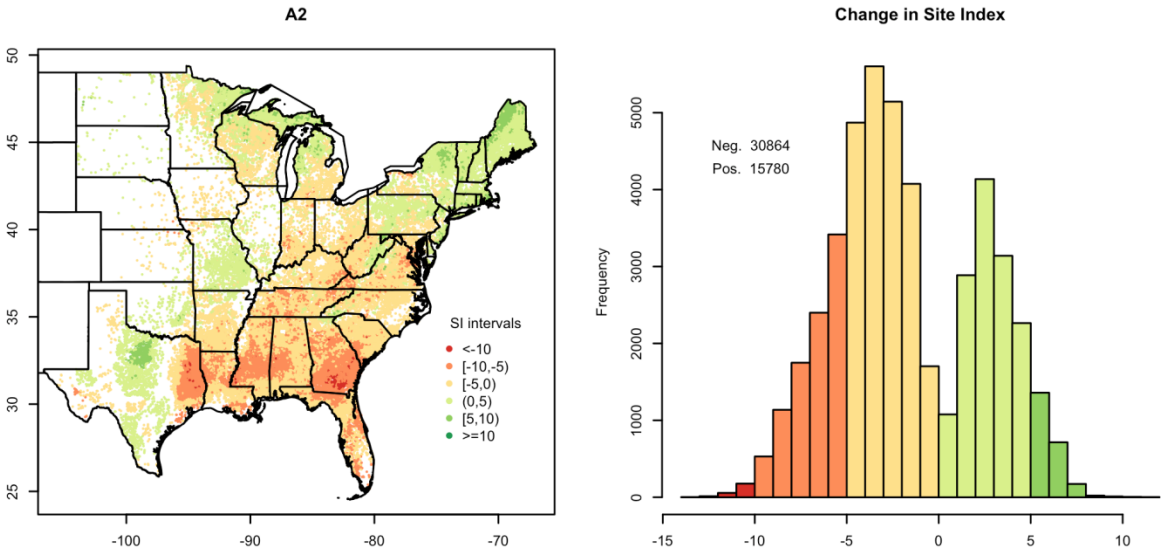


Figure 2.7 Spatial map and histogram of predicted change of site index (m) for conifers in 21st century for four climate change scenarios. Numbers of FIA conifer plots (n = 46,654) having significant ($p < .05$) negative (Neg.) or positive (Pos.) predicted ΔSI are noted.





(Figure 2.8, continued on next page)

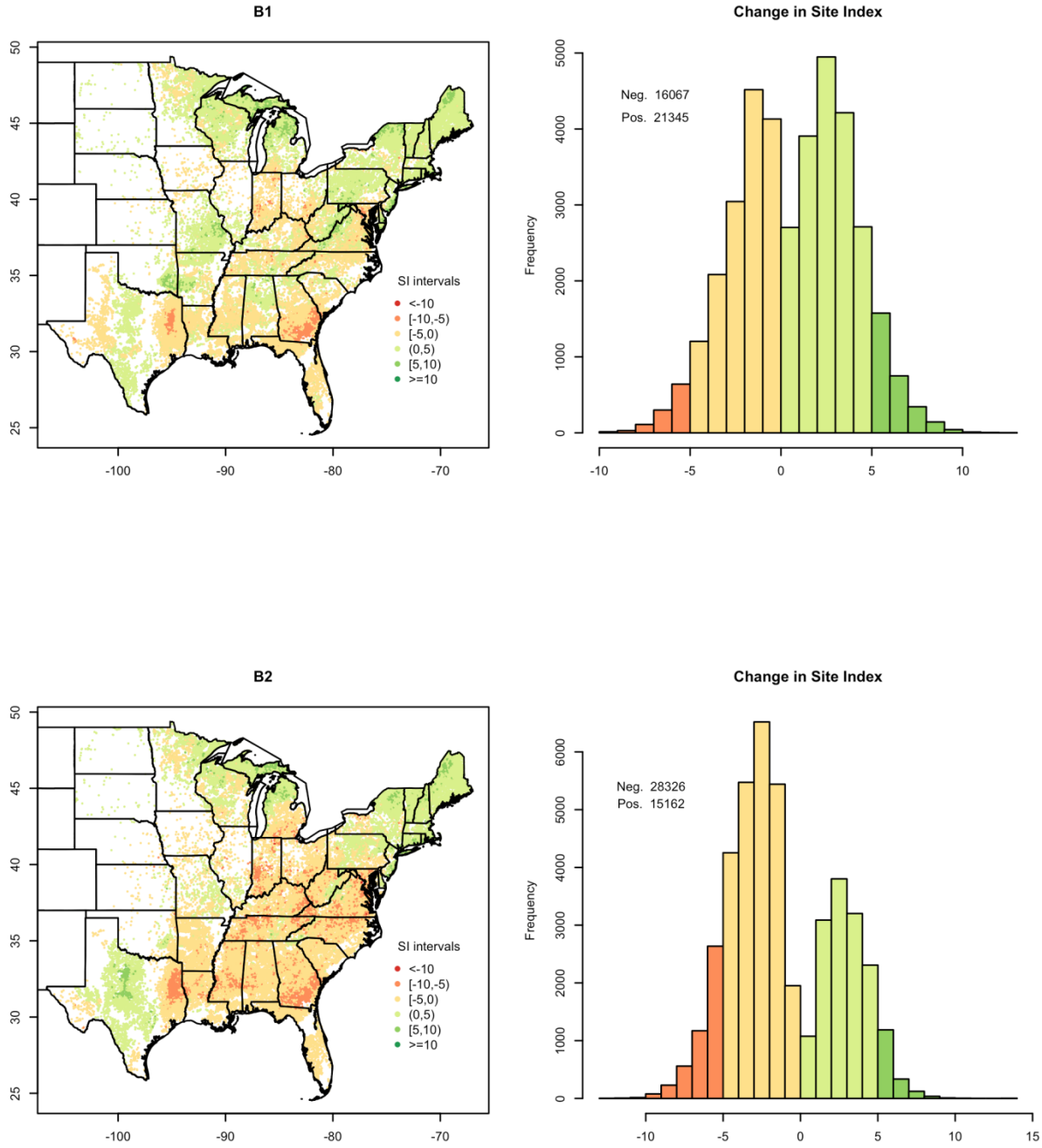


Figure 2.8 Spatial map and histogram of predicted change of site index (m) for hardwoods in 21st century for four climate change scenarios. Numbers of FIA hardwood plots (n = 71,871) having significant ($p < .05$) negative (Neg.) or positive (Pos.) predicted ΔSI are noted.

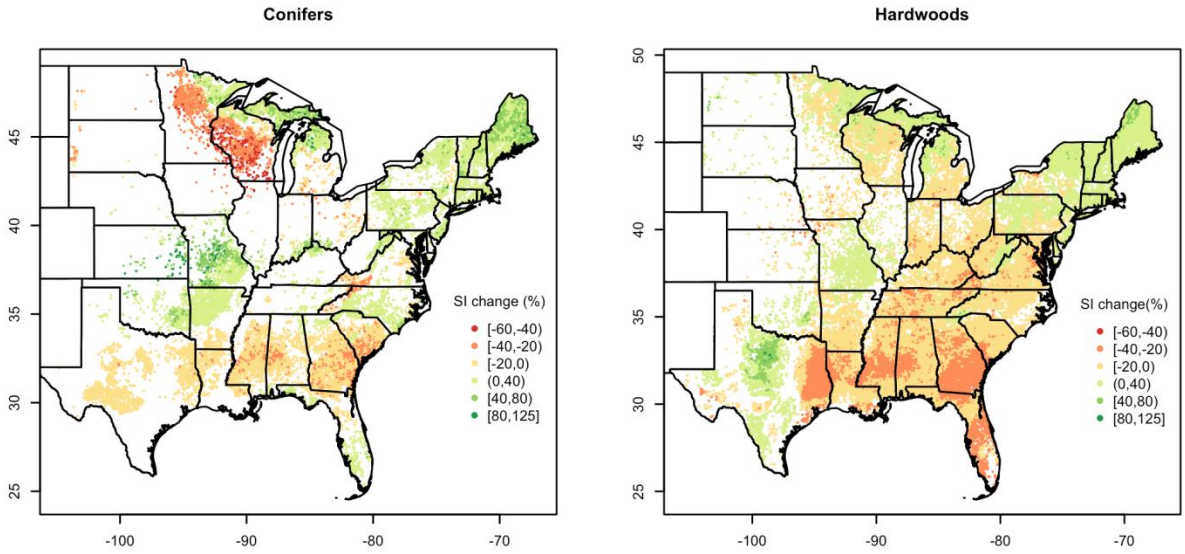


Figure 2.9. Relative change in site index over the 21st Century under A2 climate change scenario.

Chapter 3 Climate sensitive models of individual tree mortality: Comparison of modeling approaches

Abstract

Detecting climate-induced effects in forest ecosystems become increasingly significant as more evidence of greenhouse-gas-related climate change were founded. Predicting climate-induced individual tree mortality is a meaningful but also a difficult task since death is a rare event. Our goal was to develop climate-sensitive mortality models to predict tree's periodic survival probability (PSP) between [0,1] over a measurement interval for 20 most common eastern species using three widely used classification methods, namely, logistic regression model, artificial neural networks, and Random Forests. Specifically, they were presented as logistic regression dealing with equal (LR1) or unequal measurement intervals (LR2) from forest inventory field data, Random Forests model on the basis of the full set of predictor (RFo) or PCA component variables (RF), artificial neural networks (ANN). For each predictive method, individual-tree PSP was predicted using a suite of environmental indicator variables including climate, soils and tree-related information, which were available for most of the FIA plots covering eastern U.S. Average and the standard deviation on MAD and AUC on 40 sets of validation data which were created by bootstrap sampling B = 40 times. Results from each of the predictive methods were compared against actual mortality observations for evaluating the model prediction accuracy showed that RFo had a general better performance than other models with respect of properties of bootstrap MAD and AUC. In addition, RF and RFo models were observed to keep sensitivity and specificity equally higher than other models for most of species at chosen thresholds.

Key words: model performance comparison, logistic regression, artificial neural networks, Random Forests, survival probability prediction

3.1 Introduction

Tree mortality plays a significant role in the functioning of forest ecosystems, directly affecting forest structure, composition, and resource availability at a wide range of spatial and temporal

scales (Franklin et al. 1987). Mortality is related to a host of processes and agents ranging from biotic factors including between-tree competition, herbivory, insects, or diseases, to abiotic factors such as drought, windthrow, fire, or lightning strikes. Often a combination of factors contribute to the death of trees, such as when drought stress and competition act together to make trees susceptible to insect outbreaks (Negron et al. 2009). Mortality is also a key concern in forest management, for example in self-thinning when some trees die – especially those in subordinate canopy positions – freeing up critical resources that are ultimately used by others to survive and grow (Kenkel et al. 1997). Monitoring mortality is important since it may signal a transition in the life history of a particular stand of trees that warrants management intervention (Reynolds and Ford 2005). Mortality affects production-related attributes such as net growth and growing stock (Vanclay and Sands 2009), along with the potential for seedling recruitment, which can be important in uneven-aged or advance-regeneration-based silvicultural systems (Hofmeyer et al. 2010; Steiner et al. 2008). Further, mortality directly influences the health and diversity of organisms that may depend on dead trees for habitat or nutrients (Breda et al. 2006; Rouault et al. 2006).

Because of its direct effect on forest growth and yield, mortality is an essential component in most management-oriented prediction models. Some modeling approaches separate mortality into one of two types, with the first, regular mortality, being defined as that caused by competition or age-related factors. The second, irregular mortality, is defined as that caused by catastrophic events or disturbances such as storms or insect outbreaks (Lee 1971). Many forest models treat the presence of regular mortality as a continual process; while irregular mortality events are either ignored or treated as stochastic events triggered by external factors (Fortin et al. 2008). Variation in regular tree mortality rates is frequently explained using some measures of tree size or age, measures of competition including stand density or individual-tree competition indices, measures of vigor such as crown ratio or tree growth rate, and differential survival rates among species in mixed forest conditions (Fan et al. 2006).

Individual tree mortality models have been developed using a variety of data and approaches. One approach involves the use of cumulative probability distributions such as Weibull, Richards, logistic, gamma, or negative exponential to calculate empirically-derived probabilities of tree

survival or mortality over time (Buford 1983; Somers and Langdon 1980). Logistic regression has been widely used in data sets with equal-length measurement periods (Hamilton 1986; Monserud and Sterba 1999; Radtke et al. 2012; Yang et al. 2003), since it allows for the prediction of a binary response (Hamilton and Edwards 1976). A generalized logistic function has been used where data were collected over unequal measurement intervals (e.g. Crecente-Campo et al. 2010; Monserud 1976; Yao et al. 2001). For repeatedly-measured tree survival data, generalized mixed-effects logistic regression modeling has been used to account for non-independence of residuals (Groom et al. 2012; Kiernan et al. 2009; Yang and Huang 2013). Mixed-effects modeling can also be used to account for spatially correlated errors, such as when data from multiple trees are collected on a single field plot. Survival analysis can be used to study the occurrence of tree death; however, it requires information on individual tree ages and when a tree dies, which are typically not directly observed in forest inventories (Woodall et al. 2005; Yaussy et al. 2013). Nonparametric techniques have been used in mortality prediction including classification and regression tree (CART) analysis (Dobbertin and Biging 1998; Fan et al. 2006) and artificial neural networks (ANN; Guan and Gertner 1991a; Guan and Gertner 1991b; Hasenauer et al. 2001). One advantage of these methods is that they do not require *a priori* knowledge or assumptions about underlying tree mortality processes or functional forms.

Whereas considerable efforts have been conducted to model regular mortality using different approaches, few studies have been conducted to evaluate differences in model performance and suitability. Monserud (1976) compared discriminant analysis, probit analysis, and generalized logistic analysis to find the best strategy to predict tree mortality in northern hardwood stands in Wisconsin. King et al. (2000) compared logistic regression, ANN and a support vector machine for classifying individual tree mortality in West Virginia. Hasenauer et al. (2001) evaluated different types of neural networks and compared them to logistic regression for predicting tree mortality in Norway spruce. Kiernan et al. (2009) showed that logistic regression modeling using generalized estimating equations (GEE) for predicting tree mortality over time in uneven-aged hardwood stands gave improved results compared to logistic regression models fitted by generalized linear modeling. The primary advantage of model fitting with GEE is the ability to account for correlated observations even though the nature of such correlations may be unknown (Liang and Zeger 1986).

ANN and Random Forests (RF) are two widely-used nonparametric methods, which have been developed for classification purpose due to several noted advantages including ability to detect nonlinear relationship between predictor and response variables, the ability to reproduce interactions between predictors variables, and the flexibility of being unconstrained by a predefined functional relationship (Cutler et al. 2007; Tu 1996). The simplest ANN model formulation, known as a McCulloch-Pitts model (McCulloch and Pitts 1943), consists of a single neuron, that simply accepts a set of input values, assigns weights to each value and sums them, then transforms the summed value by a activation function into an output value. Typical forms used in activation functions include linear, threshold, threshold linear, and sigmoidal. The sigmoid activation function is by far the most frequently used in neural networks due to its desired asymptotic properties (Bishop 1994). It allows prediction values to be continuous ranging from 0 to 1, which is useful for binary classification problems. There are also many situations where the neural networks have at least one hidden layer with multiple neurons and an output layer with multiple classes. An example of a well-explained mechanism of neural networks with a more complex structure including multiple layers and neurons is given by Guan and Gertner (1991a). The nature of neural network model in modeling interactions and nonlinearities implicitly lead to overfitting a training data set (Tu 1996). A recommended method to avoid overfitting involves using a large enough network to avoid underfitting, then limiting the number of iterations of the fitting procedure by cross-validation or bootstrapping, such as using weight decay to control overfitting.(Moisen and Frescino 2002). Random Forests(Breiman 2001) are classification-and-regression-tree-based ensemble machine learning tools capable of modeling complex relationships between many variables in large data sets. The predictors may consist of either continuous or categorical variables, which is also true of the response variables. One factor affecting internal estimates of out-of-bag error is the number of trees in the ensemble environment (Breiman 2001). As the size of collections of trees increases, out-of-bag error decreases and tends to stabilize at some point. The other factor influencing out-of-bag error rate is the number of variables selected to split each node (m_{try}). The default value for m_{try} is $\sqrt{p}$ for classification problem, however, in practice, the best value for this parameters will depend on the dataset, especially in the presence of a large number of noise predictors, where randomly selected $\sqrt{p}$ predictors may be non-informative (Hastie et al. 2009).

Strong correlation among predictors in the linear regression will inflate the standard errors of the estimates of the model coefficients and may lead to unreliable results (Hosmer and Lemeshow 1989). Principal component analysis (PCA), as a common variable reduction procedure, transforms correlated variables into a set of linearly uncorrelated components using an orthogonal transformation, thereby reducing predictor multicollinearity (Kuhn and Johnson 2013). PCA has been used to reduce the dimensionality of predictors, including measures of soil properties (de Toledo et al. 2012) and climate (Cailleret et al. 2014; Ruiz-Benito et al. 2013). Because PCA relies on an assumption of linearity in relationships between predictors, its effectiveness may be limited when nonlinear patterns or complex interactions among variables exist (Chen et al. 2015; Muñoz and Felicísimo 2004; Quinn and Keough 2002).

The goal of this research was to compare the performance of three approaches for modeling tree mortality in twenty forest tree species from the eastern United States, namely logistic regression (LR), ANN, and RF. Models were fitted to observations of individual-tree survival and mortality made over unequal measurement intervals from forest inventory field data. Auxiliary records of climate-related variables and soil properties paired with the field locations were also included as predictors. To account for the relatively large number (> 40) of possible predictor variables where correlated predictors might adversely affect regression results, PCA was employed; however, a baseline Random Forests model (RFo) was trained to the full set of predictors without any variable reduction using PCA to test a known advantage of the Random Forests method when a large number of possibly correlated predictors is used for predictive modeling. In pursuing these goals, our aim was to compare how well the different techniques performed at predicting mortality across an extensive geographic area, and to identify any notable strengths or weaknesses of the methods tested.

3.2 Materials and Methods

3.2.1 Study area

The study area was defined to span the geographic ranges of most forest tree species growing in the eastern U.S., which we defined as those states east of the Great Plains physiographic province (Fenneman 1946). Data were compiled from 37 states that included areas east of 100°

W longitude, comprising states of North and South Dakota, Nebraska, Kansas, Oklahoma, Texas, and all states further east.

3.2.2 Vegetation data

Vegetation data were compiled through the publicly-available online database of the USDA Forest Service, Forest Inventory and Analysis (FIA) national program[1]. Individual trees observed on 7.3 m (24 foot) radius subplots that were alive at the beginning of a growth interval and either dead or alive at the end of the interval were included. Only trees having diameters at breast height (1.37 m; DBH) ≥ 13.0cm were used in the analysis. Trees were omitted from any plots where human-caused damage or geologic disturbances, e.g., landslides, were recorded. The remeasurement interval varied from plot-to-plot with a range from 0.2 to 16.4 years between inventories; however, seventy percent of tree records had measurement intervals between 4 and 6 years. For each plot, several tree and stand level statistics measured or calculated at the beginning of the growth interval were used, including number of trees per hectare (N, trees ha^{-1}); stand basal area (BA, m^2 ha^{-1}); Tree diameter at breast height, i.e. at 1.37m, (DBH, cm) and its inversed and squared forms DBH^{-1} and DBH^2; quadratic mean diameter (QMD, cm); RDBH, ratio of individual tree DBH to plot QMD; and the percentage of stand basal area by species component (BASC, %). Other factors deemed as potentially useful predictors of mortality included the measurement interval length (L, years) along with a suite of climate and soils variables.

The twenty species occurring most frequently in the FIA database were selected for mortality model development (Table 3.3). In addition to ensuring very large sample size, the twenty species' nominal ranges covered virtually all of the Eastern U.S. Five species exhibited mortality rates below 1.5%, which was considered to be low relative to the other species studied. Four exhibited mortality of 5% or greater, which was considered to be relatively among the species studied. We considered the other eleven species as exhibiting relatively moderate rates of mortality among the twenty species studied.

[1] Available online at http://apps.fs.fed.us/fiadb-downloads/datamart.html (last accessed on 22 March, 2015)

3.2.3 Climate and soils data

Contemporary climate data were comprised of climate station records spanning three decades (1961 to 1990), chosen to overlap with the development of many present-day intermediate-aged and mature forests in the eastern U.S. (Crookston 2012). A suite of climate variables (Table 3.1) was compiled for contemporary based on averaged monthly values for maximum, mean, and minimum daily temperatures, monthly total precipitation, and several derived annual climate variables potentially related to tree growth, such as growing season precipitation (Rehfeldt et al. 2006).

Soil types and their associated attributes were compiled from the USDA Soil Survey Geographic (SSURGO) online database, linked to FIA field plots based on plot geographic coordinates. Available soil survey data for the eastern U.S. were downloaded from the USDA NRCS Geospatial Data Gateway server (USDA Natural Resources Conservation Service 2013). At the time of downloading, soils data were unavailable for some parts of several states in the eastern U.S., most notably Alabama, Georgia, Maine, Minnesota, Mississippi, New Hampshire, and New York. Most void areas were small – the size of one or two counties – except for a larger area in northwest Maine and several relatively large tracts of public lands in northern Minnesota. Various soil properties related to the growth or survival of forest tree species based on the findings of Iverson et al. (2008) were included as predictors in model development (Table 2.4).

3.3 Modeling and Analysis

To avoid severe multicollinearity problems involving the 20 climate and 11 soils variables, prior to development and comparison of LR, ANN, and RF mortality models, PCA was applied to project the predictor variables into new linearly uncorrelated components (Quinn and Keough 2002). According to the tests from 20 most common species, the first five components were retained as new predictors for later analysis, because those components were able to account for at least 80 percent of the total variation in 20 climate and 11 soil variables. Those five components, PC_1 - PC_5, along with tree & stand-related variables were then used as predictors in the mortality models developed for comparison of LR, ANN, and RF. Since logistic regression model is sensitive to multicollinearity problem, the variables with variance inflation factor larger

than 10 were removed from the tree & stand-related variables as well as PC1 - PC5, and such procedure was conducted species by species. The remaining variables (Table 3.4) used in LR1 and LR2 were also applied into ANN and RF models for a fair model comparison. Results from each of the predictive methods were compared against actual mortality observations for evaluating the model prediction accuracy, where the MAD, receiver operating characteristic (ROC) plots, and associated area under the ROC curve (AUC) were conducted as the criteria for model comparison.

Random Forests models are generally unaffected by correlated predictors in the same way parametric models are due to the use of bootstrap sampling of both predictors and observations used in training and validating models. They are also robust to overfitting by nature of how predictor subsets are selected by bootstrap sampling, even in extreme situations where the number of predictors exceeds the number of observations. Moreover, computational problems are seldom encountered when training RF models involving a large number of predictors and many observations due to the simplicity of the training algorithm. To realize these potential advantages, a baseline Random Forests model (RFo) was trained using all the original variables in the training data without any variable selection or dimensionality reduction using PCA.

Predictions of individual tree survival (1) or mortality (0) were compared to observed survival and mortality to evaluate model accuracy, where mean absolute deviation (MAD) as well as threshold-independent ROC plots and associated Area Under the Curve (AUC) (Fawcett 2006) were used as evaluation criteria. Mean absolute deviation is a measure of model accuracy expressed as MAD = $\frac{\sum_{i=1}^{n_{OOB}} |p_i - \hat{p}_i|}{n_{OOB}}$, *w*here n_{OOB} is the number of out-of-bag (OOB) or validation observations generated from a bootstrap sample, p_i is the the i^{th} OOB observed value, and $\hat{p}_i$ is the model-predicted value corresponding to p_i. In the case of individual-tree survival, observed or predicted values for p_i are binary (0, 1). The ROC curve plots the true positive rate versus false positive rate for a model of interest over the range of threshold values, usually 0 to 1 for models that predict the probability of a binary outcome. The ROC curve can be used to find an optimal prediction threshold based on relative costs of false-positive or false-negative predictions and the relative proportions of observed zeros and ones in the training data (Fawcett 2006; Streiner and Cairney 2007). For comparing different models and modeling approaches AUC was

used because of its simple interpretation and the ability to quantify overall model performance in a single scalar value regardless of the choice of prediction threshold (Marrocco et al. 2008). A higher AUC score corresponds to a better-performing model. In contrast, lower MAD only compares model performance based on a specific choice of threshold or thresholds without distinguishing between false positive or negative errors.

3.3.1 Model specification

Three widely used binary classification methods, LR, ANN, and RF were chosen as candidate approaches for developing models to predict tree status (live or dead) over a measurement time interval. Given that L was not identical on all field plots in the FIA data, two variants of LR were used, one (LR1) that did not explicitly include L in the model formulation and one (LR2) that did.

Logistic model and the general logistic model

Denoting k as the number of predictor variables measured at the beginning of a measurement period and P as the probability a tree survives during the measurement period, the following logistic model (LR1) was proposed and fitted using the "glm" function in the R package "stats" (R Core Team 2012).

$$P = \frac{1}{1 + e^{-(\beta_0+\beta_1 x_1+\cdots+\beta_k x_k)}} \quad [3.1]$$

Monserud (1976) proposed a generalized logistic model, similar to [3.1] but with L included as a predictor, thus allowing for predicting mortality over a variable time interval rather than adopting a single fixed interval. The model, denoted as LR2, was fitted using the "mle2" function in the R package named "Tools for general maximum likelihood estimation (bbmle)" (R Core Team 2012).

$$P = \left(\frac{1}{1 + e^{-(\beta_0+\beta_1 x_1+\cdots+\beta_k x_k)}}\right)^L \quad [3.2]$$

A quasi-Newton method, BFGS, was chosen as a preferred optimization method since it generally performed best among available optimization methods in "bbmle" package (results not shown).

Neural networks

Neural networks were built and trained with "nnet" package in R (Venables and Ripley 2002). The "nnet" function implements neural networks with multiple neurons in a single hidden layer, and the output employs a sigmoid function for activation function to guarantee output values ranging from 0 to 1. The use of a zero weight decay value results in overfitted networks; in contrast, an infinitely large weight decay results in underfitted networks (Hastie et al. 2009). To investigate the optimal weight decay, experiments were performed by altering the value of weight decay. A simple data splitting was applied to reduce computation time, even though it was not as robust as k-fold cross validation. Data set for each species was divided into two parts: two thirds were placed in a training set and one third in a validation set.

The optimal number of hidden neurons was also examined by varying the number from 1 to 20 while checking the resulting model validation accuracy. A goal in tuning or adjusting the number of hidden neurons was to select a number large enough to ensure that the functional forms of relationships between inputs and outputs are sufficiently flexible. In addition, hidden nodes ensured that any interactions between input variables that affected the model output are accounted for. Once an adequate number of hidden neurons were included, adding more leads to no additional gain in accuracy. Further, too many hidden neurons can result in a model that overfits the training data.

Data were normalized before being used in the algorithm. Normalization involved standardizing each continuous columns through dividing the centered column by its own standard deviation, which standardized all values to have mean zero and standard deviation one (Terrin et al. 2003). Five neural networks were created by applying a set of weights (0.005, 0.01, 0.03, 0.05 and 0.1) in the training data species by species. MAD and AUC as two model criteria were calculated on the base of test data for each neural network to find an "optimal" weight.

Random forests

Random Forests models were trained using the "randomForest" package in R. In order to reduce computation time, data set for each species was simply split into two parts: two thirds as training set and one third as test set. To find an "optimal" m_{try} for each species, nine Random Forests models were developed at all possible values of m_{try} on the base of training set, which was ranged from 2 to 10. The performance of those models was evaluated in validation data by the model criteria MAD and AUC.

3.3.2 Assessing performance

Bootstrap cross-validation was used to estimate model accuracy, due to its lower variance property than k-fold cross-validation or leave-one-out techniques (Efron 1982; Lendasse et al. 2005). The number of repetitions needed to obtain a reliable estimate was detected as follow. For each bootstrap trail, a sample of size n (Table 3.3) was drawn with replacement to represent a training set for constructing LR1, LR2, ANN, RF and RFo models. The remaining observations were used for the validation of the models. Several trials showed that $\overline{MAD}$ and $\overline{AUC}$ tended to be stable when the number of bootstrap sample replications (B) was set to 40 or more; therefore B = 40 was used in subsequent analyses.

3.3.3 Model performance test for each species

In order to compare model accuracy statistics using paired sample data, a single collection of B = 40 bootstrap samples was generated. Each model was fitted to the same bootstrap samples to achieve greater power in testing for differences in model performance using pairwise comparisons (Kohavi 1995; Sakai et al. 2007). Model performances for each species were ranked based on MAD and AUC values averaged over B = 40 bootstrap samples of LR1, LR2, ANN, RF and RFo models. A model with a lower value of $\overline{MAD}$ or a higher value of $\overline{AUC}$ was deemed to have a better performance, and was ranked accordingly. Because the paired sample size (B) was larger than 30, pairwise comparisons based on Student's paired t tests with Bonferroni corrections for multiple testing were used to check if the performance of any two models differed significantly from each other over repeated bootstrap sampling. When no

evidence was observed to note that one model outperformed another, the two models were assigned the same rank.

For each bootstrap sample, MAD and AUC were calculated using the bootstrap out-of-bag observations compared to predicted values form each of the different fitted models. Ten possible combinations were available for calculating pairwise differences. One pair of model accuracy from one bootstrap was calculated as below

$$\begin{aligned} \text{MADD}_{ij} &= \text{MAD}_i - \text{MAD}_j; \\ \text{AUCD}_{ij} &= \text{AUC}_i - \text{AUC}_j, \; i \neq j \in \{\text{LR1}, \text{LR2}, \text{ANN}, \text{RF}, \text{RFo}\} \end{aligned} \quad [3.3]$$

$\overline{MADD}_{ij}$ and $\overline{AUCD}_{ij}$ are respectively mean value of 40 pairs of MADD_{ij} and AUCD_{ij} [3.3]. We set H_0: the mean difference between paired MAD, $\mu_{MADDij} = \mu_{MADi} - \mu_{MADj} = 0$ and H_0: the mean difference between paired AUC, $\mu_{AUCDij} = \mu_{AUCi} - \mu_{AUCj} = 0$ separately with significance level of 5%. The performance of model *i* was deemed to be significantly better or worse than model *j* only when a null hypothesis of "no difference" as to AUC or MAD was rejected ($\alpha = .05$). In such a case the relative rankings of models *i* and *j* based on $\overline{MADD}_{ij}$ criterion were adjusted accordingly. The same procedure was used to rank differences in models using the $\overline{AUCD}_{ij}$ criterion." Then LR1, LR2, ANN, RF, and RFo were ranked as to the $\overline{MAD}$ and $\overline{AUC}$ separately. The above procedure was repeated for all 20 species.

3.3.4 Model performance test regardless of species

Pairwise comparison were also made to find out the general model performance among LR1, LR2, ANN, RF and RFo models regardless of species. Since the above model performance comparison were made for each of 20 species, 800 pairwise differences MADD_{ij} and AUCD_{ij} were calculated for each pair of models. We denoted the overall mean of 800 MADD_{ij} pairwise differences as $\overline{\overline{MADD}}_{ij}$.Correspondingly, $\overline{\overline{AUCD_{ij}}}$ denoted the overall mean of 800 pairwise AUCD_{ij} differences over 20 species used to rank model performance. Null hypotheses comparing model performance regardless of species were the same as the species-specific null hypotheses but using overall mean $\overline{\overline{MAD}}$ and $\overline{\overline{AUC}}$ from 800 samples to rank model

performance($\alpha = 0.05$). Finally, LR1, LR2, ANN, RF and RFo were ranked according to the $\overline{MAD}$ and $\overline{AUC}$ separately.

3.3.5 Model comparison on specificity and sensitivity

Specificity (true negative rate) and sensitivity (true positive rate) were also examined as measures of classification performance. For binary classifiers based on a continuous response on [0, 1], new observations were assigned predicted classifications of live or dead if the model response either exceeded or fell below a specified threshold τ. Using bootstrap sample OOB observations of trees' survival status, specificity and sensitivity were computed for each OOB observation. As the threshold was lowered, sensitivity necessarily increased along with a corresponding decrease in model specificity. Based on the assumption that the cost of making a false positive prediction was the same as that for making a false negative prediction, an optimal threshold was chosen by finding the threshold value where the cost of the sensitivity and the specificity of model predictions on testing data were equal. Two curves were plotted for describing how sensitivity and specificity changed with the increasing value of threshold. The intersection point was set as the preferred threshold that balanced the two accuracy measures equally.

3.3.6 Periodic survival probability prediction under climate change

The RFo models were applied to the climate predictions for future climate scenarios at eastern sites to demonstrate the possible changes in PSP over space and time. RFo was developed based on all indicators including stand & tree variables (BA, N, BASC, QMD, L, IDBH and DBH2), contemporary climate (Table 3.1) and soil (Table 2.4) variables. Contemporary predictions of tree PSP for each FIA plot were obtained using the RF out-of-bag prediction function. Future site index predictions were obtained using RF models with future climate conditions in 2090 specified as inputs. Soil properties, as defined by the SSURGO data, were assumed to remain unchanged at a particular plot location from contemporary to future conditions, so SSURGO data were used as inputs without alteration between contemporary and future scenarios. For each IPCC development scenario, projections based on inputs from one or more GCMs were averaged (Table 2.3 GCMs and scenarios used in data downloaded from Moscow FSL, with precipitation

and temperature summaries[†] for each scenario to account for differences between different GCM predictions run under the same scenario. To simplify reporting, only results from A2 climate scenario are presented here. Two native species, red maple and eastern white pine were chosen as example species for further investigation. On each plot, the predicted PSPs of each tree under contemporary and future climates were averaged to reflect the overall level of the species' mortality on the plot area. The prediction difference per tree or plot between future and contemporary period served as a measure of how future climate change may influence mortality for these species.

The PSP of trees decreased across large geographic areas over time due, with changing precipitation and temperature being likely drivers. The information collected from the plots having the greatest average differences in PSP over the century was chosen to show the relationship between tree diameter and decreasing survival. For eastern white pine, plots with 20 or more trees and mean decrease in PSP $\leq$ -0.1 were selected. For red maple, the plots with 20 or more trees and mean decrease in PSP $\leq$ -0.3 were selected. Variable importance scores were ranked according to permutation accuracy and node impurity measures to assess which predictors were most meaningful in predicting climate-drive mortality. Within variable importance plots, higher positions indicate greater importance of predictor variables.

3.4 Results

3.4.1 Parameter tuning

For most species, MAD and AUC in the test data both increased as weight values were increased from 0.005 to 0.1. A compromise value of 0.03 was selected for ANN weights so as to obtain a balance between low values of MAD and high values AUC (Figure 3.1). Based on fitting the ANN with a series of increasing numbers of hidden neurons, twenty hidden neurons were used to ensure maximum prediction accuracy in the ANN. The number of classification trees for RF models was set to 500, since only a slight improvement could be achieved by adding more than that. Given 500 trees, a lower value of m_{try} generally corresponded to a lower value of MAD leading to a better model performance (Figure 3.1). Unlike MAD, the value of AUC did not vary appreciably over the values of m_{try} tested; however, the smallest values of m_{try} showed a slightly

lower model AUC. Thus, setting m_{try} to the default setting (m_{try} = 3) gave near optimum results for most species based on balancing values of MAD and AUC. The default value of m_{try} was also found to be suitable for use in training the RFo model.

3.4.2 Model accuracy and stability

Over all models tested, AUC and MAD averaged over multiple bootstrap samples of the test data varied inversely, i.e. higher values of AUC corresponded to lower values of MAD, both of which indicated better model performance (ρ = -0.73). All five models that were tested performed better in species with low observed mortality than those with high mortality (Figure 3.2). The five species having the lowest observed mortality – ACSA, PIST, TSCA, FRAM, and FAGR, all of which had mortality lower than about 1.5% – also had among the lowest values of $\overline{MAD}$ and highest values of $\overline{AUC}$. Poorer model performance corresponding to relatively high values of $\overline{MAD}$ was seen in species having higher mortality, including PINU, LIST, PIEC and QUNI (Figure 3.2; left panel). Among those species PINU, LIST and QUNI also had relatively low values of $\overline{AUC}$ (Figure 3.2; right panel).

Lower values of the standard deviation of MAD and AUC computed across 40 bootstrap samples indicated which models exhibited greater stability. The LR models were most stable, having consistently lowest values of standard deviation of MAD and AUC for all 20 species (Figure 3.3). The standard deviations of MAD and AUC for RF and RFo models were generally as low as LR for all 20 species, except for several noticeably high values observed in standard deviation of AUC for species with the lowest observed mortality. ANN exhibited the least stability of the models tested as seen from its comparatively high standard deviations for MAD and AUC across all 20 species.

3.4.3 Model performance test for each species

The $\overline{MAD}$ statistic showed that ANN models performed best in 19 of the 20 species examined; however, ANN models never ranked higher than third best as measured by $\overline{AUC}$ (Table 3.5). Both LR models performed poorly compared to ANN or RF models as measured by $\overline{MAD}$. Random Forests models consistently ranked either first or second highest in terms of $\overline{AUC}$,

except for when used to model species having observed mortality below about 1.5% or 2%. In species with the lowest observed mortalities, even though rankings between modeling approaches were deemed to be statistically significant, the magnitudes of differences between best and worst ranked models were relatively small. Between the two LR models developed, the generalized LR2 model was ranked as high as or above LR1 based on $\overline{AUC}$ for 17 of 20 species.

3.4.4 Model performance test regardless of species

Besides investigating model performance for each species, the overall performances for 20 species were studied as well (Table 3.6). A negative $\overline{\overline{MADD}}_{ij}$or a positive $\overline{\overline{AUCD}}_{ij}$ would suggest that a model on the row has a better performance than a model on the column, since a lower value of $\overline{\overline{MAD}}$ or a higher value of $\overline{\overline{AUC}}$denotes a better model performance. A positive $\overline{\overline{MADD}}_{ij}$or a negative $\overline{\overline{AUCD}}_{ij}$would suggest the opposite. Student's paired t tests with Bonferroni corrections for multiple testing (α = 0.05) all resulted in the rejection of null hypotheses of no difference in $\overline{\overline{MAD}}$ or $\overline{\overline{AUC}}$ among all pairwise models comparison. It was concluded that ANN models had the best performance based on $\overline{MAD}$, followed by RFo, RF, LR2 and LR1 regardless of species. Based on $\overline{AUC}$, RFo models showed the best performance, followed by RF, LR2, LR1and ANN regardless of species.

3.4.5 Model comparison on specificity and sensitivity

Species PINU and QUNI with relatively high observed mortality, species ACRU and PRSE with intermediate observed mortality rates, and species ACSA and FRAM with relatively low mortality were chosen for example in Figure 3.4. A frequent phenomenon was observed that as the value of threshold is increased from 0 to 1, RF and RFo model exerts a slower drop of sensitivity but a continuously higher rise of specificity, compared with other models. No matter of species having low or high mortality, RF and RFo model generally generates a higher value of specificity over thresholds, but only sacrifices a relatively smaller value of sensitivity. The threshold was set to make sure the value of sensitivity being equal to specificity for each species. As is shown in Table 3.7, higher values of thresholds were seen in species with lower mortality,

with extremely high value as 1 appeared in ANN models. RF and RFo models generally had higher value on sensitivity and specificity among five models at the chosen thresholds.

3.4.6 Prediction of periodic survival probability of individual tree

For red maple trees, more than 85% of plots had average individual-tree PSP ≥ 0.95 for contemporary climate predictions (Figure 3.5a). Decreasing PSP under climate scenario A2 in red maple in the 2090s was noted across a broad area of the south area and in central states including Indiana, Ohio, and Kentucky (Figure 3.5b). The decrease of average PSP over a century was severe over time that about 15% of plots had dropped at least 0.3 in average PSP (Figure 3.5c). About 8% of plots have increasing PSP, with the increases were predicted in northern-tier states from Wisconsin to Maine. For eastern white pine trees, most plots showed average PSP ≥ 0.98 for the contemporary period (Figure 3.6a). Under the A2 climate change scenario, the lower values of PSP appeared in most of eastern white pine growing regions, mainly centered in the Appalachian mountain with future PSP intervals ranging from 0.5 to 0.9 (Figure 3.6b). Only 4% of plots had rising PSPs and most of those were located in New England, northern Michigan and Wisconsin over the next century. Within the remaining of 96% of plots, PSP in about two-thirds of them changed between [0.0, -0.1], indicating that that the impact of global warming under the A2 scenario could have a marked effect – although not necessarily catastrophic – on eastern white pine survival across much of its current-day range (Figure 3.6c).

Diameter-related variables were the only tree attributes used to distinguish individual trees within a plot in the RFo model used here. Therefore it was reasonable to see that DBH2 was among the most important variables either for eastern white pine or red maple according to the Random Forests variable importance plot (Figure 3.7). The spline was used to describe the trend of PSP against DBH2 of trees, and the trees per plot will be represented by a different solid spline (Figure 3.8). Within each plot, the difference of mean PSP generally increased with increasing DBH2 until to a vertex and decreased afterward for red pine, which meant that compared with the tree having medium diameter around 1000 cm^2, the tree with extreme large diameter had lower PSP in the 2090s. For eastern white pine, the difference of mean PSP generally increased as DBH2 increased until the tendency tended to flat, which showed that until DBH2 was close to 1000 cm^2, the larger diameter of the tree was, the lower probability of the

tree survived in the 2090s. The PSP seems insensitive to increasing DBH2 when the DBH2 was larger than 1000 cm^2.

PRATIO was relatively important among climate variables on both MeanDecreaseAccuracy and MeanDecreaseGini for red maple (Figure 3.7). So was TDIFF for Eastern white pine. Marginal plots and rank correlation coefficients were generated using predicted difference of average PSP per plot from contemporary to the 2090s ($\Delta PSP = PSP_{2090} - PSP_{contemporary}$) along with climate variable changes from 1990 to 2090 ($\Delta PRATIO = PRATIO_{2090} - PRATIO_{contemporary)}$ and ($\Delta TDIFF = TDIFF_{2090} - TDIFF_{contemporary}$; Figure 3.9). The ΔPRATIO was negatively corresponded to the difference of PSP, which meant that when the ratio of summer precipitation to total precipitation increased in future, average PSP of red maple tended to decrease from contemporary to the 2090s. The ΔTDIFF was negatively corresponded to the difference of PSP, which implied that if summer-winter temperature differential went higher since contemporary period, average PSP of eastern white pine decreased from now.

3.5 Discussion

Individual tree mortality prediction involves the analysis of imbalanced data since observations of tree mortality constitute a very small portion of the data. Most classification algorithms make the assumption that the class distribution in the data set is uniform. When a model is trained on imbalanced data, the learning algorithm tends to optimize the overall prediction accuracy. As a result, the prediction for new observations tends more often toward the majority class unless the difference between the classes is distinctive and the model is trained on a large data set (Bekkar and Alitouche 2013). Threshold and sampling approaches are commonly used to handle this kind of problem. Sampling adjustment creates an artificial balance between two or more classes by oversampling the minority class and/or downsampling the majority class. Down-sampling the majority class may lead to loss of information (Weiss 2004). Oversampling is prone to result in computational difficulty and does not necessarily present a solution to the fundamental issue of lack of data, since it does not introduce new observations (Chen et al. 2004). With these considerations, we used all data set to build model and implemented prediction thresholds to transform probability values into binary alive–dead status that accounted for the imbalanced

nature of the training data. The choice of a decision threshold may be based on prior knowledge of mortality probabilities, or on the average observed mortality rate of the species of interest (Monserud and Sterba 1999). Alternatively, optimal thresholds can be determined from the ROC convex hull when cost and class-distribution information is available (Fawcett 2006). Since the cost of making false decisions was not known in this study, the thresholds were computed by setting sensitivity equal to specificity for all species. The resulting thresholds tended to be close to the average observed mortality rates in the species of interest.

MAD and AUC are commonly used to measure the quality of a classification algorithm. It is not surprising to see that MAD and AUC gave different ranking result for different models, since past work has shown that algorithms designed to minimize error rates may not lead to the best possible AUC values (Cortes and Mohri 2003). MAD is the mean of absolute differences between predictions and corresponding observed values. While it is an effective measure in detecting the probability estimates over all observations, it does not take class affiliations into account. This leads to the result that error rates are not computed separately for each class. Even though a high MAD is achieved by a prediction model, we can't guarantee the model makes a good prediction for both classes. In other words, low values of MAD indicate overall high model accuracy, but the result may represent a very high true positive rate combined with a low true negative rate or *vice versa*. Especially in imbalance data sets, a classifier generally projects new observations more often to the majority class (Janitza et al. 2013); therefore, a classifier is prone to have a high classification accuracy for the majority class, but its accuracy for the minority may be unacceptable. An alternative way of evaluating performance of a classifier that accounts for the false positive and false negative error rates is based on the ROC curve and its associated AUC, which are independent of the decision threshold, the prior class distribution, and costs of misclassification (Bradley 1997). ROC and AUC analyses are therefore thought to be more reliable under conditions where imbalanced training data or misclassification costs are present. (Marrocco et al. 2008; Provost et al. 1998). Although AUC was more applicable in this study due to the unbalanced nature of the training data, we used both due to their common use in past studies and their own merits in comparing algorithms.

The models evaluated here may be compared in terms of their ease of use in terms of data preparation, computational intensity, and ease of use or interpretation. In terms of data preparation, both logistic regression model forms require input variables to be nearly independent and linearly related to the log-odds. Due to possible multicollinearity among variables, a variable selection or data compression scheme, e.g., PCA, should be used to address correlated predictors. Data normalization is required in training artificial neural networks to prevent the premature saturation of hidden nodes, which impedes the learning process (Basheer and Hajmeer 2000). While principal components were used as predictors in training the RF model, only the unprocessed raw data were used to train the RFo model. In terms of computation intensity, the parameters of logistic regression models were estimated using maximum likelihood. Compared with the standard logistic regression LR1, the generalized LR2 model required a longer computation time due to its more complex structure and the possibility of non-convergence when the number of variables is large (Flewelling and Monserud 2002). Adjusting start values of parameters to solve non-convergence and choosing the most suitable optimization method are strongly recommended to seek the best maximum likelihood estimate. Artificial neural networks share several structure similarities to logistic regression model and are identical to logistic regression when no hidden layers are specified (Tu 1996). Unlike logistic regression models, ANN are not designed to fit a logistic curve, only to minimize the sum of squared errors (Hastie et al. 2009). More computation time is thus required because connection weights are continuously adjusted based on calculated errors, which involves back-propagation from the output layer through the hidden layers to the input layer until no obvious additional error reduction is achieved (Guan and Gertner 1991b). In the Random Forests algorithm, each regression tree is grown to its largest extent with recursive splitting of data into two distinct subsets (Ziegler and König 2014). When more trees are included, a greater computational burden is added. In the work done here, standard logistic regression modeling was the most time-efficient among these approaches and Random Forests modeling took the most computer processing time. In terms of the ease of use or interpretation, ANN and Random Forests models did not require the specification of model functional formulas prior to model fitting. They also allowed for greater flexibility in modeling interactions and complex nonlinearities. Predictor variables could be of any type and the techniques are little influenced by outliers, missing data,

or collinearity between predictors (Tu 1996; Ziegler and König 2014). Model flexibility unfortunately increases proneness to overfitting. In order to prevent overfitting, one layer of hidden neurons is deemed sufficient for classifying most data sets. Limiting the number of hidden neurons or restricting the magnitude of weights are also effective ways to prevent overfitting (Dreiseitl and Ohno-Machado 2002). Random Forests are known to be relatively unaffected by problems of overfitting because of the reliance on OOB observations and bootstrap-selected predictors in training (Breiman 2001; Prasad et al. 2006). Logistic model coefficients are readily interpretable based on their signs and magnitudes, making them suitable for ascertaining ecological meaning of predictors. Neural networks are described as a "black-box" approach due to the uninterpretable connection weights. Random Forests are also weak in terms of model interpretability since the individual trees cannot be examined separately without great effort. The use of predictor importance measures, however, makes Random Forests much more interpretable than neural networks. The name "gray-box" has thus been used to describe them (Prasad et al. 2006).

Based on our data, results showed that the model performance was highly influenced by the mortality rates of the different species. Although there were some differences among species, LR1, LR2, ANN, RF and RFo generally perform better for species with lower mortality than for those with higher mortality based on either MAD or AUC criteria. According to the ranks of model performance based on AUC criteria for each species, LR1 and LR2 were the most accurate methods for modeling mortality in species with the lowest observed mortality, whereas RF and RFo were more accurate for species having higher observed mortality

Although the ANN model showed the best performance among the models tested based on MAD for 19 of the 20 species studied, it appeared to be the least accurate based on the AUC. In other words, for this imbalanced data set, ANN was prone to predict individual tree status as live, i.e. the majority class, leading to high classification accuracy for the live class, but low accuracy for trees that died. Clearly the overall accuracy of ANN models comes at the expense of misclassification of the minority class, i.e. low accuracy in predicting mortality. Besides this weakness, ANN tends to have poorer model stability according to the high standard deviations of MAD and AUC in model validation by bootstrap sampling. Much like the results observed here,

Terrin et al. (2003) found that neural networks had poorer performance under external validation than logistic regression with piecewise-linear and quadratic terms or standard logistic regression. In contrast King et al. (2000) showed that ANN were superior to logistic regression and two forms of support vector machine classifiers for classifying noncatastrophic individual tree mortality in West Virginia. Similarly Guan and Gertner (1991b) reported that ANN achieved a lower total sum of squared errors than logistic regression in modeling individual tree mortality in mixed oak-hickory forests in Missouri. Reasons why these studies did not necessarily reach the same conclusions as those drawn here likely include: a) examining overall prediction error rates compared to within-class error rates as was done here; b) differences in the degree of unbalancedness in model development data sets; and c) differences in error rate calculations, e.g., the use of a single instance of data withholding compared to the use of bootstrap-based cross validation.

For mortality data collected with unequal measurement intervals, L, a common approach is to modify the standard logistic function, LR1 to a generalized logistic model, LR2 (Yao et al. 2001). In LR2, PSP can be projected using a variable L rather than a fixed yearly interval. From the properties of LR2 model Eq. [3.2], as L gets large the PSP of a live tree will get small, approaching zero as $L \rightarrow \infty$. In the data used here L varied among species, with averages values ranging between 4.9 and 6.8, and standard deviations ranging from 0.8 to 4.4 years (Table 3.3). Observed MAD values for LR2 were $\leq$ MAD values for LR1 models in all 20 species. Further, AUC values for LR2 were $\geq$ AUC for LR1 models. Both logistic regression methods exhibited comparable model stability based on either MAD or AUC criteria. Our results illustrated that LR2 was preferable over LR1 in predicting tree mortality in data with unequal measurement intervals, which is consistent with findings of Monserud (1976), who concluded that despite only a slight improvement, prediction of individual overstory tree mortality in managed northern hardwoods in Wisconsin was improved by using the length of growing period as a variable exponent rather than as a concomitant explanatory variable.

The fundamental assumption for LR2 in this study is a constant annual mortality or survival rate over the measurement interval L. Nevertheless, some studies considered the PSP as well as diameter or height growth for each tree to vary every year, so they estimated tree survival rates

using iterative methods (Cao 2000; Crecente-Campo et al. 2010). The iterative approach was not adopted here since the length of the periods in this study was very similar and generally quite short – about 5 to 7 years – the differences between the two approaches should be small.

Both Random Forests modeling approaches used here generally achieved favorable results compared to the other models tested. Further, the RFo approach simplified model development in that it did not require the intermediate step of using PCA to eliminate collinearity and reduce the number of predictors to be included in the models(Massy 1965). Our result showed that Random Forests with all original variables, RFo, significantly performs better than Random Forests with PCA factors either on AUC or MAD, although the difference between RF and RFo with respect to AUC is hardly noticeable from Figure 3.2 and Figure 3.3. It is likely because interactions among predictors and nonlinearity are masked when using projected variables (Muñoz and Felicísimo 2004). Random Forests with all original variables saves efforts to conduct projecting procedure and is able to predict more accurate result than Random Forests with PCA factors.

The impact of competition on mortality has been demonstrated in many studies (Bravo-Oviedo et al. 2006; Collet and Le Moguedec 2007; Yaussy et al. 2013). Monserud and Sterba (1999) conducted a conclusion from six major forest species of Austria that a larger diameter tree relatively had a lower mortality rate than a small diameter tree. But from this study, the PSP of the tree in the same plot was not always decreasing as the diameter was larger. The relationship between tree mortality and tree diameter size under climate change scenario was differed species by species. The reason was likely because apart from diameter difference, the tree mortality was dependent on other tree factor as well, such as crown ratio and crown class.

Some important variables such as site index, crown class, recent tree DBH growth were reported to improve the ability of tree mortality ability (Shifley et al. 2006). Since the aim of this paper was for comparing model performance on parametric and nonparametric method based on the same baseline on input information, therefore we paid less attention on finding best variables to attain best fit for each model. These variables were not put into models also due to its specific reason. For example, site index information was less useful when species for each plot was not same as the targeted species; Crown class label was missing in thousands of tree records. We

would lose other useful tree information at the same time when we delete missing record; Diameter in previous inventory was not straightforwardly provided in recent inventory FIA record database and the accuracy of previous record need to be verified. Due to the above reason, we neglect the recent diameter growth in the model.

3.6 Summary

Three modelling techniques, LR, ANN, and RF, were compared for predicting individual tree mortality for 20 eastern species in Unites States with PCA factors. Compared with other models, LR1 and LR2 had better model performance for tree species with low mortality due to a higher AUC and a consistently lower standard deviation for all species. LR2 generally behaved better than LR1 for most species based on either MAD or AUC. ANN was not a good method in class affiliation due to a poor behavior in terms of AUC, although it has the best performance as to MAD. RFo, Random Forests model with all indicators included, generally performed well no matter with MAD or AUC, especially sound for species with medium and high mortality. At a chosen threshold, RFo frequently achieved the equally highest value of sensitivity and specificity among five approaches. Moreover, Random Forests model was more user-friendly since it could incorporate all variables into model without PCA transformation, resulting in a significant better model performance either on MAD or AUC regardless of species. Periodic survival probability prediction of trees based on red maple and eastern white pine showed that tree mortality at lower latitude was likely to be severer than that at higher latitude under climate change A2 scenario.

3.7 Reference

Basheer IA, Hajmeer M, 2000. Artificial neural networks: fundamentals, computing, design, and application. Journal of Microbiological Methods 43(1):3-31.

Bekkar M, Alitouche TA, 2013. Imbalanced data learning approaches review. International Journal of Data Mining & Knowledge Management Process 3(4):15-33.

Bishop CM, 1994. Neural networks and their applications. Review of Scientific Instruments 65(6):1803-1832.

Bradley AP, 1997. The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognition 30(7):1145-1159.

Bravo-Oviedo A, Sterba H, del Río M, Bravo F, 2006. Competition-induced mortality for Mediterranean Pinus pinaster Ait. and P. sylvestris L. Forest Ecology and Management 222(1–3):88-98.

Breda N, Huc R, Granier A, Dreyer E, 2006. Temperate forest trees and stands under severe drought: a review of ecophysiological responses, adaptation processes and long-term consequences. Ann. For. Sci. 63(6):625-644.

Breiman L, 2001. Random Forests. Machine Learning 45(1):5-32.

Buford MA, 1983. Probability distributions as models for mortality in loblolly pine (pinus taeda l.) plantations Ann Arbor: North Carolina State University. 73-73 p.

Cailleret M, Nourtier M, Amm A, Durand-Gillmann M, Davi H, 2014. Drought-induced decline and mortality of silver fir differ among three sites in Southern France. Ann. For. Sci. 71(6):643-657.

Cao V, 2000. Prediction of annual diameter growth and survival for individual trees from periodic measurements. Forest Science 46(1):127-131.

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

Chen G, Metz MR, Rizzo DM, Dillon WW, Meentemeyer RK, 2015. Object-based assessment of burn severity in diseased forests using high-spatial and high-spectral resolution MASTER airborne imagery. ISPRS Journal of Photogrammetry and Remote Sensing 102(0):38-47.

Collet C, Le Moguedec G, 2007. Individual seedling mortality as a function of size, growth and competition in naturally regenerated beech seedlings. Forestry 80(4):359-370.

Cortes C, Mohri M, 2003. AUC Optimization vs. Error Rate Minimization. Presented at the Neural Information Processing Systems.

Crecente-Campo F, Soares P, Tomé M, Diéguez-Aranda U, 2010. Modelling annual individual-tree growth and mortality of Scots pine with data obtained at irregular measurement intervals and containing missing observations. Forest Ecology and Management 260(11):1965-1974.

Crookston N, 2012. Details on spatial extents, temporal information and data elements http://forest.moscowfsl.wsu.edu/climate/details.php (accessed July 9, 2013,).

Cutler DR, Edwards TC, Jr., Beard KH, Cutler A, Hess KT, Gibson J, Lawler JJ, 2007. Random forests for classification in ecology. Ecology 88(11):2783-2792.

de Toledo JJ, Magnusson WE, Castilho CV, Nascimento HEM, 2012. Tree mode of death in Central Amazonia: Effects of soil and topography on tree mortality associated with storm disturbances. Forest Ecology and Management 263(0):253-261.

Dobbertin M, Biging GS, 1998. Using the non-parametric classifier cart to model forest tree mortality. Forest Science 44(4):507-516.

Dreiseitl S, Ohno-Machado L, 2002. Logistic regression and artificial neural network classification models: a methodology review. Journal of Biomedical Informatics 35(5):352-359.

Efron B. 1982. The jackknife, the bootstrap, and other resampling plans. Society for Industrial and Applied Mathematics, Philadelphia, PA. 35-38.

Fan Z, Kabrick JM, Shifley SR, 2006. Classification and regression tree based survival analysis in oak-dominated forests of Missouri's Ozark highlands. Canadian Journal of Forest Research 36(7):1740-1748.

Fawcett T, 2006. An introduction to ROC analysis. Pattern Recognition Letters 27(8):861-874.

Fenneman NM, and Johnson, D.W., 1946. Physiographic divisions of the conterminous United States Washington, D.C.: U.S. Geological Survey (USGS). Special map series, scale 1:7,000,000.

Flewelling JW, Monserud RA, 2002. Comparing methods for modelling tree mortality. In: Second forest vegetation simulator conference: USDA For. Serv., Rocky Mountain Research Station.

Fortin M, Bedard S, DeBlois J, Meunier S, 2008. Predicting individual tree mortality in northern hardwood stands under uneven-aged management in southern Quebec, Canada. Ann. For. Sci. 65(2).

Franklin JF, Shugart HH, Harmon ME, 1987. Tree death as an ecological process. Bioscience 37(8):550.

Groom JD, Hann DW, Temesgen H, 2012. Evaluation of mixed-effects models for predicting Douglas-fir mortality. Forest Ecology and Management 276:139-145.

Guan BT, Gertner G, 1991a. Modeling red pine tree survival with an artificial neural network. Forest Science 37(5):1429-1440.

Guan BT, Gertner G, 1991b. Using a parallel distributed processing system to model individual tree mortality. Forest Science 37(3):871-885.

Hamilton DA, 1986. A logistic model of mortality in thinned and unthinned mixed conifer stands of northern idaho. Forest Science 32(4):989-1000.

Hamilton DA, Edwards BM. 1976. Modeling the probability of individual tree mortality Intermountain Forest and Range Experiment Station, Forest Service, U.S. Dept. of Agriculture, Ogden, Utah. 2.

Hasenauer H, Merkl D, Weingartner M, 2001. Estimating tree mortality of Norway spruce stands with neural networks. Advances in Environmental Research 5(4):405-414.

Hastie T, Tibshirani R, Friedman JH. 2009. The elements of statistical learning: data mining, inference, and prediction Springer, New York, NY. 393-399.

Hofmeyer PV, Kenefic LS, Seymour RS, 2010. Historical Stem Development of Northern White-Cedar (Thuja occidentalis L.) in Maine. North. J. Appl. For. 27(3):92-96.

Hosmer DW, Lemeshow S. 1989. Applied logistic regression Wiley, New York.

Iverson LR, Prasad AM, Matthews SN, Peters M, 2008. Estimating potential habitat for 134 eastern US tree species under six climate scenarios. Forest Ecology and Management 254(3):390-406.

Janitza S, Strobl C, Boulesteix A-L, 2013. An AUC-based permutation variable importance measure for random forests. BMC Bioinformatics 14:119.

Kenkel NC, Hendrie ML, Bella IE, 1997. A Long-Term Study of Pinus banksiana Population Dynamics. Journal of Vegetation Science 8(2):241-254.

Kiernan D, Bevilacqua E, Nyland R, Zhang L, 2009. Modeling tree mortality in low- to medium-density uneven-aged hardwood stands under a selection system using generalized estimating equations. Forest Science 55(4):343-351.

King SL, Bennett KP, List S, 2000. Modeling noncatastrophic individual tree mortality using logistic regression, neural networks, and support vector methods. Computers and Electronics in Agriculture 27(1–3):401-406.

Kohavi R, 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. Presented at the Proceedings of the 14th international joint conference on Artificial intelligence - Volume 2, Montreal, Quebec, Canada.

Kuhn M, Johnson K. 2013. Applied Predictive Modeling Springer, New York. 35-37.

Lee Y, 1971. Predicting mortality for even-aged stands of lodgepole pine. The Forestry Chronicle 47(1):29-32.

Lendasse A, Simon G, Wertz V, Verleysen M, 2005. Fast bootstrap methodology for regression model selection. Neurocomputing 64(0):161-181.

Liang KY, Zeger SL, 1986. Longitudinal data analysis using generalized linear models. Biometrika 73(1):13-22.

Marrocco C, Duin RPW, Tortorella F, 2008. Maximizing the area under the ROC curve by pairwise feature combination. Pattern Recognition 41(6):1961-1974.

Massy WF, 1965. Principal Components Regression in Exploratory Statistical Research. Journal of the American Statistical Association 60(309):234-256.

McCulloch W, Pitts W, 1943. A logical calculus of the ideas immanent in nervous activity. Bulletin of Mathematical Biophysics 5(4):115-133.

Moisen GG, Frescino TS, 2002. Comparing five modelling techniques for predicting forest characteristics. Ecological modelling 157(2–3):209-225.

Monserud RA, 1976. Simulation of forest mortality. For. Sci. :22:438-444.

Monserud RA, Sterba H, 1999. Modeling individual tree mortality for Austrian forest species. Forest Ecology and Management 113(2–3):109-123.

Muñoz J, Felicísimo ÁM, 2004. Comparison of statistical methods commonly used in predictive modelling. Journal of Vegetation Science 15(2):285-292.

Negron J, McMillin J, Anhold J, D C, 2009. Bark beetle-caused mortality in a drought-affected ponderosa pine landscape in Arizona, USA. Forest Ecology and Management:1353–1362.

Prasad AM, Iverson LR, Liaw A, 2006. Newer classification and regression tree techniques: bagging and Random Forests for ecological prediction. Ecosystems 9(2):181-199.

Provost FJ, Fawcett T, Kohavi R, 1998. The case against accuracy estimation for comparing induction algorithms. Presented at the International Conference on Machine Learning.

Quinn GP, Keough MJ. 2002. Experimental design and data analysis for biologists Cambridge University Press, Cambridge, UK. 453.

R Core Team, 2012. R: A language and environment for statistical computing Vienna, Austria: R Foundation for Statistical Computing.

Radtke PJ, Herring ND, Loftis DL, Keyser CE, 2012. Evaluating forest vegetation simulator predictions for southern appalachian upland hardwoods with a modified mortality model. Southern Journal of Applied Forestry 36(2):61-70.

Rehfeldt GE, Crookston NL, Warwell MV, Evans JS, 2006. Empirical analyses of plant‐climate relationships for the western United States. International Journal of Plant Sciences 167(6):1123-1150.

Reynolds JH, Ford ED, 2005. Improving competition representation in theoretical models of self-thinning: a critical review. Journal of Ecology 93(2):362-372.

Rouault G, Candau JN, Lieutier F, Nageleisen LM, Martin JC, Warzee N, 2006. Effects of drought and heat on forest insect populations in relation to the 2003 drought in Western Europe. Ann. For. Sci. 63(6):613-624.

Ruiz-Benito P, Lines ER, Gomez-Aparicio L, Zavala MA, Coomes DA, 2013. Patterns and drivers of tree mortality in iberian forests: climatic effects are modified by competition. PLOS ONE 8(2):e56843.

Sakai S, Kobayashi K, Toyabe S-i, Mandai N, Kanda T, Akazawa K, 2007. Comparison of the levels of accuracy of an artificial neural network model and a logistic regression model for the diagnosis of acute appendicitis. J Med Syst 31(5):357-364.

Shifley, Fan Z, Kabrick JM, Jensen RG, 2006. Oak mortality risk factors and mortality estimation. Forest Ecology and Management 229(1-3):16-26.

Somers GL, Langdon OG. 1980. Predicting mortality with a weibull distribution.

Steiner KC, Finley JC, Gould PJ, Fei SL, McDill M, 2008. Oak regeneration guidelines for the central Appalachians. North. J. Appl. For. 25(1):5-16.

Streiner DL, Cairney J, 2007. What's under the ROC? an introduction to receiver operating characteristics curves. Canadian Journal of Psychiatry 52(2):121-128.

Terrin N, Schmid CH, Griffith JL, D'Agostino Sr RB, Selker HP, 2003. External validity of predictive models: a comparison of logistic regression, classification trees, and neural networks. Journal of Clinical Epidemiology 56(8):721-729.

Tu JV, 1996. Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. Journal of Clinical Epidemiology 49(11):1225-1231.

USDA Natural Resources Conservation Service, 2013. Geospatial data gateway http://datagateway.nrcs.usda.gov/ (accessed July, 24, 2013).

Vanclay JK, Sands PJ, 2009. Calibrating the self-thinning frontier. Forest Ecology and Management 259(1):81-85.

Venables WN, Ripley BD. 2002. Modern applied statistics with S. Fourth Edition. Springer, New York.

Weiss GM, 2004. Mining with rarity: a unifying framework. SIGKDD Explor. Newsl. 6(1):7-19.

Woodall CW, Grambsch PL, Thomas W, 2005. Applying survival analysis to a large-scale forest inventory for assessment of tree mortality in Minnesota. Ecological Modelling 189(1–2):199-208.

Yang Y, Huang S, 2013. A generalized mixed logistic model for predicting individual tree survival probability with unequal measurement intervals. Forest Science 59(2):177-187.

Yang Y, Titus SJ, Huang S, 2003. Modeling individual tree mortality for white spruce in Alberta. Ecological Modelling 163(3):209-222.

Yao X, Titus SJ, MacDonald SE, 2001. A generalized logistic model of individual tree mortality for aspen, white spruce, and lodgepole pine in Alberta mixedwood forests. Canadian Journal of Forest Research 31(2):283-291.

Yaussy DA, Iverson LR, Matthews SN, 2013. Competition and climate affects US hardwood-forest tree mortality. Forest Science 59(4):416-430.

Ziegler A, König IR, 2014. Mining data with random forests: current options for real-world applications. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 4(1):55-63.

Table 3.1 Climate variables used as predictors in model development.

Acronym	Definition	Max	Min	Mean	SD
MAT	Mean annual temperature (°C)	24.9	0.5	11.9	5.5
MTCM	Mean temperature in the coldest month	20.4	-18.1	-1.3	8.0
MMIN	Minimum temperature in the coldest month	16.7	-24.4	-7.2	7.8
MTWM	Mean temperature in the warmest month	30.5	13.3	23.6	3.4
MMAX	Maximum temperature in the warmest month	38.4	17.7	30.0	3.3
MAP	Mean annual precipitation (mm)	2216	223	1094	272.7
GSP	Growing season precipitation, April-September	1106	161	604	106.3
TDIFF	Summer-winter temperature differential, (MTWM – MTCM)	37.6	8.2	24.9	5.2
DD5	Degree-days > 5°C	7217	849	3302	1284
DD0	Degree-days < 0°C	1900	0	437.0	521.1
MMINDD0	Minimum degree-days <0°C	2958	0	926.2	819.1
SDAY	Julian date of the last spring freeze	175	0	109.1	28.8
FDAY	Julian date of the first autumn freeze	365	241	292	23.4
GSDD5	Degree-days >5°C accumulating within the frost-free period	7217	632	2813	1187
D100	Julian date the sum of degree-days >5°C reaches 100	158	7.0	81.8	39.1
AMI	Annual moisture index, DD5/MAP	22.8	0.5	3.1	1.5
SMI	Summer moisture index, GSDD5/GSP	27.1	0.8	4.7	2.1
SMRPB	Summer precipitation balance	3.5	0.5	1.1	0.2
SMRSPRPB	Summer/Spring precipitation balance	5.5	0.4	1.1	0.4
PRATIO	Ratio of summer precipitation to total precipitation, GSP/MAP	0.8	0.4	0.6	0.1

Note:[†]Temperature-related variables defined in units of °C and precipitation values in mm.

Table 3.2 Soils variables used as predictors in model development.

Abbreviation	Soil Properties	Max	Min	Mean	SD
SBD	Soil bulk density (g/cm^3) - weighted average of components	2.2	0.3	1.6	0.1
SILT	Percent silt (0.002 to 0.05 mm) - weighted average of components	90.0	0.0	32.4	18.5
SAND	Percent sand (0.05mm to 2.0 mm) - weighted average of components	99.0	0.4	46.1	26.4
CLAY	Percent clay (< 0.002 mm) - weighted average of components	80.8	0.0	21.5	13.7
KFFACT	Soil erodibility factor - weighted average of components	0.6	0.0	0.3	0.1
NO10	Percent soil passing #10 sieve (coarse) - weighted average of components	100.0	19.1	85.8	15.2
NO200	Percent soil passing #200 sieve (fine) - weighted average of components	100.0	1.8	51.7	23.9
OM	Organic matter content (% by weight) - weighted average of components	98.0	0.0	1.2	2.0
pH	Soil pH- weighted average of components	9.3	2.8	5.6	0.9
TAWC	Total available water capacity - weighted average of components (0 to 100 cm)	0.5	0.0	0.1	0.0
KSAT	Saturated hydraulic conductivity - weighted average of components (0 to 100 cm)	615.6	0.1	26.6	33.6

Table 3.3 Summary statistics and mortality status over one remeasurement interval for 20 most frequently occurring eastern species in the FIA database

Species code	Scientific name	Common name	Obs	Plots	Mortality	L		QMD (cm)	
					%	Mean	sd	Mean	sd
PINU	loblolly pine	*Pinus taeda*	147134	11296	5.0	6.6	3.7	20.6	4.9
ACRU	red maple	*Acer rubrum*	92139	18964	2.3	5.2	1.6	24.5	4.6
ACSA	sugar maple	*Acer saccharum*	51635	8171	0.4	5.0	0.8	25.6	4.5
QUAL	white oak	*Quercus alba*	42026	11208	2.2	5.2	2.1	25.9	4.7
LIST	sweetgum	*Liquidambar styraciflua*	40508	10085	5.3	6.3	3.4	24.8	5.4
LITU	yellow-poplar	*Liriodendron tulipifera*	25640	7103	4.4	5.3	2.3	26.1	5.7
QURU	northern red oak	*Quercus rubra*	25174	8082	1.9	5.0	1.4	26.0	5.0
PIST	eastern white pine	*Pinus strobus*	19504	4016	0.9	4.9	0.9	25.1	5.2
PIEC	shortleaf pine	*Pinus echinata*	19289	4026	6.9	6.0	3.6	23.5	4.3
QUST	post oak	*Quercus stellata*	19064	4880	3.8	6.7	4.4	23.8	4.3
QUPR	chestnut oak	*Quercus prinus*	18470	3398	2.4	5.0	1.4	26.2	4.6
PRSE	black cherry	*Prunus serotina*	18115	7631	2.5	5.3	1.8	24.2	5.1
QUVE	black oak	*Quercus velutina*	17671	6082	3.8	5.4	2.4	25.5	4.9
TSCA	eastern hemlock	*Tsuga canadensis*	17084	2990	0.7	5.0	0.8	25.9	4.6
JUVI	eastern redcedar	*Juniperus virginiana*	15407	3738	3.6	5.4	2.5	22.2	4.0
QUNI	water oak	*Quercus nigra*	15379	5359	7.0	6.8	3.6	24.9	5.7
FRPE	green ash	*Fraxinus pennsylvanica*	14719	4563	3.7	5.9	2.8	25.2	5.6
FRAM	white ash	*Fraxinus americana*	14279	5359	1.4	5.2	1.4	24.9	5.0
FAGR	American beech	*Fagus grandifolia*	13811	4221	0.8	5.2	1.5	26.3	5.1
ULAM	American elm	*Ulmus americana*	11604	6148	2.9	5.5	2.3	25.4	5.7

Reference link for species code at http://plants.usda.gov/java/downloadData?fileName=plantlst.txt&static=true

Table 3.4 Variables for model development

	Variables
Shortleaf pine and Black cherry	N, BASC, IDBH,DBH2, QMD, L PC1, PC2, PC3, PC4, PC5
Other species	BA, N, BASC, QMD, L, IDBH, DBH2, PC1, PC2, PC3, PC4, PC5

Table 3.5 Ranks[†] (R) of model performance, by species, as measured by mean absolute difference (MAD) and area under the ROC curve (AUC).

	$\overline{MAD}$									$\overline{AUC}$								
Species	LR1	R	LR2	R	ANN	R	RF	R	RFo	R	LR1	R	LR2	R	ANN	R	RF	R	RFo	R
PINU	0.0936	5	0.0933	4	0.0858	3	0.0759	2	0.0698	1	0.6552	4	0.6192	5	0.7002	3	0.7648	2	0.7670	1
ACRU	0.0422	5	0.0420	4	0.037	1	0.0389	3	0.0375	1	0.8741	3	0.8780	3	0.8743	3	0.8837	1	0.8816	2
ACSA	0.0077	4	0.0076	4	0.0062	1	0.0074	3	0.0073	2	0.9389	2	0.9451	1	0.8937	3	0.8676	4	0.8580	4
QUAL	0.0431	5	0.0428	4	0.031	1	0.0424	3	0.0415	2	0.7607	4	0.7675	3	0.7145	5	0.7912	2	0.7983	1
LIST	0.0989	5	0.0986	4	0.0847	1	0.0945	3	0.0901	2	0.5811	4	0.5562	5	0.6024	3	0.6725	1	0.6709	1
LITU	0.0829	5	0.0825	4	0.0631	1	0.0787	3	0.0761	2	0.6586	4	0.6705	3	0.6553	4	0.7369	2	0.7416	1
QURU	0.0352	5	0.0347	4	0.0284	1	0.0339	3	0.0332	2	0.8797	3	0.8906	1	0.8338	5	0.8768	3	0.8831	2
PIST	0.0176	5	0.0174	4	0.0149	1	0.0158	3	0.0152	1	0.9004	2	0.9079	1	0.8333	5	0.8823	3	0.8865	3
PIEC	0.1239	5	0.1224	4	0.0999	1	0.1077	3	0.1026	2	0.7084	5	0.7239	3	0.7296	3	0.7982	1	0.7979	1
QUST	0.071	5	0.0705	4	0.053	1	0.0658	3	0.0634	2	0.673	4	0.7009	3	0.6448	5	0.7402	1	0.7411	1
QUPR	0.0473	5	0.0472	4	0.0344	1	0.0437	3	0.0413	2	0.675	4	0.7032	3	0.6388	5	0.7322	1	0.7337	1
PRSE	0.0459	5	0.0457	4	0.0385	1	0.0447	3	0.0439	2	0.8542	4	0.8580	3	0.8302	5	0.8718	2	0.8751	1
QUVE	0.0702	5	0.0692	4	0.0571	1	0.0661	3	0.0651	2	0.8068	4	0.8283	3	0.7828	5	0.8473	2	0.8508	1
TSCA	0.014	4	0.0140	4	0.0103	1	0.01	1	0.0093	3	0.9261	1	0.9250	1	0.8038	5	0.912	3	0.9008	3
JUVI	0.0686	5	0.0681	4	0.0537	1	0.0644	3	0.0616	2	0.7303	4	0.7449	3	0.6846	5	0.7672	2	0.7793	1
QUNI	0.1285	4	0.1284	4	0.1014	1	0.1209	3	0.1171	2	0.5854	3	0.5417	5	0.5872	3	0.6699	1	0.6673	1
FRPE	0.0673	5	0.0672	4	0.0548	1	0.0635	3	0.0618	2	0.788	3	0.7887	3	0.7429	5	0.8153	1	0.8168	1
FRAM	0.0274	3	0.0272	3	0.0227	1	0.0273	3	0.0267	2	0.8451	2	0.8603	1	0.756	5	0.8224	4	0.8491	2
FAGR	0.0158	3	0.0158	3	0.0126	1	0.0157	3	0.0153	2	0.8582	1	0.8570	1	0.7248	5	0.8086	4	0.8219	3
ULAM	0.0538	4	0.0537	4	0.0427	1	0.0507	3	0.0497	2	0.8196	3	0.8204	3	0.7712	5	0.861	2	0.8735	1

[†]only differences deemed significant by Student's paired t-test using Bonferroni corrections for multiple testing ($\alpha = .05$) are assigned different ranks.

Table 3.6 Student's t test on model comparison over 20 species: The overall mean, $\overline{\overline{MADD}}$, of 800 paired differences (MADD = MAD $_{row}$ – MAD $_{column}$); the median, $\overline{\overline{AUCD}}$ of 800 paired differences (AUCD = AUC $_{row}$- AUC $_{column}$)

	LR2	ANN	RF	RFo
LR1	$\overline{\overline{MADD}}$= 0.0003*	$\overline{\overline{MADD}}$ = 0.011*	$\overline{\overline{MADD}}$ = 0.004*	$\overline{\overline{MADD}}$ = 0.006*
	$\overline{\overline{AUCD}}$ = -0.003*	$\overline{\overline{AUCD}}$ 0.036*	$\overline{\overline{AUCD}}$ = -0.03*	$\overline{\overline{AUCD}}$ = -0.034*
LR2		$\overline{\overline{MADD}}$= 0.01*	$\overline{\overline{MADD}}$= 0.004*	$\overline{\overline{MADD}}$= 0.006*
		$\overline{\overline{AUCD}}$ = 0.039*	$\overline{\overline{AUCD}}$ = -0.027*	$\overline{\overline{AUCD}}$ = -0.03*
ANN			$\overline{\overline{MADD}}$ = -0.007*	$\overline{\overline{MADD}}$ = -0.005*
			$\overline{\overline{AUCD}}$ = -0.066*	$\overline{\overline{AUCD}}$ = -0.07*
RF1				$\overline{\overline{MADD}}$ = 0.002*
				$\overline{\overline{AUCD}}$ = -0.004*

Note: * indicates significant difference between models based on Student's paired t-test using Bonferroni corrections for multiple testing (α = 0.05).

Table 3.7 Model comparison on sensitivity and specificity at selected thresholds

	Sensitivity					Threshold				
	LR1	LR2	ANN	RF	RFo	LR1	LR2	ANN	RF	RFo
PINU	0.618	0.584	0.579	0.704	0.713	0.952	0.954	0.971	0.980	0.988
ACRU	0.794	0.803	0.776	0.803	0.808	0.969	0.973	0.968	0.984	0.990
ACSA	0.887	0.893	0.698	0.843	0.901	0.991	0.990	1.000	0.998	0.998
QUAL	0.707	0.729	0.676	0.718	0.736	0.975	0.976	0.979	0.984	0.986
LIST	0.557	0.536	0.551	0.610	0.602	0.947	0.950	0.952	0.964	0.974
LITU	0.611	0.628	0.635	0.667	0.690	0.958	0.959	0.963	0.968	0.974
QURU	0.801	0.805	0.790	0.773	0.808	0.978	0.980	0.978	0.990	0.992
PIST	0.828	0.836	0.995	0.794	0.872	0.988	0.988	1.000	0.998	0.998
PIEC	0.643	0.658	0.545	0.740	0.725	0.932	0.932	0.957	0.944	0.962
QUST	0.608	0.666	0.586	0.700	0.681	0.962	0.963	0.978	0.978	0.986
QUPR	0.612	0.639	0.639	0.676	0.675	0.975	0.974	0.980	0.990	0.994
PRSE	0.772	0.767	0.755	0.795	0.806	0.970	0.976	0.985	0.972	0.978
QUVE	0.719	0.753	0.689	0.765	0.766	0.961	0.961	0.966	0.968	0.974
TSCA	0.844	0.843	0.703	0.891	0.904	0.990	0.990	1.000	0.996	0.998
JUVI	0.671	0.676	0.649	0.719	0.726	0.964	0.967	0.969	0.974	0.982
QUNI	0.563	0.527	0.540	0.627	0.635	0.929	0.935	0.942	0.952	0.962
FRPE	0.689	0.718	0.689	0.760	0.756	0.959	0.963	0.962	0.970	0.976
FRAM	0.753	0.785	0.737	0.771	0.816	0.983	0.982	0.987	0.990	0.988
FAGR	0.767	0.737	0.847	0.745	0.753	0.991	0.992	1.000	0.996	0.998
ULAM	0.724	0.741	0.606	0.756	0.783	0.967	0.970	0.996	0.978	0.982

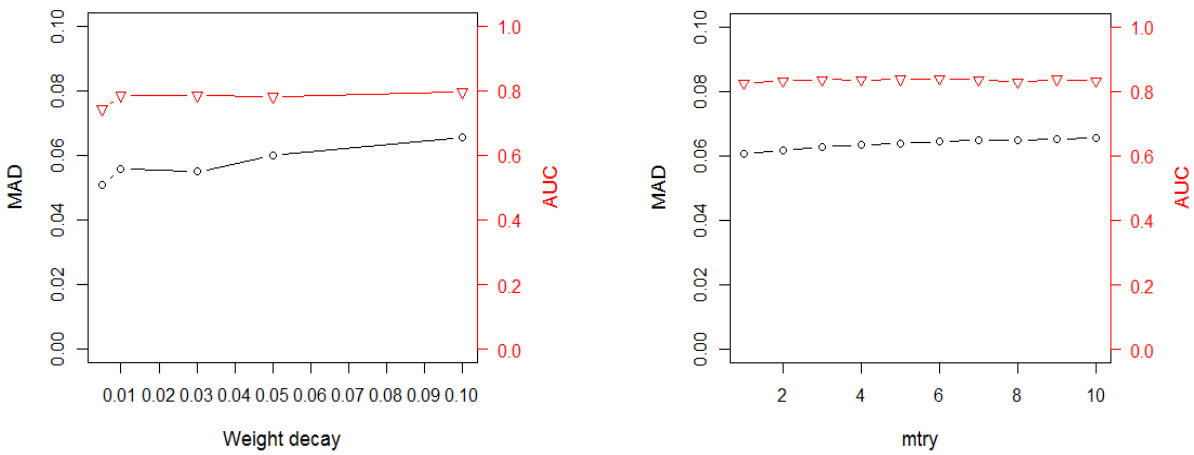


Figure 3.1Parameter tunning plots for training ANN and RF (Species FRPE as example).

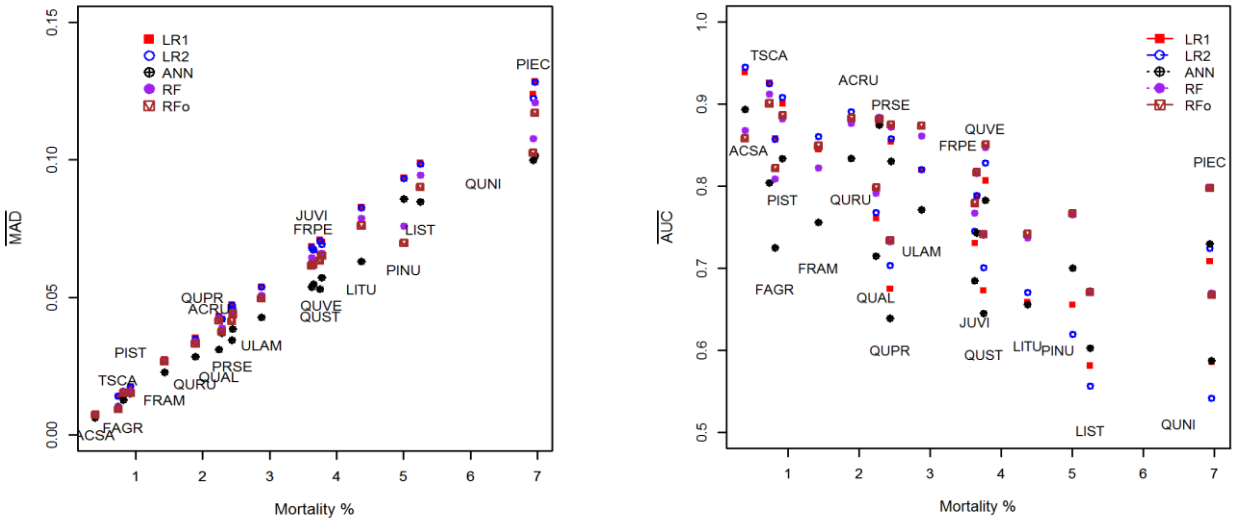


Figure 3.2 Model accuracy statistics $\overline{MAD}$ and $\overline{AUC}$ computed over 40 bootstrap samples for logistic regression (LR) models 1 and 2, artificial neural network (ANN) models, and Random Forests (RF) models.

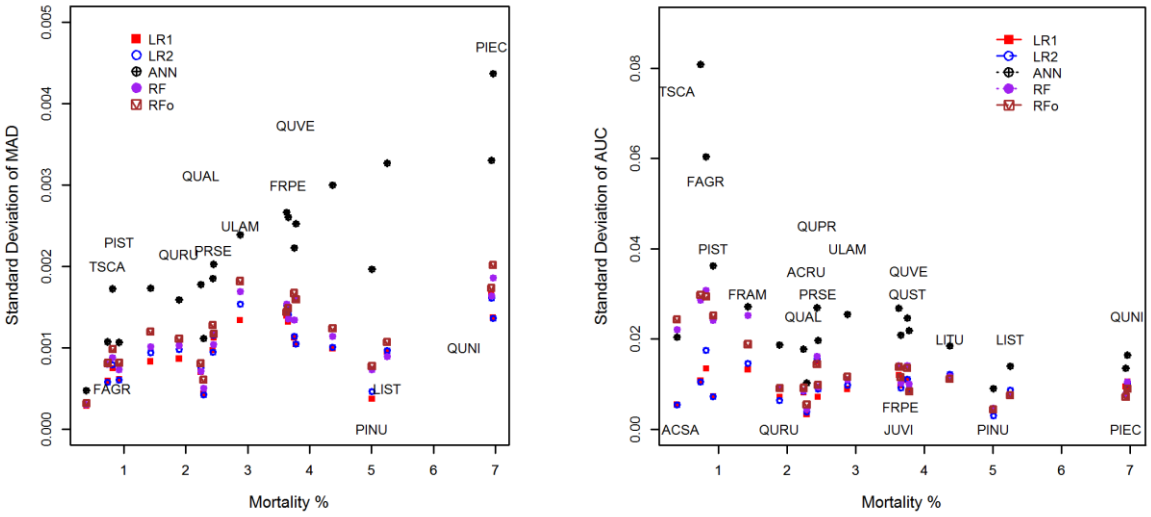


Figure 3.3 Model stability measured as the standard deviation of MAD and AUC statistics computed over 40 bootstrap samples.

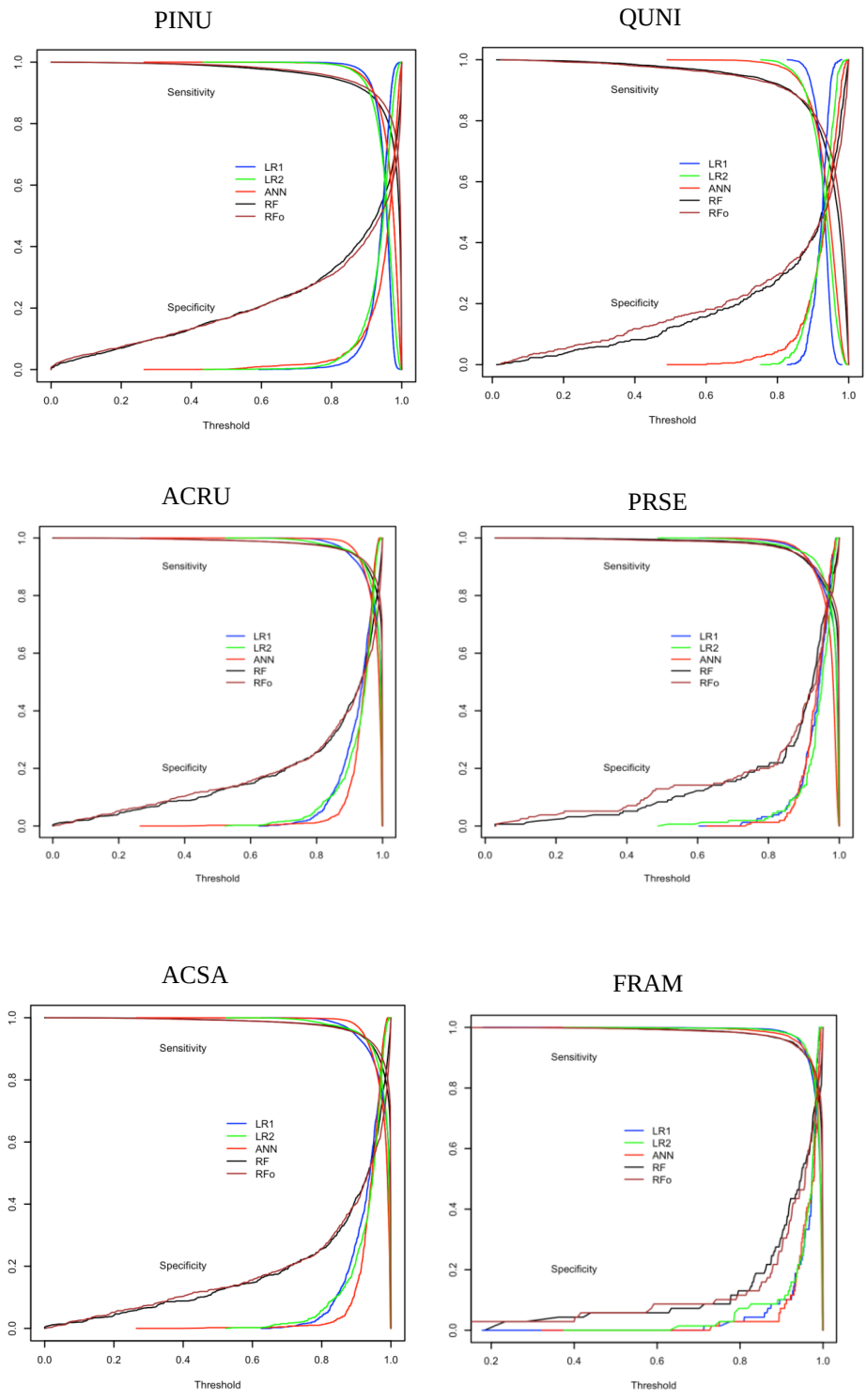


Figure 3.4 The ROC curve performance of models developed for selected species with relatively high mortality (PINU top left; QUNI top right), intermediate mortality (ACRU middle left; PRSE middle right), and low mortality (ACSA bottom left; FRAM bottom right).

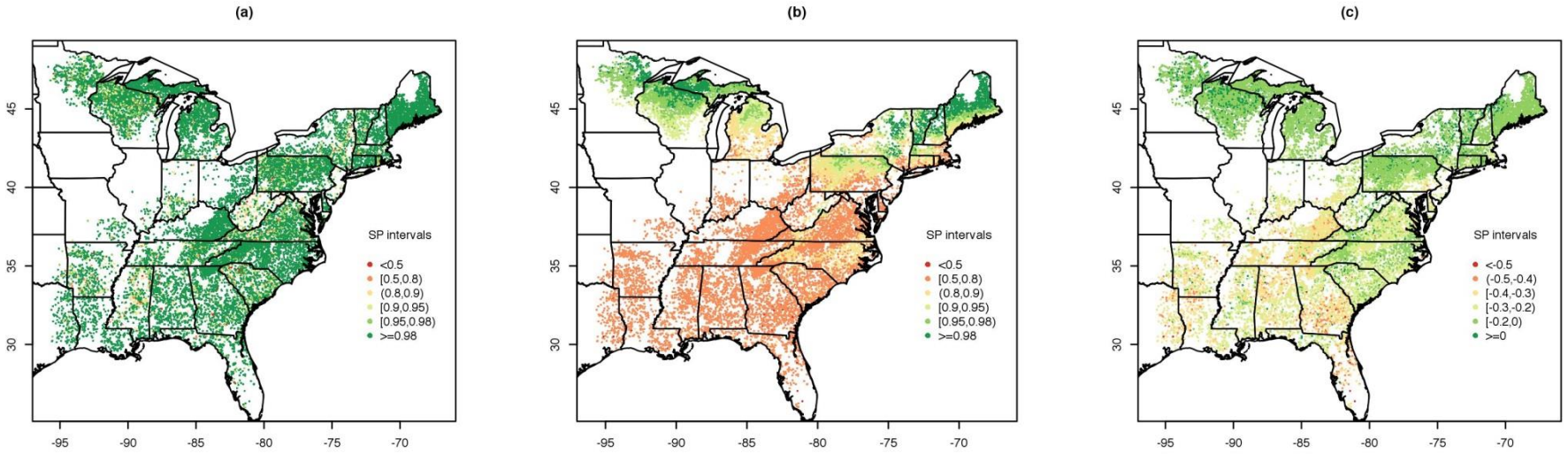


Figure 3.5 Spatial map of predicted average PSP at the contemporary (a), in the 2090s under scenario A2 (b), and predicted change of average PSP of red maple (ACRU) over a century under climate change scenario A2.

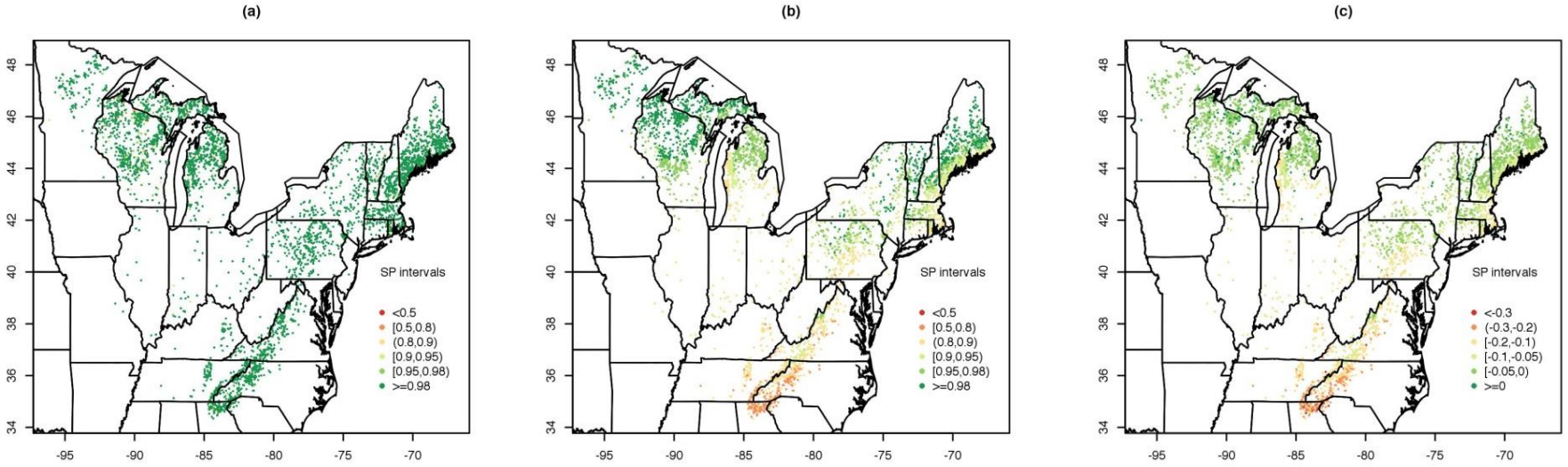


Figure 3.6 Spatial map of predicted average PSP at the contemporary (a), in the 2090s under scenario A2 (b), and predicted change of average PSP of Eastern white pine (PIST) over a century under climate change scenario A2.

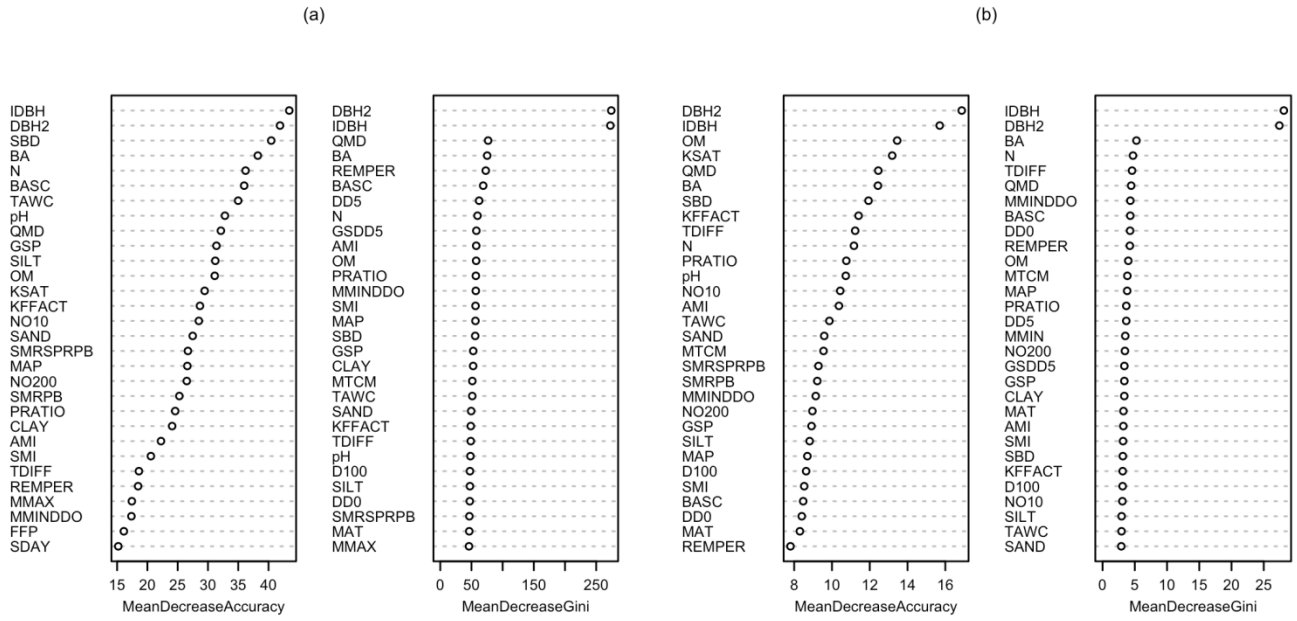


Figure 3.7 Variable importance plots of red maple (ACRU) (a) and white eastern pine (PIST) (b)

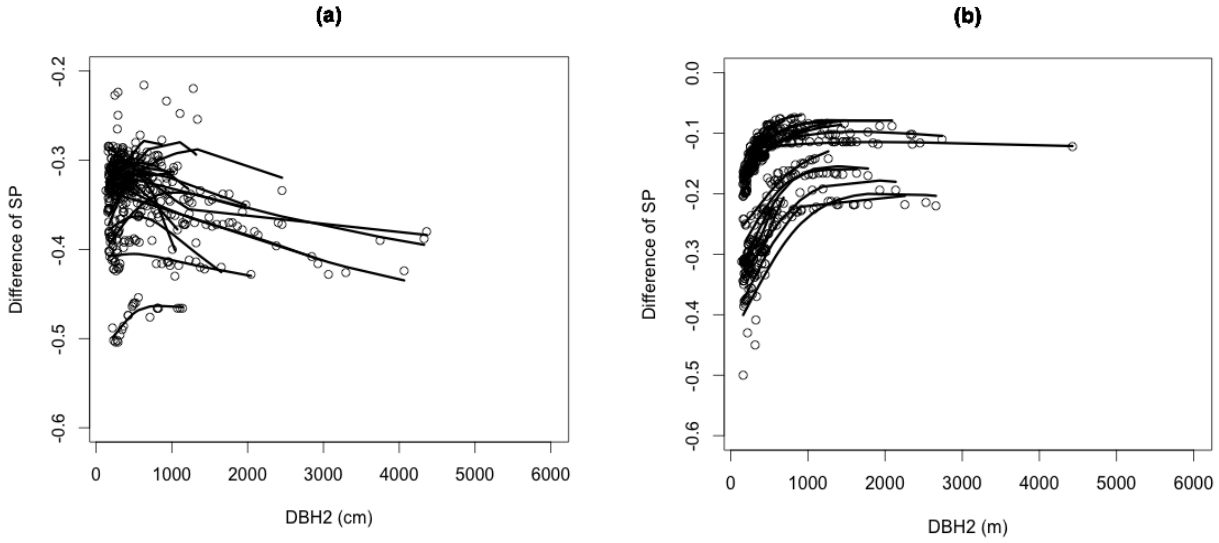


Figure 3.8 The relationship between DBH2 and the difference of PSP of trees in selected plots over one century ($PSP_{2090} - PSP_{contemporary}$)

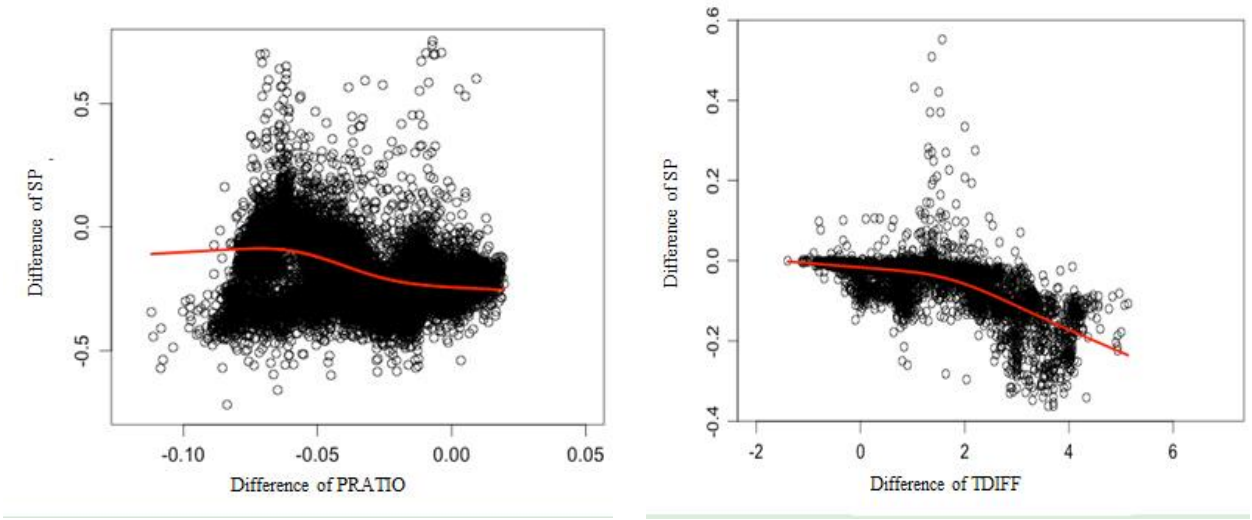


Figure 3.9 The relationship between climate variables and the difference of PSP of trees in selected plots over one century ($PSP_{2090} - PSP_{contemporary}$)

Chapter 4 Climate and Soils-based Models of tree diameter increment in Eastern U.S. Tree Species

Abstract

Concerns of the effect of climate change on forest diameter growth have urged the need to accurately predict forest diameter growth from climate, edaphic and tree and stand related variables. Although traditional linear or nonlinear regression models have been widely-used to simulate and predict tree growth, the challenge on modeling nonlinear relationship and interaction among variables has limited its application in dealing with complex plant-climate relationships. An alternative nonparametric approach, Random Forests (RF) was proposed to develop climate-sensitive annual diameter increment (ADI) model. The multiple linear regression model (MLR) was considered as a baseline model to compare relative model performance. Species-specific tree and stand related information from the USDA Forest Service Forest Inventory and Analysis program were linked to contemporary climate data and soil properties mapped in the USDA Soil Survey Geographic SSURGO database for model construction. The comparison result conducted for 20 most common tree species in eastern U.S. area showed that a general better model prediction capability was seen in RF models as to RMSE ranged between 0.157 and 0.399 as well as R_a^2 proposed by Kvalseth (1985) ranged between 0.483 and 0.936 on validation data. RF model provided a more reasonable prediction in average ADI of individual tree per plot at climate change A2 scenario that global warming effect exerted an optimistic effect on diameter growth of red maple in northern states.

Key words: Random Forests; multiple linear regression; annual diameter increment; future prediction

4.1 Introduction

Numerous studies have been conducted on global warming and potential change of temperature and precipitation point to a wide change of effects on the tree growth. As an important indicator of tree growth, individual tree diameter increment, and its subsequent use play tremendous roles in forest growth and yield simulators (Dixon 2002). Numerous studies have been done to evaluate possible effects of climate change on diameter increment (e.g. Du et al. 2007; Gebrekirstos et al. 2008; Kasson and Livingston 2012; Maxime and Hendrik 2011), further knowledge of improving model prediction accuracy is of growing importance.

A variety of modeling approaches have been developed and applied to predict diameter and basal area increment from a wide range of environmental and tree & stand related variables (e.g. Adame et al. 2008; Mailly et al. 2003; Ritchie and Hamann 2008); however most models were developed based on small datasets or limited to one species or a small geographic extent. National forest inventory originally designed to determine the extent, condition, volume, growth of trees on the Nation's forest land provide large amount of data for growth and yield model development. Although the sampling methodology was not specifically designed to develop growth and yield model, large sample could also compensate for the deficiency of measurement error, assuming that measurement errors are random (Monserud and Sterba 1996). Multiple diameter growth models on the basis of national forest inventory data (Adame et al. 2008; Andreassen and Tomter 2003; Lessard et al. 2001; Monserud and Sterba 1996; Pokharel and Froese 2008; Wykoff 1990) have been developed using linear or nonlinear regression techniques, which only allow a limited number of regressors remaining in the model. Thus the use of parametric regression analysis was challenged while attempting to put a large number of environment (e g. Climate and soils) factors into model.

As a classification and regression tree based nonparametric method, Random Forests has shown greater accuracy and a wider application range in ecological prediction than traditional parametric models (Prasad et al. 2006). It is robust to modeling challenges such as correlated predictors, nonlinear or non-monotonic relationships, and variable interactions, along with the ability to include either categorical or continuous variables (Prasad et al. 2006; Wang et al. 2005). The models are also robust to overfitting by nature of how predictor subsets are selected

by bootstrap sampling, even in extreme situations where the number of predictors exceeds the number of observations. The successful application of classification and regression based method has been shown in the ecological areas, including species habitat distribution (Prasad et al. 2006; Rehfeldt et al. 2008), site index (Crookston et al. 2010; Jiang et al. 2014; Weiskittel et al. 2011), mortality (Dobbertin and Biging 1998), but it has not been applied into diameter or basal area prediction.

The objective of this study was to develop diameter increment model of individual trees for 20 eastern tree species on the base of FIA data using RF. These models were intended for use in identifying the potential responses of eastern forests to climate change over a broad geographic range in eastern U.S. In order to demonstrate RF model performance, the multiple linear regression model (MLR) was considered as a baseline and was compared with the model based on RF. The following steps were taken in order to accomplish this goal: 1) developing multiple linear regression (MLR) and Random Forests (RF) models that related annual change in squared diameter (DDS) to tree-related, stand-related, climate and soils variables, 2) examining the prediction bias of the MLR and RF by making comparison of root mean squire error (RMSE), R_a^2 proposed by Kvalseth (1985) as well as quantiles of the distribution of residuals, 3) evaluating and ranking variable importance on the basis of MLR and RF models, 4) predicting how diameter increment may change in the eastern U.S. over the 21st century under climate change scenario A2 using two models (red maple was used as an example).

4.2 Data

Study area

The study area was defined to span the geographic ranges of most forest tree species growing in the eastern U.S., which we defined as those states east of the Great Plains physiographic province (Fenneman 1946). Data were compiled from 37 states that included areas east of 100° W longitude, comprising states of North and South Dakota, Nebraska, Kansas, Oklahoma, Texas, and all states further east.

Vegetation and climate & soils data

Vegetation data were compiled through the publicly-available online database of the USDA Forest Service, Forest Inventory and Analysis (FIA) national program[2]. Individual trees observed on 7.3 m (24 foot) radius subplots that were alive at the beginning of a growth interval and either dead or alive at the end of the interval were included. Only trees having diameters at breast height (1.37 m; DBH) ≥ 13.0cm were used in the analysis. Dead trees at the first measurement were removed from data analysis since it is assumed that the observed trees are alive at the beginning of the growth interval. The twenty species occurring most frequently in the FIA database were selected for diameter increment model development (Table 4.1). In addition to ensuring very large sample size, the twenty species' nominal ranges covered virtually all of the Eastern U.S.

Both tree-level and stand-level variables were included in diameter growth model as potential explanatory variables. At the tree level, diameter at breast height (PREVDIA) at the beginning of growing interval is used as the major predictor variable in the growth model for predicting annual diameter increment. Its subsequent forms involving log transformed DBH (LPREVDIA) and the square of PREVDIA (PREVDIA2) were used for better model fitting. Other individual tree variables include initial crown ratio and crown class. Crown ratio (CR) was the ratio of crown to total tree height. Crown class (CCLCD1-CCLCD5) was recorded in five ordered categories ranging from open-grown to suppressed. Crown ratio and crown class were measured at the end of the growth period and assumed not to have changed significantly during the past years. Plot-level variable included the number of trees per hectare (N), total basal area per hectare (BA), the percentage of stand basal area by species component (BASC, %), and quadratic mean diameter (QMD) at the beginning of the growth interval. Other factors deemed as potentially useful predictors of diameter increment are a suite of climate and soils variables (Table 3.1and Table 3.2) and stand information including longitude (LON) and latitude (LAT).

Logarithm of the annual increase in diameter squared ln (DDS) was set as dependent variable rather than the variable diameter increment since the correlation between DDS with initial diameter was greater than that for diameter growth with initial diameter (West 1980). The log

[2] Available online at http://apps.fs.fed.us/fiadb-downloads/datamart.html (last accessed on 22 March, 2015)

transformation was used not only to equalize the variance over the entire range of response values, but also the logarithm transformed DDS was more linearly related to indicators CR and PREVDIA. In order to make the training dataset more representative, the data for each species was equally divided into four groups according to the size of PREVDIA from smallest to the largest. Eighty percent of individual trees from each four groups was selected and combined together into training dataset. Within the remaining data, the individual trees from plots other than ones existing in the training data set were kept as test dataset. RF and MLR models were developed based on the same training dataset and validated based on the same test dataset for each species group.

4.3 Methods

4.3.1 Random Forests

Random Forests (RF ;Breiman 2001) is a classification and regression-tree-based ensemble machine learning technique capable of modeling complex relationships between many variables in large data sets and is able to account for nonlinear relationships and variable interactions in noisy data set (Prasad et al. 2006). Besides that, RF avoids overfitting and is robust to multicollinearity present in between-predictor relationships (Ziegler and König 2014). Diameter increment model for each species was developed with tree-related, stand- related, climate and soils variables as input.

Random Forests models were trained for predicting basal area increment using the "randomForest" algorithm implemented in R. According to past training experience on similar database, desirable modeling performance could be achieved when the number of trees for RF models was chosen at 500 and m_{try} was set as one third the total number of variables, which was also the default value of Random Forests controls.

4.3.2 Multiple linear regression models

In order to demonstrate RF model performance, the multiple linear regression model (MLR) is considered as a baseline and is compared with the model based on RF. The development of MLR for each species group involved using principal component analysis to identify two new

components to be included in the model, selecting best set of variables via stepwise regression, applying weighted least square adjustment for keeping constancy of error variance. MLR was built as formula [4.1], where DDS was calculated as equation [4.2].

Due to high correlation between climate variables involving the 21 climate variables, PCA was applied to project the predictor variables into new linearly uncorrelated components (Quinn et al., 2002). To carry out a principal component analysis (PCA) on a multivariate data set, the function "prcomp" implemented in R was used while setting "scale" and "mean" being true to standardize the variables. According to the tests from 20 most common species, the first two components were retained as new predictors for later analysis, because those components were able to account for at least 80 percent of the total variation in climate variables. Those two components, PC1 and PC2, along with tree -related variables were then used as predictors in the MLR. Correlation coefficient between principal component and each climate variable were computed by R package 'cor" using pearson method, which can be used to evaluate the ecological meaning of each component.

Assuming that ln (DDS) is linearly related with the above indicators and any possible combination of indicators was prepared for constructing multiple linear regressions. Using stepwise model selection in "stepAIC" function in R, the model formula with the smallest value of AIC was identified for each species. Since MLR is sensitive to multicollinearity, variance inflation factors (VIF's) were computed for each of the remaining coefficients to detect collinearity among the independent variables. The variables with variance inflation factor larger than 10 were removed among the tree & stand related variables as well as PC1 and PC2.Such procedure was conducted species by species.

$$\ln(DDS) = X\beta \qquad [4.1]$$

$$DDS = (DI + PREVDIA)^2 - PREVDIA^2 \qquad [4.2]$$

We found that the assumption of resulting MLR models was voilated for most speceis since heterogeneity of variance still appears even though DDS was logarithm transformed at the beginning. Weighted regression was thus applied to stabilize residual variance along the different value of ln (DDS).The unweighted basal area increment model was first fit to the data for a single speceis. Then regression analysis as in equation [4.3] was conducted on the resulting residuals.

$$|e_i| = \alpha_1 + \alpha_2 \ln(\widehat{DDS_i}) \qquad [4.3]$$

, where $|e_i|$is the absolute value of each residuale_i. The fitted values from an ordinary least square regression of $|e_i|$ against $\widehat{DDS_i}$ could be used to calculate the weight for each observation (Kutner 2005), where weight was the inverse of this square of this fitted value from equation.

4.3.3 Model validation and evaluation

Three well-known error statistics were calculated to measure the difference between the observed and predicted annual diameter increment. Because annual measurements were not available in FIA data set, diameter growth is assumed to be the same over the growth period. Thus ADI was calculated as the ratio of DBH difference between the two measurements and total number of years during the two measurement periods. For maintaining an adequate database, only the observations with ADI larger than 4 were deleted from data base for each species group. The predicted ADI was calculated as formula [4.4]. Because the residuals contain a number of large values which may influence the values of the common parametric statistics, such as root mean-squared error (RMSE, [4.5]), nonparametric statistics including 25th percentile, median, and 75th percentile for the distribution of residuals were adopted to examine the difference of ADI and predicted ADI. Instead of using the coefficient of determination R^2 statistics, an alternative goodness- of- fit statistics recommended by Kvalseth (1985), R_a^2 [4.6] for its robustness to outliers in the diameter increment model was calculated for each species group. As a measure of the proportion of the total variability explained by the fitted model, a higher value of R_a^2 indicates a better model fitting within the range between 0 and 1.

$$\widehat{ADI} = \sqrt{(PREVDIA)^2 + \widehat{DDS}} - PREVDIA \quad [4.4]$$

$$RMSE = \left(\sum_{i=1}^{n} (\widehat{ADI} - DI)^2 / n \right)^{1/2} \quad [4.5]$$

$$R_a^2 = 1 - [\frac{Med(|ADI - \widehat{ADI}|)}{Med(|ADI - \widehat{ADI}|)}]^2 \quad [4.6]$$

4.3.4 Variable importance measure

For RF model, the importance of variable was ranked according to the permutation accuracy importance measures. Permutation accuracy importance was used to assess which variables were most meaningful in predicting basal area increment. The "randomForest" package reports this measure as "%incMSE," the percentage increase in OOB MSE computed for nonpermuted versus randomly permuted predictors (Genuer et al. 2010). The measure of variable importance known as node impurity was not used here because of its stronger sensitivity to within-predictor correlation and differences in category frequencies as well as less stability than those using permutation accuracy (Nicodemus 2011). The t statistic in MLR is the coefficient on related variable divided by its standard error, which is commonly used to show how strongly each independent variable is associated with the dependent variable. A larger absolute value of t statistics means a bigger effect of the independent variable on the dependent variable. The

variable with a larger absolute value of t statistics among all variables was deemed as a more important variable, and then ranked accordingly.

4.3.5 Future ADI prediction based on RF model

The resulting MLR and RF models for each species group were applied to the ADI predictions for future climate change A2 scenario at eastern sites to demonstrate the possible changes in ADI over space and time. Future condition of ADI of individual trees for each FIA plot across eastern U.S. were predicted respectively using the MLR and RF models with future climate conditions in the 2090s specified as inputs. Soil properties and tree conditions, were assumed to remain unchanged at a particular plot location from contemporary to future conditions, so those data were used as inputs without alteration between contemporary and future scenarios. For each IPCC development scenario, projections based on inputs from three GCMs were averaged to account for differences between different GCM predictions run under the same A2 scenario. To simplify reporting, only red maple was chosen as an example species for further investigation. The predicted ADI of trees within a plot were averaged to reflect an overall level of diameter increment, where predicted ADI was calculated from the predicted DDS using equation [4.4]. The difference of average predicted ADI at a plot between future and contemporary period was serve to reflect how future climate changing influences on tree diameter growth.

4.4 Results

4.4.1 Model validation

Scatterplots of studentized residuals from ordinary and weighted least squares regression versus the predicted DDS were plotted in Figure 4.1. The slope of red trend line was closer to zero after weight transformation, which meant that the residuals were less related to the fitted ln (DDS) so that the assumption of constant variance was met.

The tree- level residuals of ADI for the validation data sets were listed in

Table 4.2. The median of residuals for the validation data were generally less than 0.04 cm/year with negative signs for RF and MLR. General negative values of the median residuals for 20 species suggested that both models slightly underpredicted annual diameter growth; however, the magnitude of the underprediction was small. Although the median residuals of RF models were more deviated away from zero than that of MLR for many species, the range of half of residuals derived by RF models were narrower that that by MLR models based on the interquantile reading of residuals. A general better model fitting capability of RF models than MLR models was also demonstrated by model criteria RMSE and R_a^2. R_a^2 computed by RF models was higher than it was achieved by MLR models for all species except for shortleaf pine with a slight better fitting predicted by MLR. The value of RMSE based on RF models was mostly smaller than that based on MLR models, except sweetgum, balsam fir and shortleaf pine.

4.4.2 Coefficient estimation and variable importance evaluation

Variables whose corresponded parameter estimates were not statistically different from zero were shown in Table 4.3. The result from 20 MLR species models showed that the sign of parameters for N, BA, intermediate and overtopped crown level in CCLCD were generally negative, in contrast with the sign of LPREDIA, CR and TAWC commonly being positive, which indicated that a crowd and dense forest stand will hold back the growth of tree diameter, especially for suppressed trees. It also suggested that the bigger of tree diameter was, the higher proportion of soil water capability was, the larger proportion of crown relative to tree height was, the faster trees grew with respect of diameter.

Variable importance ranking result (Table 4.3) from MLR showed that CR as well as diameter transformed variables PREVDIA2 and LPREVIDA were noted as the most important variables for most tree species models. According to rank of variables, we concluded that tree-level variables were the most important variables for all species, which were followed by climate and soils indicators sequentially for most species, except for yellow-poplar, balsam fir, sweetgum and paper birch.

PCIPcom1 was generally more highly correlated with SMI and temperature related variables including MAT,MTCM,MMIN,MTWM,MMAX,SDAY,FDAY,FFP,DD5,GSDD5,D100, DD0,

MMINDD0 and TDIFF(Table 4.4). It was shown that no matter with any species, the group including MAT, MTCM, MMIN, MTWM, MMAX, FDAY, FFP, DD5, GSDD5 and SMI as well as the group containing SDAY, D100, DD0, MMINDD0 and TDIFF always had contrast relationship with PCIPcom1. Specifically, if PCIPcom1was positively correlated with the former group then negatively related to the latter group, and *vice versa*. PCIPcom2 was generally more correlated with one or more precipitation related variables depending on different species (Table 4.5**)**. The sign of parameter of principal components (Table 4.3) and the sign of correlation coefficient between principal components and individual climate indicators (Table 4.4 and Table 4.5) worked together to explain whether climate indicators have positive or negative influences on tree diameter growth. Take loblolly pine for example, decreasing value of MAT and increasing value of TDIFF will lead to an increase of tree diameter growth. The climate variables with absolute values of correlation coefficient larger than 0.6 were assumed to be highly correlated with PCIPcom1 or PCIPcom2.

The most important ten variables concluded from RF models (Table 4.6) for each species showed that the importance of tree and stand related variables were more substantial than soils and climate variables. Crown ratio and CCLCD were found to be among the most two important climate predictors in most species. The relative rank location of climate and soils variables differed species by species. Compared with temperature related variables, which seldom appear in the top 10 important variables, precipitation-related variables such as MAP, PRATIO, GSP and SMRSPRPB were more influential in effecting diameter growth. Even though the diameter increment of loblolly pine, eastern redcedar and red maple tree species were not directly related to precipitation information, the rank of TAWC indicated that tree diameter growth was highly influenced by total available water capacity in soils, which was also a reflection of the levels of physiological satisfaction of water intake.

4.4.3 Mapping average ADI for red maple

Contemporary condition concluded from existing FIA records for red maple showed that the average ADI was frequently higher (green color) in the south area and Southern Wisconsin (Figure 4. 2a). A lower value (red color) was seen at most areas of New England, northern Michigan, Minnesota, and Mid-Atlantic region (Figure 4. 2a). The resulting prediction under

climate change A2 scenario based on RF showed that average ADI in the 2090s was above 0.25 cm/year at most areas of eastern U.S. (Figure 4. 2b). It was noted that the biggest value of the increase in average annual diameter increment over time was mainly located in Northern Michigan, Minnesota (Figure 4. 2c), where the increase was above 0.1cm/year. The spatial plot of future prediction of average ADI based on MLR (Figure 4. 2d) showed that lower diameter growth was located in the Appalachian Mountains and northern lake states with average ADI less than 0.2cm/year. When this future prediction is compared with contemporary conditions, we found that the decrease in average annual diameter increment over time had been dominant in all eastern states (Figure 4. 2e).

According to the rank of variable importance based on RF, climate variables MAP, PRATIO, and GSP were all influential in effecting the change of ADI. It was shown that the difference of ADI was negatively related with the difference of PRATIO; but was positively correlated with the difference of MAP before a point at which the difference of MAP was zero and decreasing afterward, which was also same with the relationship between the difference of ADI and the difference of GSP (Figure 4.3). Those trends indicated that when the ratio of summer precipitation to total precipitation was increased in future, average ADI of red maple tends to decrease from contemporary to the 2090s. It was also showed that the closer to contemporary level the mean annual precipitation or growing season precipitation in the 2090s was, the higher annual diameter increment could be achieved in the 2090s.

4.5 Discussion

For meeting regession assumption puspose, the DDS variable was log transformed to equalize the variance along with increasing value of response (Strimbu 2012). Systematic underestimation of basal area increment was introduced whenever the antilogarithm used to convert log-normality distributed estimates back to original units (Baskerville 1972; Flewelling and Pienaar 1981), which thus is the reason why the residual median for the validation data were generally negative for most species. A widely used statistical tool, the logarithmic correction factor recommended by Baskerville (1972), was attempted to counteract the systematic bias (Strimbu 2012); however, this bias correction was not applied eventually since its application led to large overprediction of DDS in this study. Basal area difference between two periods was

set as the response variable. Since individual tree diameter increment plays a tremendous role in forest growth and yield simulators, in general predicted basal area was converted to DBH in subsequent equations, which inevitably brought bias in DBH estimation.

Due to the nature of differences between parametric and nonparametric statistics, it was not surprising to observe that performance of R_a^2 was not always consistent with the performance of RMSE for each species. For the situation of a good performance of R_a^2 in contrast of a poor performance of RMSE at a species level, it was likely because a number of large values of response variables existing in data set so that the mean and median of the values of residuals differed greatly. In order to maintain a large amount of information in the training data set, we only disregard a few outliers in some species group. Thus those large values were still kept in the dataset. The result of R_a^2 and RMSE with sweetgum and paper birch showed that R_a^2 derived by RF was higher than that acquired by MLR, but parametric statistics RMSE was on the contrary. It is likely that a poor prediction of RF models happened at several extreme large value of response variables in test data, which were out the range of observed response variable in training data. Those bad predictions for extreme value of data point will not affect the evaluation of R_a^2 but RMSE.

It was shown from Table 4.2 that the performance of RF model was generally better than MLR under study. The superior capabilities of RF models have been demonstrated by many ecological studies either for classification (Rehfeldt et al. 2006; Rehfeldt et al. 2008) or regression purpose (Jiang et al. 2014; Prasad et al. 2006; Rehfeldt et al. 2006; Sabatia and Burkhart 2014; Weiskittel et al. 2011). In spite of RF models being given more and more attention in ecological field due to its advantages in dealing complex relationship among environmental variables, it has not been applied in the diameter growth field in our knowledge. Therefore it is wise to compare the relative performance of RF models with other methods in order to prove the feasibility of RF models in large-scale prediction of tree diameter growth. The R_a^2 computed in data validation by RF models seem desirable than the diameter increment models proposed by Lessard et al. (2001), which incorporated a gamma function using DBH as the independent variables and a modifier to adjusting the predicted growth values for individual tree and stand conditions. R_a^2 achieved by that study was generally above 0.3 on residual analysis of validation data based on Minnesota

FIA data validation. But R_a^2 of RF models under study achieved a better result, which was ranged from 0.483 to 0.936. The ratio between RMSE and the mean diameter increment both on the base of validation data was ranged between around 0.2 to 0.7 depending on different species.Simlar value of the ratio could be found in residuals analysis of nonlinear mixed model for Slash Pine in Florida (Timilsina and Staudhammer 2013); or better result found in coppices oak located in northwest Spain based on mixed model (Adame et al. 2008). The reason why our result as to RMSE seems not as desriable as some related studies becaues large values were not deleted from the data set thus resulting in a higher value of RMSE. The common value of mean annual diameter growth is about 0.25cm/year (e.g. Pokharel and Froese 2008), however we only treated data points with the value of annual diameter incrment larger than 4 cm/year as abnormal for seeking a more complete dataset for each speceis.

The main idea of variable importance measures of MLR and RF models were both based on the fact that how much of RMSE was increased when a specific variable was added into model while other rest variables already existing in the model (Genuer et al. 2010; Kutner 2005). The agreement on top important variables between two models differed species by species (Table 4.6). Species such as quaking aspen had the most similarity on variable importance, where CR, CCLCD, BASC, NO10, CLAY, BA, KFFACT and GSP were among most ten important variables. Species such as balsam fir has the least agreement on evaluating variable importance, where CR, LON, BA and NO200 were among the most ten important variables. It was difficult to decide which method was wiser to decide the most influential variables since they both had drawback in ranking variable importance. In order to avoid multicollinerity among independent variables in MLR, variables sharing duplicated information were deleted before model development, even though the variables were still useful in explaining the variance of response variables (Cutler et al. 2007). Thus only selected explanatory variables with less correlation were put into variable importance analysis, even though the excluded variables may also be equally important (Nicodemus et al., 2010). Although RF predictions are known to be relatively unaffected when correlations exist between predictors, using permutation accuracy as a measure of variable importance has been shown to give preference to correlated predictor variables and categorical predictors having large numbers of classes (Strobl et al. 2008). Take Balsam fir for example, the variables NO10 and pH were not statistical significantly in effecting basal area

growth in MLR models but they were belonging to the most ten important variables in RF models. It is likely because they were highly correlated with another influential predictor variable and equally well suited for splitting as the truly influential predictor variable even though they are not related with the response. It may also answer the reason why catigorical variable CCLCD was detected to not be siginificant in MLR models but be influential in RF model (Nicodemus et al. 2010; Strobl et al. 2007). The conditional variable importance measure proposed by Strobl et al. (2008) is preferable for use in ranking the importance of predictors in RF models; however, its use here was impractical due to the computational requirements of its implementation in the R package "*party*" (Strobl et al. 2009).

Judging from the difference of spatial map on average ADI for red maple based on MLR and RF model (Figure 4. 2 b and d), MLR model led to a lower value of future prediction on ADI of individual tree than RF model. The predicted change of average ADI from contemporary to the 2090s was distinct between two models. The most outstanding difference was at northern states of U.S., since RF model predicted an increase of average ADI over time at these areas but MLR had a contrary result. A wealth of studies have modeled the impact of climate change on forest growth. The results vary by forest region, climate scenario, and modeling approach (e.g. Battles et al. 2008; Duchesne et al. 2012; Savva et al. 2006). Although few studies have examined potential mean diameter increment in the eastern US so comparisons with other studies are somewhat limited, recent studies on potential habitat (Iverson et al. 2008; Joyce and Rehfeldt 2013)and site index (Jiang et al. 2014) have indicated that northern area was more appealing than southern area for tree growth under the influence of global warming. Thus as to the future prediction result of average ADI for red maple, RF model creates more consistent result with past studies than MLR model. Climate- diameter growth relationship has been demonstrated only for three most important variables (Figure 4.3). The prediction induced from RF model showed that a suitable amount of rainfall with balanced distribution through the year had an excellent influence on diameter growth of red maple. A positive sign on PCIPcom1 in MLR model (Table 4.3) and a negative relationship between PCIPcom1 and PRATIO (Table 4.4) showed that PRATIO had an adverse effect on diameter growth of red maple. The interpretation of PRATIO based on MLR was same with what was found from RF model. Since the trend of correlation

relationship between ADI and GSP or ADI and MAP was nonlinear (Figure 4.3), so it was not appropriate to make comparison with result concluded from MLR.

4.6 Summary

Nonparametric RF and parametric MLR model were compared and evaluated for predicting diameter increment of individual tree for 20 species in eastern U.S. area. Unlike MLR, the development of which included variable transformation, PCA data dimension reduction, variable selection and weight transformation, RF was free of data preparation or any model assumption requirement, that was easy to build and implement in "randomForest" of R environment. Both models may have system bias in underpredicting ADI of individual trees but the magnitude was slight in general. RF also gained advantages in predicting capability no matter in model validation or creating a more reasonable future projection for eastern forests. The agreement on top important variables between two models differed species by species, but they both agreed that tree-level variables had more influential roles than climate and soils factors. The issue of heavy computational requirement in conditional variable importance measure in big data set, however, leads to the application of Random Forests in forest growth simulation being more prone to prediction purpose not to variable importance interpretation.

4.7 Reference

Adame P, Hynynen J, Cañellas I, del Río M, 2008. Individual-tree diameter growth model for rebollo oak (Quercus pyrenaica Willd.) coppices. Forest Ecology and Management 255(3–4):1011-1022.

Andreassen K, Tomter SM, 2003. Basal area growth models for individual trees of Norway spruce, Scots pine, birch and other broadleaves in Norway. Forest Ecology and Management 180(1–3):11-24.

Baskerville GL, 1972. Use of logarithmic regression in the estimation of plant biomass. Canadian Journal of Forest Research 2(1):49-53.

Battles J, Robards T, Das A, Waring K, Gilless JK, Biging G, Schurr F, 2008. Climate change impacts on forest growth and tree mortality: a data-driven modeling study in the mixed-conifer forest of the Sierra Nevada, California. Climatic Change 87(1):193-213.

Breiman L, 2001. Random Forests. Machine Learning 45(1):5-32.

Crookston NL, Rehfeldt GE, Dixon GE, Weiskittel AR, 2010. Addressing climate change in the forest vegetation simulator to assess impacts on landscape forest dynamics. Forest Ecology and Management 260(7):1198-1211.

Cutler DR, Edwards TC, Jr., Beard KH, Cutler A, Hess KT, Gibson J, Lawler JJ, 2007. Random forests for classification in ecology. Ecology 88(11):2783-2792.

Dixon GE, 2002. Essential FVS: a user's guide to the forest vegetation simulator. In: Internal Rep. Fort Collins, CO: USDA Forest Service, Forest Management Service Center. 220.

Dobbertin M, Biging GS, 1998. Using the non-parametric classifier cart to model forest tree mortality. Forest Science 44(4):507-516.

Du S, Yamanaka N, Yamamoto F, Otsuki K, 2007. The effect of climate on radial growth of Quercus liaotungensis forest trees in Loess Plateau, China. Dendrochronologia (Verona) 25(1):29-36.

Duchesne L, Houle D, D'Orangeville L, 2012. Influence of climate on seasonal patterns of stem increment of balsam fir in a boreal forest of Québec, Canada. Agricultural and Forest Meteorology 162–163(0):108-114.

Fenneman NM, and Johnson, D.W., 1946. Physiographic divisions of the conterminous United States Washington, D.C.: U.S. Geological Survey (USGS). Special map series, scale 1:7,000,000.

Flewelling JW, Pienaar LV, 1981. Multiplicative Regression with Lognormal Errors. Forest Science 27(2):281-289.

Gebrekirstos A, Mitlöhner R, Teketay D, Worbes M, 2008. Climate–growth relationships of the dominant tree species from semi-arid savanna woodland in Ethiopia. Trees 22(5):631-641.

Genuer R, Poggi J-M, and Tuleau-Malot C, 2010. Variable selection using random forests. Pattern Recognition Letters 31(14):2225-2236.

Iverson LR, Prasad AM, Matthews SN, Peters M, 2008. Estimating potential habitat for 134 eastern US tree species under six climate scenarios. Forest Ecology and Management 254(3):390-406.

Jiang H, Radtke PJ, Weiskittel AR, Coulston JW, Guertin PJ, 2014. Climate- and soil-based models of site productivity in eastern US tree species. Canadian Journal of Forest Research 45(3):325-342.

Joyce DG, Rehfeldt GE, 2013. Climatic niche, ecological genetics, and impact of climate change on eastern white pine (Pinus strobus L.): Guidelines for land managers. Forest Ecology and Management 295:173-192.

Kasson MT, Livingston WH, 2012. Relationships among beech bark disease, climate, radial growth response and mortality of American beech in northern Maine, USA. Forest Pathology 42(3):199-212.

Kutner MH. 2005. Applied linear statistical models McGraw-Hill Irwin, Boston. 355.

Kvalseth TO, 1985. Cautionary Note about R2. The American Statistician 39(4):279-285.

Lessard VC, McRoberts RE, Holdaway MR, 2001. Diameter Growth Models Using Minnesota Forest Inventory and Analysis Data. Forest Science 47(3):301-310.

Mailly D, Turbis S, Pothier D, 2003. Predicting basal area increment in a spatially explicit, individual tree model: a test of competition measures with black spruce. Canadian Journal of Forest Research 33(3):435-443.

Maxime C, Hendrik D, 2011. Effects of climate on diameter growth of co-occurring Fagus sylvatica and Abies alba along an altitudinal gradient. Trees 25(2):265-276.

Monserud RA, Sterba H, 1996. A basal area increment model for individual trees growing in even- and uneven-aged forest stands in Austria. Forest Ecology and Management 80(1–3):57-80.

Nicodemus KK, 2011. Letter to the Editor: On the stability and ranking of predictors from random forest variable importance measures. Briefings in Bioinformatics 12(4):369-373.

Nicodemus KK, Malley JD, Strobl C, Ziegler A, 2010. The behaviour of random forest permutation-based variable importance measures under predictor correlation. BMC Bioinformatics 11:110.

Pokharel B, Froese RE, 2008. Evaluating alternative implementations of the Lake States FVS diameter increment model. Forest Ecology and Management 255(5–6):1759-1771.

Prasad AM, Iverson LR, Liaw A, 2006. Newer classification and regression tree techniques: bagging and Random Forests for ecological prediction. Ecosystems 9(2):181-199.

Rehfeldt GE, Crookston NL, Warwell MV, Evans JS, 2006. Empirical analyses of plant‐climate relationships for the western United States. International Journal of Plant Sciences 167(6):1123-1150.

Rehfeldt GE, Ferguson DE, Crookston N, 2008. Quantifying the Abundance of Co-occurring Conifers Along Inland Northwest (USA) Climate Gradients. Ecology [H.W. Wilson - GS] 89(8):2127.

Ritchie M, Hamann J, 2008. Individual-tree height-, diameter- and crown-width increment equations for young Douglas-fir plantations. New Forests 35(2):173-186.

Sabatia CO, Burkhart HE, 2014. Predicting site index of plantation loblolly pine from biophysical variables. Forest Ecology and Management 326(0):142-156.

Savva Y, Oleksyn J, Reich P, Tjoelker M, Vaganov E, Modrzynski J, 2006. Interannual growth response of Norway spruce to climate along an altitudinal gradient in the Tatra Mountains, Poland. Trees 20(6):735-746.

Strimbu B, 2012. Correction for bias of models with lognormal distributed variables in absence of original data. Annals of Forest Research 55(2):265-279.

Strobl C, Boulesteix A-L, Kneib T, Augustin T, Zeileis A, 2008. Conditional variable importance for random forests. Bmc Bioinformatics 9:307.

Strobl C, Boulesteix A-L, Zeileis A, Hothorn T, 2007. Bias in random forest variable importance measures: Illustrations, sources and a solution. Bmc Bioinformatics 8:25.

Strobl C, Hothorn T, Zeileis A, 2009. Party on! The R Journal 1(2):14-17.

Timilsina N, Staudhammer CL, 2013. Individual Tree-Based Diameter Growth Model of Slash Pine in Florida Using Nonlinear Mixed Modeling. Forest Science 59(1):27-37.

Wang Y, Raulier F, Ung C-H, 2005. Evaluation of spatial predictions of site index obtained by parametric and nonparametric methods—A case study of lodgepole pine productivity. Forest Ecology and Management 214(1–3):201-211.

Weiskittel AR, Crookston NL, Radtke PJ, 2011. Linking climate, gross primary productivity, and site index across forests of the western United States. Canadian Journal of Forest Research 41(8):1710-1721.

West PW, 1980. Use of diameter increment and basal area increment in tree growth studies. Canadian Journal of Forest Research 10(1):71-77.

Wykoff WR, 1990. A Basal Area Increment Model for Individual Conifers in the Northern Rocky Mountains. Forest Science 36(4):1077-1104.

Ziegler A, König IR, 2014. Mining data with random forests: current options for real-world applications. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 4(1):55-63.

Table 4.1 Summary statistics of ADI for 20 most frequently occurring eastern species in the FIA database

Scientific name	Common name	ADI (cm/year)							
		Training Data				Validation Data			
		obs	mean	max	min	obs	mean	max	min
Loblolly pine	Pinus taeda	77009	0.76	3.98	0.02	446	1.01	3.46	0.05
Red maple	Acer rubrum	59482	0.30	3.63	0.03	1533	0.34	3.30	0.04
Sugar maple	Acer saccharum	35566	0.27	3.99	0.03	672	0.34	1.89	0.04
white oak	Quercus alba	27605	0.31	3.99	0.02	945	0.40	2.26	0.05
Quaking	Populus tremuloides	22767	0.38	2.54	0.04	400	0.47	1.40	0.04
Sweetgum	Liquidambar styraciflua	20417	0.37	3.63	0.02	792	0.45	2.50	0.03
N.white-cedar	Thuja occidentalis	18908	0.19	2.54	0.04	101	0.28	1.47	0.05
Northern red oak	Quercus rubra	18099	0.41	3.99	0.03	902	0.49	1.75	0.04
Yellow-poplar	Liriodendron tulipifera	16870	0.54	3.75	0.03	762	0.64	2.86	0.04
Balsam fir	Abies balsamea	14975	0.31	2.08	0.04	385	0.38	1.62	0.05
Red pine	Pinus resinosa	14883	0.35	2.10	0.04	96	0.53	1.84	0.05
E. white pine	Pinus strobus	13944	0.45	3.18	0.04	377	0.63	3.30	0.04
Chestnut oak	Quercus prinus	12666	0.30	3.87	0.03	262	0.38	1.78	0.03
Eastern hemlock	Tsuga canadensis	12233	0.31	3.27	0.04	307	0.40	1.59	0.04
Black oak	Quercus velutina	11548	0.37	3.89	0.03	618	0.42	2.11	0.04
Black cherry	Prunus serotina	11430	0.38	2.95	0.03	873	0.37	3.26	0.03
Shortleaf pine	Pinus echinata	10380	0.36	2.93	0.02	372	0.39	1.61	0.03
White ash	Fraxinus americana	9803	0.35	3.06	0.03	648	0.38	2.59	0.03
Paper birch	Betula papyrifera	9631	0.19	1.94	0.04	487	0.22	1.23	0.04
Eastern redcedar	Juniperus virginiana	9411	0.29	3.03	0.03	346	0.39	2.54	0.04

Table 4.2 Residual statistics for annual diameter increment model fit to FIA validation tree-level data

	RF					MLR				
Species	R_a^2	RMSE	25th percentile	Median	75th percentile	R_a^2	RMSE	25th percentile	Median	75th percentile
Loblolly pine	0.781	0.298	-0.204	-0.031	0.117	0.619	0.379	-0.255	-0.023	0.168
Red maple	0.928	0.153	-0.083	-0.011	0.030	0.697	0.225	-0.143	-0.012	0.080
Sugar maple	0.798	0.182	-0.114	-0.015	0.056	0.716	0.190	-0.123	-0.014	0.068
white oak	0.834	0.186	-0.114	-0.018	0.060	0.769	0.205	-0.131	-0.016	0.737
Quaking aspen	0.714	0.177	-0.125	-0.017	0.072	0.596	0.195	-0.148	-0.020	0.084
Sweetgum	0.744	0.264	-0.170	-0.018	0.082	0.687	0.288	-0.184	-0.017	0.093
N.white-cedar	0.790	0.102	-0.064	-0.007	0.038	0.717	0.111	-0.073	-0.008	0.045
Northern red oak	0.842	0.206	-0.133	-0.015	0.074	0.786	0.383	-0.161	-0.018	0.091
Yellow-poplar	0.799	0.328	-0.212	-0.030	0.100	0.721	0.380	-0.240	-0.023	0.132
Balsam fir	0.695	0.154	-0.115	-0.019	0.059	0.641	0.164	-0.121	-0.016	0.065
Red pine	0.791	0.147	-0.090	-0.015	0.052	0.565	0.183	-0.127	-0.015	0.082
E. white pine	0.898	0.243	-0.161	-0.026	0.067	0.872	0.265	-0.179	-0.021	0.083
Chestnut oak	0.836	0.198	-0.112	-0.013	0.062	0.791	0.207	-0.116	-0.009	0.068
Eastern hemlock	0.803	0.177	-0.108	-0.013	0.060	0.749	0.187	-0.116	-0.008	0.076
Black oak	0.795	0.204	-0.140	-0.020	0.069	0.733	0.392	-0.150	-0.016	0.085
Black cherry	0.773	0.259	-0.178	-0.024	0.079	0.678	0.284	-0.205	-0.021	0.095
Shortleaf pine	0.717	0.192	-0.128	-0.019	0.068	0.603	0.216	-0.150	-0.019	0.083
White ash	0.795	0.217	-0.017	0.068	0.634	0.638	0.289	-0.175	-0.018	0.104
Paper birch	0.621	0.123	-0.086	-0.009	0.048	0.550	0.130	-0.090	-0.006	0.052
Eastern redcedar	0.743	0.187	-0.114	-0.016	0.060	0.651	0.205	-0.126	-0.014	0.070

Note: RMSE and percentiles are defined in units of cm/year.

Table 4.3 Coefficient estimation and importance rank of variables for 20 species models

Indicators	Loblolly pine	Red maple	Sugar maple	White oak	Quaking aspen	Sweetgum	N. white-cedar	N. red oak	Yellow poplar	Balsam fir
Intercept.	-4.2960	-2.5720	2.8860	1.6890	1.6963	-2.8510	-1.1160	3.3650	2.6980	1.4150
N	-0.0004	-0.0010 [4]	-0.0021 [3]	-0.0020 [6]			-0.0005 [3]	-0.0022 [4]		
BA	-0.0321 [3]	-0.0123 [6]		-0.0085	-0.0129 [5]	-0.0280 [3]		0.0072 [10]	-0.0248 [3]	-0.0258 [4]
LPREVDIA	0.9262 [2]	1.0250 [2]	1.1120 [2]			0.6339 [2]	1.2040 [2]			
PREVDIA				0.0730[1]	0.0376 [2]	0.0002				
PREVDIA2	-0.0001 [8]	0.0001 [9]	0.0000	-0.0005 [4]		0.0002 [4]		0.0004 [1]	0.0003 [1]	0.0008 [2]
BASC		-0.1048 [8]	0.0556	-0.3096 [5]	0.0507 [10]	-0.0822	-0.2763 [4]	-0.5633 [2]	-0.0664 [10]	0.1173 [7]
QMD	-0.0185		-0.0062 [8]	-0.0154 [8]		0.0341				0.0160 [5]
PCIPcom1	-0.0151[10]	0.0309 [3]	-0.0215 [7]	0.0071 [2]	-0.0045		-0.0054	-0.0045	0.0107 [9]	0.0032
PCIPcom2	0.0423 [5]	-0.0101	-0.0513 [5]	-0.0456	0.0294 [3]	0.0341 [8]	-0.0297 [5]	-0.0199 [8]		
CR	0.0238 [1]	0.0186 [1]	0.0192 [1]	0.0102[3]	0.0242 [1]	0.0177 [1]	0.0132 [1]	0.0102 [3]	0.0203 [2]	0.0193 [1]
LAT	0.0554		-0.0757 [4]	0.0102		0.0264 [7]	-0.0239			
LON	-0.0299 [4]	-0.0071 [5]	0.0086 [6]	0.0056 [9]	0.0127 [4]	-0.0138 [9]	-0.0068 [6]	0.0134 [5]	0.0140 [4]	0.0148 [3]
pH		0.0310 [10]			0.0206	0.0228				
TAWC	1.2320 [7]			1.6550 [7]			-0.7954 [9]			0.9458 [6]
SBD	0.0975		0.1323 [10]		0.1078	0.1620	0.1992	0.1992 [9]	0.3930 [6]	
CLAY	-0.0029		-0.0035 [9]		-0.0034 [6]			-0.0018		
KFFACT		0.1611	-0.1201	-0.5102	0.3505 [9]	-0.2782	-0.2984			
OM	0.0049		-0.0054			0.0271	0.0024		-0.0061	-0.0048 [8]
KSAT	-0.0004	0.0006	0.0006			0.0008	-0.0014 [7]			
SAND						0.0060 [6]				
NO200	0.0022		0.0008			0.0037 [10]		0.0045 [6]	-0.0013	-0.0013 [9]
SILT	0.0018	0.0028 [7]		0.0021		0.0082 [5]	0.0022 [10]		0.0039 [8]	-0.0013 [10]
NO10	0.0025	0.0008 [9]			0.0023 [8]		-0.0022 [8]		0.0041 [7]	
CCLCD2	0.1104	0.5759								
CCLCD3	0.0879	0.4488								
CCLCD4	-0.5476			-0.4266	-0.5158	-0.6156		-0.6116	-0.9918	
CCLCD5	-0.9475		-0.4296	-0.7776 [10]	-0.6573 [7]	-0.8517		-1.0950 [7]	-1.4490 [5]	

Note: The rank for the first ten important variables was labeled at upper right.

Table 4.3, continued on next page

Indicators	Red Pine	E. white pine	Chestnut oak	Eastern hemlock	Black oak	Black cherry	Shortleaf pine	White ash	Paper birch	Eastern redcedar
Intercept	1.5590	4.1920	-2.6470	-0.1586	2.7440	-0.7343	0.8384	3.6160	0.2884	-1.6320
N		-0.0005	-0.0021 [5]	-0.0005 [9]	-0.0015 [3]	0.0008	-0.0011 [7]	-0.0015 [3]	-0.0007 [4]	-0.0007 [10]
BA	-0.0236 [4]	-0.0212 [7]		-0.0193 [3]		-0.0358 [5]	-0.0188 [10]		-0.0081	-0.0306 [4]
LPREVDIA		1.1580 [2]	1.2960 [1]	1.0350 [1]		0.7703 [2]			0.4868 [2]	0.9912 [2]
PREVDIA										
PREVDIA2	0.0003 [2]				0.0004 [1]	0.0001 [8]	0.0004 [1]	0.0005 [1]	0.0003 [9]	
BASC		-0.0990	-0.3520 [2]	-0.1155	-0.0735 [9]	0.0693		-0.1792 [6]	-0.2063 [6]	-0.0758
QMD	0.0083 [9]	-0.0098	-0.0173 [8]		-0.0041 [7]	0.0142 [6]	-0.0160 [8]	0.0038		
PCIPcom1	-0.0289 [3]	0.0479 [4]	0.0147 [6]	0.0120 [4]		0.0228 [3]	0.0158 [6]	-0.0505 [5]	-0.0115 [8]	0.0280 [3]
PCIPcom2	-0.0119 [8]	-0.0177	-0.0315 [4]	0.0132	-0.0268 [5]	0.0196 [9]	-0.0349 [4]		-0.0297 [7]	-0.0172
CR	0.0215 [1]	0.0242 [1]	0.0073 [3]	0.0184 [2]	0.0149 [2]	0.0202 [1]	0.0205 [2]	0.0229 [2]	0.0175 [1]	0.0145 [1]
LAT		-0.0996 [3]						-0.0664 [4]		
LON		0.0129 [5]	-0.0152 [7]	0.0055 [8]	0.0138 [4]		-0.0141 [5]		0.0052	-0.0044
pH	0.0457		0.0328	-0.0464 [10]	0.0275	0.0337	-0.0383			-0.0543 [6]
TAWC	1.2860 [7]	1.1320 [10]	-0.7316	1.0290 [7]	0.6876	1.5410 [7]	1.2240			1.5130 [8]
SBD	-0.3502 [5]					0.1598			0.2115 [10]	0.2709 [7]
CLAY	-0.0049 [10]	-0.0020	0.0041 [9]			-0.0031	-0.0085 [3]	-0.0048 [7]		
KFFACT		-0.2338		-0.3454						-0.4258
OM		-0.0321 [9]	0.0257		-0.0085	-0.0201		-0.0232 [8]		-0.0342
KSAT		-0.0005		-0.0013 [6]				0.0015 [10]		0.0008
SAND				0.0024 [5]		-0.0023		-0.0016		
NO200					0.0028 [8]			0.0030 [9]	0.0053 [5]	-0.0016
SILT	-0.0020				0.0025 [6]				0.0057 [3]	0.0024
NO10		0.0033 [8]	0.0019 [10]	0.0016		-0.0042 [4]	0.0037 [9]		0.0020	0.0034 [9]
CCLCD2					0.6561 [10]			1.0440		
CCLCD3				-0.5745				0.9009		
CCLCD4	-0.3780	-0.6413		-0.7368		-0.4509			-0.7580	-0.4007
CCLCD5	-0.7450 [6]	-0.9628 [6]	-0.5775	-1.0020	0.2482	-0.7778 [10]	-0.9346		-0.8893	-0.6647 [5]

Note: The rank for the first ten important variables was labeled at upper right.

Table 4.4 The correlation coefficient matrix for measuring PCIPcom1 and climate variables

	Loblolly pine	Red maple	Sugar maple	White oak	Quaking aspen	Sweetgum	N. white-cedar	N. red oak	Yellow poplar	Balsam fir
MAT	0.994	0.998	0.995	0.996	0.979	-0.994	-0.970	0.994	-0.996	0.966
MTCM	0.979	0.983	0.962	0.966	0.824	-0.976	-0.835	0.957	-0.964	0.718
MMIN	0.981	0.972	0.947	0.951	0.824	-0.979	-0.827	0.938	-0.973	0.697
MTWM	0.899	0.947	0.913	0.901	0.677	-0.871	-0.612	0.878	-0.916	0.759
MMAX	0.773	0.896	0.848	0.841	0.545	-0.753	-0.418	0.797	-0.877	0.602
SDAY	-0.965	-0.976	-0.954	-0.958	-0.653	0.963	0.699	-0.947	0.972	-0.812
FDAY	0.959	0.977	0.959	0.963	0.848	-0.958	-0.829	0.965	-0.972	0.884
FFP	0.966	0.980	0.969	0.969	0.783	-0.964	-0.828	0.963	-0.977	0.893
DD5	0.993	0.976	0.967	0.978	0.869	-0.991	-0.843	0.957	-0.990	0.899
GSDD5	0.989	0.976	0.969	0.970	0.829	-0.984	-0.868	0.953	-0.982	0.917
D100	-0.979	-0.972	-0.957	-0.973	-0.804	0.976	0.687	-0.955	0.979	-0.795
DD0	-0.836	-0.931	-0.968	-0.918	-0.877	0.814	0.881	-0.945	0.855	-0.824
MMINDD0	-0.971	-0.963	-0.971	-0.967	-0.898	0.963	0.888	-0.960	0.975	-0.837
TDIFF	-0.854	-0.886	-0.769	-0.779	-0.643	0.844	0.677	-0.764	0.655	-0.420
MAP	0.180	0.690	0.652	0.720	0.499	-0.164	-0.125	0.696	-0.277	0.006
GSP	0.327	0.617	0.480	0.510	0.199	-0.297	0.307	0.560	-0.127	-0.395
AMI	0.711	0.727	0.579	0.537	0.005	-0.760	-0.268	0.452	-0.750	0.308
SMI	0.688	0.906	0.883	0.851	0.535	-0.735	-0.805	0.840	-0.860	0.810
SMRPB	0.018	-0.219	-0.799	-0.554	-0.282	-0.111	0.264	-0.706	-0.274	-0.422
SMRSPRPB	0.112	-0.067	-0.847	-0.461	-0.732	-0.199	0.652	-0.748	-0.348	-0.590
PRATIO	0.156	-0.509	-0.573	-0.748	-0.547	-0.134	0.410	-0.623	0.442	-0.293

Table 4.4, continuing on next page

	Red Pine	E. white pine	Chestnut oak	Eastern hemlock	Black oak	Black cherry	Shortleaf pine	White ash	Paper birch	Eastern redcedar
MAT	0.979	-0.991	0.993	0.994	-0.994	0.997	0.996	0.995	-0.953	0.997
MTCM	0.771	-0.956	0.952	0.967	-0.957	0.973	0.964	0.971	-0.617	0.973
MMIN	0.767	-0.944	0.966	0.947	-0.932	0.963	0.959	0.961	-0.606	0.962
MTWM	0.742	-0.855	0.906	0.879	-0.889	0.922	0.831	0.934	-0.839	0.825
MMAX	0.627	-0.715	0.857	0.766	-0.822	0.886	0.670	0.903	-0.732	0.714
SDAY	-0.799	0.945	-0.960	-0.956	0.954	-0.964	-0.979	-0.967	0.884	-0.935
FDAY	0.874	-0.943	0.953	0.932	-0.956	0.971	0.976	0.966	-0.903	0.949
FFP	0.877	-0.954	0.960	0.958	-0.966	0.972	0.982	0.973	-0.933	0.954
DD5	0.900	-0.954	0.986	0.965	-0.971	0.974	0.991	0.975	-0.935	0.978
GSDD5	0.892	-0.956	0.971	0.969	-0.965	0.971	0.985	0.974	-0.942	0.960
D100	-0.868	0.928	-0.978	-0.929	0.967	-0.971	-0.976	-0.967	0.882	-0.973
DD0	-0.853	0.959	-0.904	-0.964	0.945	-0.933	-0.873	-0.952	0.752	-0.869
MMINDD0	-0.861	0.966	-0.976	-0.966	0.966	-0.968	-0.960	-0.970	0.773	-0.952
TDIFF	-0.507	0.824	-0.584	-0.836	0.716	-0.786	-0.789	-0.755	0.227	-0.849
MAP	0.410	-0.705	0.524	0.593	-0.735	0.690	0.226	0.592	0.078	0.678
GSP	0.045	-0.615	0.346	0.519	-0.568	0.518	-0.019	0.490	0.209	0.432
AMI	0.432	-0.224	0.493	0.426	-0.478	0.683	0.715	0.715	-0.494	0.170
SMI	0.858	-0.723	0.730	0.753	-0.847	0.891	0.853	0.905	-0.850	0.737
SMRPB	-0.376	0.528	-0.203	-0.511	0.758	-0.453	-0.193	-0.643	0.355	-0.240
SMRSPRPB	-0.742	0.556	-0.343	-0.485	0.708	-0.339	-0.082	-0.659	0.472	-0.281
PRATIO	-0.390	0.513	-0.751	-0.404	0.717	-0.628	-0.503	-0.473	0.056	-0.735

Table 4.5 The correlation coefficient matrix for measuring PCIPcom2 and climate variables

	Loblolly pine	Red maple	Sugar maple	White oak	Quaking aspen	Sweetgum	N. white-cedar	N. red oak	Yellow poplar	Balsam fir
MAT	-0.053	0.025	-0.048	-0.039	-0.137	-0.052	-0.146	-0.056	0.042	0.183
MTCM	0.113	-0.096	0.183	0.132	-0.523	0.128	-0.483	0.200	0.160	0.640
MMIN	0.119	-0.120	0.193	0.156	-0.508	0.131	-0.464	0.227	0.140	0.629
MTWM	-0.289	0.208	-0.361	-0.340	0.670	-0.355	0.620	-0.443	-0.205	-0.566
MMAX	-0.499	0.274	-0.451	-0.387	0.651	-0.528	0.670	-0.531	-0.225	-0.676
SDAY	-0.018	-0.140	0.176	0.177	-0.647	-0.011	-0.458	0.240	0.046	0.251
FDAY	0.108	0.110	-0.126	-0.064	0.191	0.094	0.070	-0.102	-0.035	-0.067
FFP	0.079	0.125	-0.162	-0.125	0.489	0.066	0.237	-0.181	-0.044	-0.166
DD5	-0.059	0.181	-0.192	-0.152	0.445	-0.061	0.425	-0.253	-0.014	-0.365
GSDD5	-0.065	0.185	-0.210	-0.195	0.535	-0.072	0.426	-0.281	-0.065	-0.333
D100	-0.049	-0.145	0.114	0.008	-0.415	-0.044	-0.550	0.129	-0.129	0.459
DD0	0.039	0.262	-0.131	-0.186	0.446	0.021	0.418	-0.226	-0.232	-0.515
MMINDD0	-0.023	0.169	-0.099	-0.134	0.392	-0.006	0.378	-0.166	-0.145	-0.476
TDIFF	-0.326	0.277	-0.506	-0.415	0.718	-0.365	0.652	-0.533	-0.432	-0.843
MAP	-0.100	-0.567	0.678	0.652	-0.720	0.000	-0.762	0.640	0.923	0.922
GSP	0.553	-0.216	0.474	0.599	-0.362	0.621	-0.218	0.399	0.788	0.600
AMI	-0.001	0.613	-0.801	-0.830	0.928	-0.086	0.889	-0.873	-0.635	-0.915
SMI	-0.484	0.252	-0.409	-0.485	0.668	-0.510	0.419	-0.478	-0.405	-0.498
SMRPB	0.959	0.830	-0.266	-0.186	-0.214	0.942	0.235	-0.333	-0.538	-0.527
SMRSPRPB	0.944	0.861	-0.333	-0.186	0.461	0.925	0.611	-0.393	-0.571	-0.720
PRATIO	0.815	0.742	-0.685	-0.488	0.741	0.755	0.863	-0.666	-0.724	-0.900

Table 4.5, continued on next page

	Red Pine	E. white pine	Chestnut oak	Eastern hemlock	Black oak	Black cherry	Shortleaf pine	White ash	Paper birch	Eastern redcedar
MAT	0.122	-0.061	-0.042	-0.032	-0.039	-0.049	0.012	-0.055	0.254	0.028
MTCM	0.615	0.152	0.140	0.116	0.213	0.122	0.226	0.123	0.751	-0.119
MMIN	0.621	0.128	0.124	0.073	0.265	0.142	0.244	0.107	0.736	-0.195
MTWM	-0.638	-0.418	-0.337	-0.308	-0.421	-0.306	-0.453	-0.288	-0.476	0.471
MMAX	-0.631	-0.552	-0.406	-0.387	-0.495	-0.340	-0.590	-0.323	-0.570	0.578
SDAY	0.517	0.183	0.091	0.102	0.186	0.199	0.059	0.172	0.258	-0.151
FDAY	0.003	-0.117	-0.080	-0.179	0.001	-0.132	-0.008	-0.137	0.004	0.014
FFP	-0.298	-0.155	-0.098	-0.148	-0.104	-0.176	-0.030	-0.161	-0.152	0.073
DD5	-0.407	-0.235	-0.119	-0.173	-0.179	-0.175	-0.061	-0.153	-0.276	0.148
GSDD5	-0.437	-0.256	-0.175	-0.197	-0.223	-0.208	-0.127	-0.186	-0.277	0.202
D100	0.396	0.130	-0.093	0.102	-0.015	0.090	-0.154	0.067	0.320	0.019
DD0	-0.491	-0.143	-0.099	-0.132	-0.188	-0.211	-0.312	-0.111	-0.625	0.224
MMINDD0	-0.485	-0.078	-0.083	-0.050	-0.173	-0.122	-0.236	-0.047	-0.582	0.231
TDIFF	-0.847	-0.369	-0.521	-0.303	-0.595	-0.413	-0.528	-0.443	-0.928	0.382
MAP	0.546	0.591	0.822	0.723	0.608	0.625	0.783	0.726	0.893	-0.664
GSP	-0.404	0.244	0.851	0.474	0.434	0.458	0.632	0.463	0.422	-0.665
AMI	-0.775	-0.940	-0.850	-0.877	-0.847	-0.700	-0.577	-0.683	-0.837	0.920
SMI	-0.194	-0.486	-0.660	-0.501	-0.454	-0.381	-0.409	-0.366	-0.392	0.602
SMRPB	0.237	-0.587	-0.309	-0.641	0.017	-0.561	0.532	-0.448	-0.490	-0.420
SMRSPRPB	-0.459	-0.624	-0.407	-0.624	0.055	-0.539	0.547	-0.416	-0.805	-0.318
PRATIO	-0.892	-0.756	-0.527	-0.722	-0.553	-0.638	-0.527	-0.753	-0.947	0.525

Table 4.6 Random Forests predictor variables and the rank of importance

Species	1	2	3	4	5	6	7	8	9	10
Loblolly pine	**CR**	**CCLCD**	pH	**NO10**	**TAWC**	**BA**	CLAY	KSAT	SAND	QMD
Red maple	**CR**	CCLCD	**LON**	**MAP**	**BASC**	**PRATIO**	**BA**	KFFACT	**GSP**	SMRSPRPB
Sugar maple	**CR**	CCLCD	BASC	**N**	BA	NO10	**PRATIO**	**QMD**	GSP	**SMRSPRPB**
white oak	**CCLCD**	**CR**	BA	**QMD**	**PREVDIA**	**AMI**	LPREVDIA	**SMI**	**PREVDIA2**	**GSP**
Quaking	**CR**	**CCLCD**	**BASC**	**NO10**	pH	**CLAY**	**BA**	KSAT	**KFFACT**	GSP
Sweetgum	CCLCD	**CR**	**BA**	QMD	SMI	**NO200**	CLAY	**SAND**	**GSP**	TAWC
N. white-cedar	**CR**	**BASC**	**N**	BA	**TAWC**	**SMRSPRPB**	LAT	pH	**LON**	PRATIO
Northern red oak	**CCLCD**	**BASC**	**BA**	**CR**	NO10	**SMRSPRPB**	**LON**	**AMI**	**PRATIO**	SAND
Yellow-poplar	**CCLCD**	**CR**	**BA**	**LON**	**BASC**	**MAP**	QMD	**GSP**	**PRATIO**	SAND
Balsam fir	**CR**	CCLCD	N	**LON**	pH	NO10	**BA**	LAT	GSP	**NO200**
Red pine	**CR**	**CCLCD**	pH	CLAY	**BA**	**NO10**	**TAWC**	BASC	NO200	QMD
E. white pine	**CR**	**CCLCD**	BA	BASC	QMD	NO200	pH	**LON**	**AMI**	SAND
Chestnut oak	CCLCD	BA	**BASC**	**QMD**	**NO10**	**CR**	NO200	**LON**	TDIFF	AMI
Eastern hemlock	**CR**	CCLCD	**BA**	BASC	NO10	**PRATIO**	**MAP**	KFFACT	CLAY	**SAND**
Black oak	**CCLCD**	**CR**	BA	**QMD**	NO200	**LON**	TAWC	**PRATIO**	SAND	**MAP**
Black cherry	**CR**	**CCLCD**	LON	**PRATIO**	CLAY	**BA**	**QMD**	SMRSPRPB	AMI	SAND
Shortleaf pine	**CR**	CCLCD	**BA**	**QMD**	BASC	**LON**	**NO10**	NO200	TAWC	SMI
White ash	**CR**	CCLCD	BA	**N**	LON	QMD	**CLAY**	PRATIO	**LAT**	TDIFF
Paper birch	**CR**	CCLCD	**N**	**BASC**	BA	pH	**GSP**	LAT	CLAY	**MAP**
Eastern redcedar	**CR**	**CCLCD**	**BA**	**TAWC**	**pH**	**N**	BASC	QMD	LON	SAND

Note: Variables shown in bold were among the most important ten indicators either based on RF or MLR.

Figure 4.1 Scatterplots of studentized residuals against predicted ln (DDS) before and after weight transformation

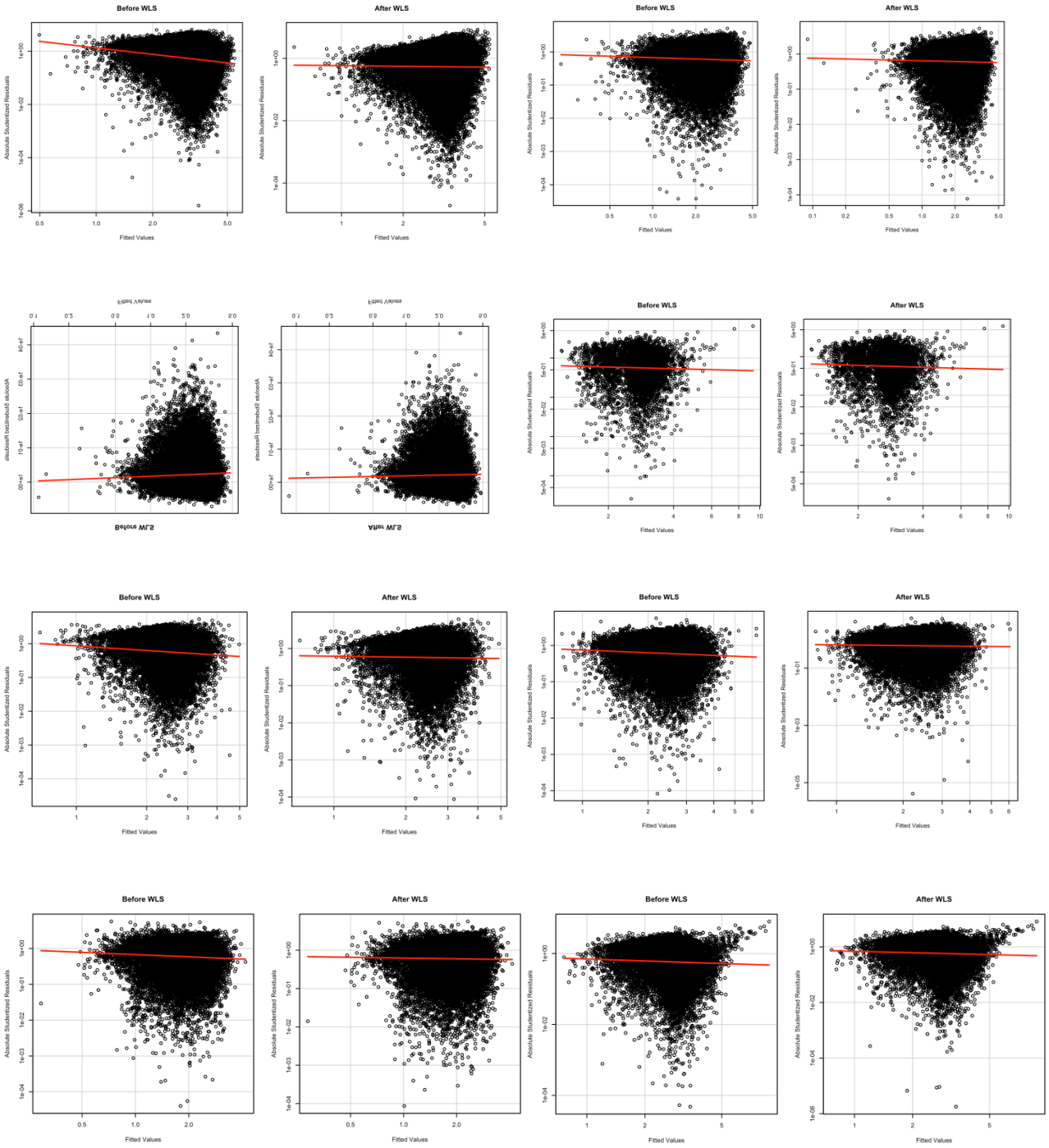


Figure 4.1, continued on next page

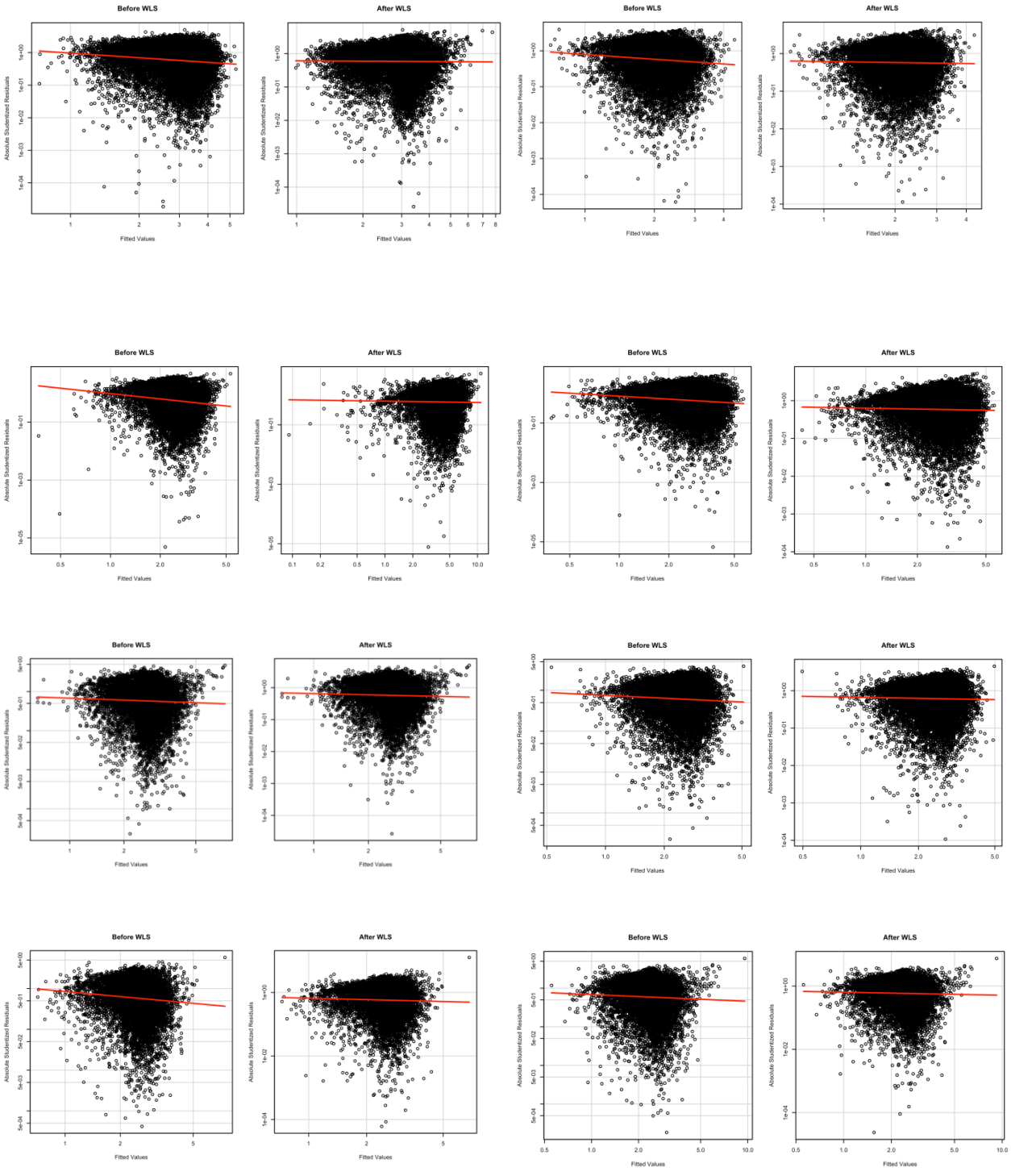


Figure 4.1, continued on next page

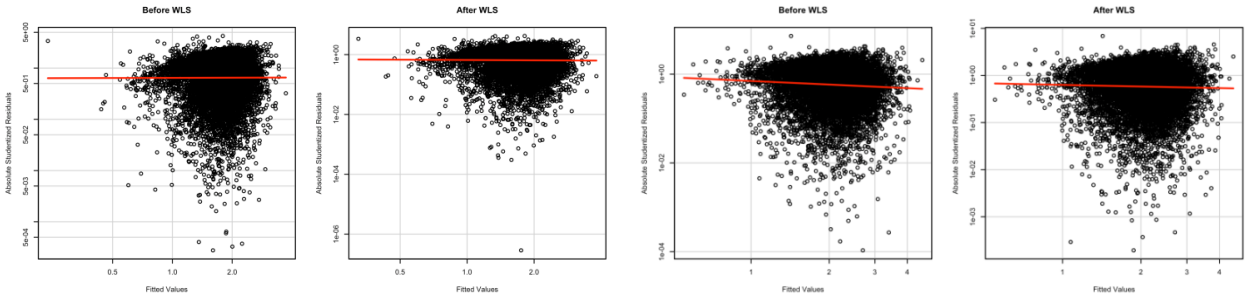

Before WLS
After WLS
Before WLS
After WLS
Absolute Studentized Residuals
Fitted Values

Figure 4. 2 Spatial map of ADI-based quantities for red maple.

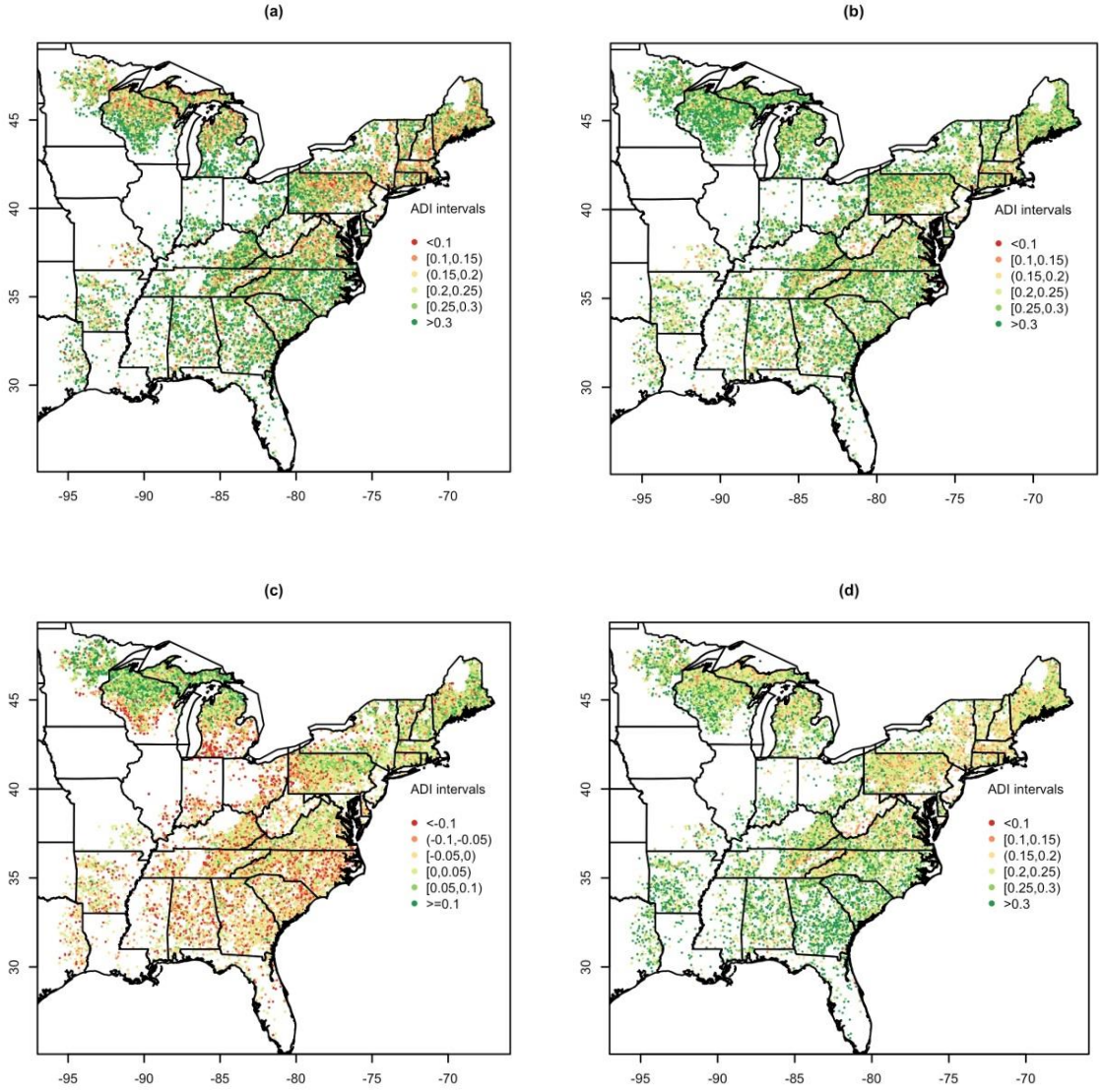
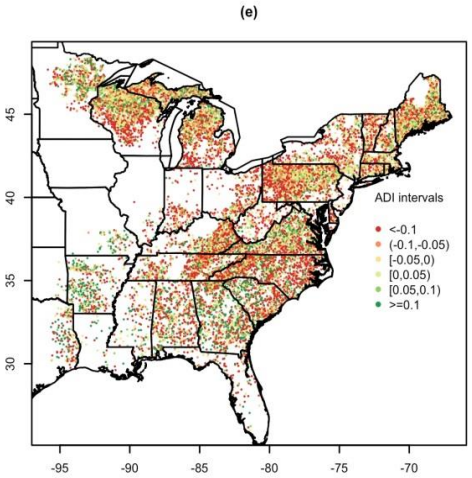
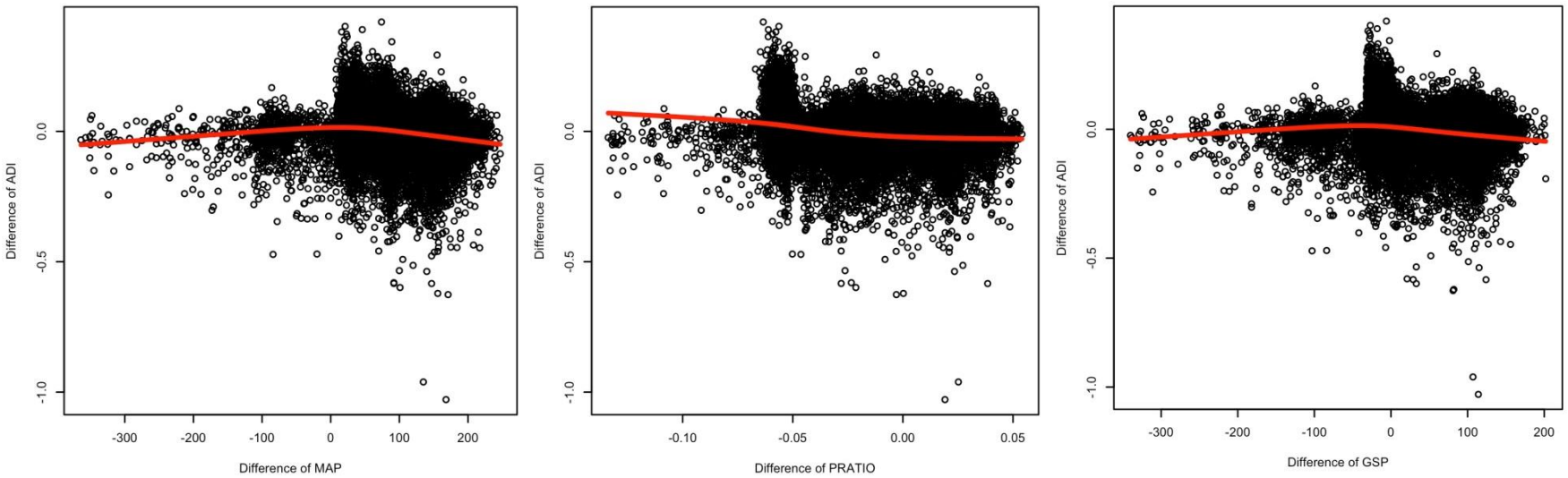


Figure 4. 2, continued on next page



Note: Spatial map of contemporary average $ADI_{contemporary}$ (cm/year) (a); predicted average ADI_{2090} (cm/year) for red maple in 21st century for A2 climate change scenarios based on RF model (b); predicted change (ADI_{2090}-$ADI_{contemporary}$) in average ADI (cm/year) for red maple based on RF model (c); predicted average ADI_{2090} (cm/year) in 21st century for A2 climate change scenarios based on MLR model (d); predicted change(ADI_{2090}-$ADI_{contemporary}$) in average ADI (cm/year) based on MLR model (e).

Figure 4.3 Marginal correlation plots for predicted red maple change of ADI versus change in climate variables for A2 development scenario.



Note: Difference of ADI = ADI_{2090}-$ADI_{comtemporary}$; Difference of MAP = MAP_{2090} – MAP $_{contemporary}$; Difference of GSP= GSP_{2090} – GSP $_{contemporary}$

Chapter 5 Summary and conclusion

The main task of this work is proposing modeling tools to assist FVS-SN in accounting for temperature and precipitation-related variables in making predictions of forest tree growth, yield, and mortality. In order to make models compatible with the FVS-SN, the data used in the analyses are national-scale and collected from the USDA Forest Service Forest Inventory and Analysis program, USDA Forest Service Rocky Mountain Research Station online climate data, and USDA Soil Survey Geographic SSURGO database. Climate-sensitive models for predicting site index, individual tree mortality, and individual tree diameter growth were developed to contribute the main components in FVS-SN. Random Forests as alternatives to other statistical regression models were used to develop climate-sensitive models for most 20 common eastern species due to its merits in modeling complex interactions or nonlinear relationships among variables. With the support of this machine learning tool, spatio-temporal relationships between climate and individual tree species was predicted under future projected climate scenarios using the Southern Variant Climate-FVS.

Although FIA plots were not limited to single-species or even-aged forests for conifers and hardwoods, evidence from spearman's rank correlation between GPP and site index, percentage of the site index measurements involved only one species of site tree on a plot, and the ranges of site tree ages on plots showed that FIA data were generally compatible with site index model development. A suite of site index models based on both climate and soil data were developed to evaluate the impact of climate change on site productivity in eastern forest species. Site index predictions for species grouped as conifers or hardwoods were nearly as precise as species-specific models for many of the most common eastern forest tree species. Soil properties were somewhat less useful in predicting site index across eastern US forests as were climate variables; however, soils and climate used together provided slightly more predictive power than either climate or soils alone. The variable importance measure showed that both soils and climate variables were among the most important predictors of site index in both conifer and hardwoods. Multiple maps were produced to illustrate contemporary patterns of site productivity across the region, and the potential for significant site index change with greenhouse-gas related global

warming. A bootstrap procedure was implemented to determine if differences between contemporary and future SI predictions were statistically significant. Even though spatial patterns of the change in site index over time varied between conifers and hardwoods and depending on the climate scenario examined, some consistent results from each scenario showed that areas of significantly decreasing site index in conifers included southern-tier states from Texas to South Carolina. Areas of potential increasing conifer site index included New England, western Lakes States, and the lower Midwest; however areas of potential increasing hardwood site index included New England, central Texas, and northern Michigan.

Three widely used classification methods with binary dependent variable, logistic regression model, artificial neural networks, and Random Forests were chosen to construct climate-sensitive mortality models (LR1, LR2, ANN, RF and RFo) to predict tree's survival probability over a specific time period for 20 most common eastern species. Bootstrap procedure was implemented to train and test each model 40 times so the mean and standard deviation of MAD and AUC over bootstraps were calculated. Results showed that LR1, LR2, ANN, RF, and RFo all performed better for species with lower mortality. LR models were the most stable models with consistently lowest values of standard deviation of MAD and AUC over bootstraps for all 20 species. The standard deviations of MAD and AUC of RF and RFo are generally as low as LR for all 20 species, except several noticeable jumps observed in standard deviation of AUC for the species with low mortality. ANN turned out to be the most instable model due to its highest values of standard deviation of either MAD or AUC at 20 species. LR1 and LR2 had finer rank and outperformed the other approaches at five species having relatively low mortality with respect of AUC; for the rest of 15 species, however, RF and RFo models had the best performance with distinctively superior behavior for species with higher mortality rate. The value of mean of AUC over bootstraps of LR2 was at least as high as that of LR1 at 16 species. Although ANN models were able to achieve the lowest mean of MAD over bootstraps for most 19 species, it had an overall poor performance as to $\overline{AUC}$ for most of the 20 species. Model performance test regardless of species showed that ANN models have the best performance as to $\overline{MAD}$, followed by RFo, RF, LR2 and LR1 regardless of species, and all of performance difference among five models are significant. In terms of $\overline{AUC}$, RFo models outperforms RF, which is significantly

better than LR1 and ANN, however, no significant difference was detected between RF and LR2 as well as LR1 and LR2.

Random Forests (RF) and Multiple Linear Regression model (MLR) were proposed to develop climate-sensitive annual diameter increment (ADI) model for individual trees for the 20 most common eastern tree species in the FIA data base. Although the median residuals of RF models were more deviated away from zero than that of MLR for many species, the range of half of residuals derived by RF models were narrower that that by MLR models based on the interquantile reading of residuals. R_a^2 computed by RF models was higher than what was achieved by MLR models for all species except for shortleaf pine with a slight better fitting predicted by MLR. The value of RMSE based on RF models was mostly smaller than that based on MLR models, except sweetgum, balsam fir and shortleaf pine. The agreement on top important variables between two models differed by species, but they both agreed that tree-level variables had more influential roles than climate and soils factors. The resulting MLR and RF models for each species group were applied to the ADI predictions for future climate change A2 scenario at eastern sites to demonstrate the possible changes in ADI over space and time. The prediction of average ADI for red maple showed that RF model created more consistent future prediction results with past studies than MLR model. The results from RF model showed that average ADI was to be frequently higher in the south area and Southern Wisconsin. A lower value was seen at most area of New England, northern Michigan, Minnesota, and Mid-Atlantic region. The resulting prediction under climate change A2 scenario based on RF showed that average ADI in the 2090s was above 0.25 cm/year at most areas of eastern U.S. The biggest value of the increase in average annual diameter increment over time was mainly located in Northern Michigan, Minnesota, where the increase was above 0.1cm/year.

As mentioned above, Random Forests based climate sensitive models were developed separately for predicting the effects of global warming on site index, individual tree mortality, and diameter growth. The developed model were species-specific and constructed based on inventory-based attributes including tree species, numbers, and sizes, plot-level climate and soils information. Public-free R package “randomForest” was used for model development so that little additional cost to model users. In addition, model validation and future project on site index, tree mortality

and diameter growth had demonstrated Random Forests was an useful tool to assist FVS-SN in simulating tree growth and yield under future projected climate scenario. The results also demonstrated the utility of using climate and soils data in predicting eastern forest growth across a large geographic region, and the potential of climate change to influence tree growth in the Eastern United States. The prediction of site quality and growth patterns of individual trees will serve policymakers and forest managers to take action to mitigate negative effects of global warming on forest ecosystems, which will allow for the selection of tree species for logging or protection as well as the estimation of cutting.