Incorporating Climate Sensitivity for Southern Pine Species into the Forest Vegetation Simulator

Melissa Dawn Shockey

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

Master of Science
In
Forestry

Philip J. Radtke
Carolyn A. Copenheaver
Stephen P. Prisley

April 17th, 2013

Keywords: Random Forest, Classification, Abundance, Climate-Soils-Vegetation modeling, Climate Change

Incorporating Climate Sensitivity for Southern Pine Species into the Forest Vegetation Simulator

Melissa Dawn Shockey

Abstract

Growing concerns over the possible effects of greenhouse-gas-related global warming on North American forests have led to increasing calls to address climate change effects on forest vegetation in management and planning applications. The objectives of this project are to model contemporary conditions of soils and climate associated with the presence or absence and abundance of five southern pine species: shortleaf pine *(Pinus echinata* Mill.*)*, slash pine (*P. elliottii* Engelm.*)*, longleaf pine *(P. palustris* Mill.*)*, pond pine (*P. serótina* Michx.), and loblolly pine *(P. taeda* L.*).* Classification and regression based Random Forest models were developed for presence-absence and abundance data, respectively. Model and diagnostics such as receiver operating curves (ROC) and variable importance plots were examined to assess model performance. Presence-absence classification models had out-of-bag error rates ranging from 6.32% to 16.06%, and areas under ROC curves ranging from 0.92-0.98. Regression models explained between 13.76% and 43.31% of variation in abundance values. Using the models based on contemporary data, predictions were made for the future years 2030, 2060, and 2090 using four different greenhouse gas emissions scenarios and three different general circulation models. Maps of future climate scenarios showed a range of potential changes in the geographic extent of the conditions consistent with current presence observations. Results of this work will be incorporated into eastern U.S. variants of the Forest Vegetation Simulator (FVS) model, similar to work that has been done for FVS variants in the West.

Table of Contents

Tables

Figures

Introduction

Growing concerns over the possible effects of greenhouse-gas-related global warming on North American forests have led to increasing calls to account for climate change in forest vegetation management and planning applications (McNulty et al. 2012). Understanding forest tree species' ranges and abundance in relation to contemporary climate and edaphic factors is a key step in developing models that can account for future climate change in predicting growth and yield (Iverson et al. 2008).

When developing management plans and prescriptions that account for climate change, it's important to know whether species can adapt and remain within their native ranges or not, whether mortality is likely to increase or decrease, and whether areas outside of existing ranges are likely to become suitable habitat for one or more species. Modeling tools capable of making predictions under climate change scenarios are therefore critical to forest managers and policymakers. Currently, most models being used for management and planning applications are not sensitive to climate inputs. They assume static climate conditions over time. Since it is generally accepted that climate conditions are changing and will continue to change, existing models may not be suitable for long-range planning. In the western United States the Forest Vegetation Simulator (FVS) has been augmented to include climate sensitivity; however, in the eastern U.S. these changes have not yet been made (Crookston et al. 2010).

The FVS model is a system of integrated analytical tools based on scientific knowledge developed from decades of natural resource research and management experience (Crookston and Dixon 2005). The USDA Forest Service maintains, supports, and develops updates and new extensions for FVS through its Forest Management Service Center (http://www.fs.fed.us/fmsc/fvs/). The first version of FVS was developed for northern Idaho and western Montana about 40 years ago and then was expanded to include all U.S. forests (Crookston and Dixon 2005; Stage 1973). At its core FVS is an individual-tree, distance-independent, growth and yield model that has been calibrated for specific geographic areas of the United States in modules called variants. The additional geographical variants were added over time, along with extensions that allowed for a wide range of management possibilities and scenarios. The variants in FVS can be used to

predict changes associated with stand dynamics, disturbance, and proposed management actions. The linkable modules, or extensions, can simulate events such as various insect and pathogen impacts, fire effects, fuel loading, snag dynamics and development of understory tree vegetation. Simulation can be performed for a wide variety of forest types, stand structures, and pure or mixed species stands (Dixon 2002). Climate-FVS is a recent addition to the FVS modeling system that provides forest managers a tool for considering the effects of climate change on forested ecosystems. It augments the predictors in FVS to take into account the effects of climate on mortality, growth, and regeneration (Crookston et al. 2010).

While no climate-sensitive version of FVS is available for the eastern U.S., considerable work has been done to demonstrate potential effects of climate change on eastern forests. Iverson and Prasad (1998) predicted the abundance of 80 tree species under various climate change scenarios in the eastern United States (areas east of the 100th meridian) using soils, climate, and vegetation data. Using Forest Service Forest Inventory and Analysis (FIA) vegetation data they estimated an average importance value for each of the target counties in their study. Using the calculated importance values, coarse soils maps, and regional climate data, they used regression trees to quantify the relationships between environmental factors and species distributions. Iverson et al. (2008) refined their methods using the Random Forests algorithm (Breiman 2001) and updated soils, climate, and vegetation data to estimate potential habitat for 134 eastern United States tree species. In their review of methods for developing species distribution models, Iverson et al. (2011) concluded that such models produced before the availability of machine-learning tools around 2005 are inferior to those produced later. A strength of recently created species distribution models that makes them superior is that they use extremely robust nonparametric statistical tools such as Random Forest (Iverson et al. 2011).

Iverson and Prasad (1998) performed their analyses at the county level using coarse (20 x 20 km), gridded data sets for soils and climate variables. Huang et al. (2011) used intermediate-resolution (4 x 4 km) gridded climate data, but did not consider soils in their climate-based model of forest productivity in loblolly pine. Finer scale (800 x 800 m) gridded climate data have been used in developing climate-sensitive forest productivity models and species distribution models in western North America, but detailed soils

information was unavailable and therefore not used (Latta et al. 2010; Rehfeldt et al. 2006). Other studies from western North America have developed climate-sensitive models of site productivity or species distributions, using either empirical methods or process-based models, but generally not incorporating detailed soils information (Coops et al. 2011; Coops and Waring 2011; Monserud et al. 2008; Weiskittel et al. 2011a)

In the western United States Rehfeldt et al. (2006) used vegetation presence-absence data instead of abundance data to make predictions of the viability potential for 9 constituent species in 25 biotic communities. The approach used by Rehfeldt et al. (2006) to define climatic conditions associated with a particular biotic community was to create a climate envelope. A climate envelope is a correlative approach for bioclimatic modeling that encompasses a realized climate niche. It is used to outline the distribution of biomes, communities, species, and populations; however, an envelope generally overestimates the realized niche (Rehfeldt et al. 2006).

In contrast to western North America, detailed soils data are widely available in the East. Iverson et al. (2008) incorporated soils data from the State Soil Geographic Data Base (STATSGO), the scale that (1:250,000) matched their county-level analyses. Finer-spatial resolution soils data for much of the eastern U.S. are also published in the Soil Survey Geographic (SSURGO) database, available at scales ranging from 1:12,000 to 1:63,360. Generally coarse resolution datasets such as STATSGO are best-suited for broad scale management and planning applications while finer resolution data such as SSURGO are suited for work involving individual land parcels (USDA 1994; USDA 1995).

The objectives of this project are to model contemporary conditions of soils and climate associated with the presence or absence of five southern pine species: shortleaf pine (*Pinus echinata* Mill.), slash pine (*P. elliottii* Engelm.), longleaf pine (*P. palustris* Mill.), pond pine (*P. serótina* Michx.), and loblolly pine (*P. taeda* L.). Data sources include (FIA) records of forest vegetation, SSURGO soils data, and Intergovernmental Panel on Climate Change (IPCC) future climate scenarios. Species presence or absence and abundance will be modeled using the classification and regression modeling methods known collectively as Random Forests. Diagnostic plots such as receiver operating curves and variable importance plots will be examined to assess goodness-of-fit and the explanatory power of

climate and soils variables as predictors. Using the models based on contemporary data, predictions will be made for the future decades starting in 2030, 2060, and 2090 using climate predictions from four different IPCC scenarios and three general circulation models.

Materials and Methods

Forest vegetation data will be linked with detailed soils and contemporary climate data to develop a series of species-specific climate-soils-vegetation relationships for five southern pines. The primary tool for characterizing vegetation-soils-climate relationships will be the Random Forests ensemble classification and regression-tree analysis software implemented in the R language and environment for statistical computing. Methods to be used will build on work that has been done in the western United States (Crookston et al. 2010; Rehfeldt et al. 2006; Weiskittel et al. 2011a) and by Iverson et al. (2008) in eastern tree species.

Target Species

Five southern pine species will be studied here because of their importance in the life cycle of the federally endangered red-cockaded woodpecker (*Picoides borealis)*, and because of the role these five species play in efforts to restore and preserve functioning native pine forest ecosystems in the North Carolina sandhills region. Old growth mixed pine stands including longleaf, loblolly and shortleaf pine are used by the red-cockaded woodpecker *Picoides borealis* (RCW) for nesting, roosting, and foraging (Santos et al. 2010). About 1% of the original red-cockaded woodpecker population remains, so the protection of this endangered species and its required habitat is a major goal of conservationists (Burns and Honkala 1990; Santos et al. 2010).

Longleaf pine is native to a wide variety of sites ranging from wet, poorly drained flatwoods to dry, rocky, mountain ridges. Red-cockaded woodpecker roosting and nesting are most-often observed in longleaf pines, with tree size and mortality playing major roles in the composition of suitable RCW nesting sites (Harding and Walters 2002). Loblolly pine is the most economically important tree species in the southern United States, comprising approximately 30 million acres of commercial forest land in the region (Smith et al. 2010). Its role in RCW nesting and roosting is second only to longleaf pine. Loblolly pine responds

well to silvicultural treatments and can be managed as either even-aged or uneven-aged natural stands, or by artificial regeneration in plantations (Schultz 1999). Shortleaf pine has the widest range of any pine species in the southeastern United States and is one of the four most important commercial conifers in that region. The species tolerates a wide range of soil and site conditions and maintains its growth rate for a relatively long period (Sabatia et al. 2008). Slash pine, a species known for its rapid early growth and frequent and abundant seed production, has the smallest native range of the four major southern pines (Powell and White 1994). Pond pine is frequently found as a major overstory species in pocosins, oligotrophic and precipitation fed coastal plain wetlands that can have a peat accumulation as deep as 2m (Bolin 2007). Pond pine stands and pocosins are considered to be a major wildlife sanctuary for wetland species (Burns and Honkala 1990).

All five pines are predominately found in the south so the primary area of focus will be the Southern (Sn) variant of FVS. Southern military lands encompassed by the Sn variant such as Fort Bragg, a large (49 km^2) Army installation in the Sandhills region of south-central North Carolina, U.S.A (Figure 1), are interested in results as they may impact future management decisions of those military lands.

Figure 1. Location of Fort Bragg, NC and the suggested range of FVS-SN

Geographic Scope

Thirty seven states from the eastern U.S. will be evaluated based on the presence or absence of the five target pine species within the study area (Table 1). For the purpose of this work the eastern U.S. will be defined by the states having any area east of 100th meridian, corresponding to forests east of the Great Plains. The states included either A) have been documented by FIA to have at least one of the target species present or B) are in a state of possible interest within this study as potentially supporting one or more of the target species as climate changes in the 21st century. In some northern tier states, the interest in these species may be very low; however, they will be included here to ensure that a wide range of climate and soils conditions are covered, both for areas where the pine species are known to be present and in areas where they are not.

Table 1. Names and abbreviations of all states included in analysis. Asterisk (*) denotes a state where at least one of the target species can currently be found in a FIA plots

States	Abbreviations	States	Abbreviations
Alabama	AL*	North Carolina	NC*
Arkansas	AR*	North Dakota	ND
Connecticut	CT	Nebraska	NE*
Delaware	DE*	New Hampshire	NH
Florida	FL*	New Jersey	NJ*
Georgia	GA*	New York	NY
Iowa	IA	Ohio	OH*
Illinois	IL*	Oklahoma	OK*
Indiana	IN*	Pennsylvania	PA
Kansas	KS	Rhode Island	RI
Kentucky	KY*	South Carolina	SC*
Louisiana	LA*	South Dakota	SD
Massachusetts	MA	Tennessee	TN*
Maryland	MD*	Texas	TX*
Maine	ME	Virginia	VA*
Michigan	MI *	Vermont	VT
Minnesota	MN	Wisconsin	WI
Missouri	MO*	West Virginia	WV*
Mississippi	MS*		

Vegetation Data

The vegetation data source used is the publicly-available online database of forest plot and tree measurements hosted by the FIA national program. The FIA program is charged by Congress with the assessment of the Nation's forests, including whether current forest management practices are sustainable and if current policies are encouraging healthy forests. Data collected through this project are freely accessible online through the FIA DataMart for all 50 US states, Puerto Rico, and the Virgin Islands in comma separated value (CSV) formatted computer files (Forest Inventory and Analysis 2012).

FIA is a continuing program mandated by Congress in the McSweeny-McNary Forest Research Act of 1928 and the Forest and Rangeland Renewable Resources Planning Act of 1974 . The primary objective of FIA is to determine the extent, condition, volume, growth, and harvest of timber from all ownership types of forest land within the United States. Starting in 1929 FIA conducted periodic forest inventories on a state by state basis. With the 1998 Farm Bill, FIA is now required to collect data annually on at least 20% of the plots within each state in the east (Reams et al. 1999). There is approximately one FIA plot for every 6,000 acres of forest land area in the United States and its territories (USDA Forest Service 2005).

Soils Data

Soil types and associated attributes were compiled from the Soil Survey Geographic (SSURGO) online database and linked to FIA's vegetation plots based on their actual geographic coordinates. Actual coordinates were used here to ensure maximum correspondence between the mapped soil types in SSURGO and the actual edaphic conditions on the FIA field plots. SSURGO data represent the most detailed level of soil mapping done by the Natural Resources Conservation Service (NRCS) with mapping scales ranging from 1:12,000 to 1:63,360. The data are available for download from the United States Department of Agriculture (USDA), NRCS, GeoSpatialDataGateway server (USDA NRCS 2011). Soils properties to be included in the analysis are listed in Table 2, namely those postulated to be relevant to the growth or survival of forest tree species by Iverson et.al (2008).

Table 2. Soil properties from SSURGO database to be used for analysis. Based on Table 2 of Iverson et al. (2008)

Variable Name	Variable Description
pawc_w	Available water capacity
pbd_w	Soil bulk density (g/cm3)
pclay_w	Percent clay (Mineral particles ranging in size from 0.002 to 0.02mm in equivalent diameter as a weight percentage of the less than 2.0mm fraction)
pkast_w	Saturated hydraulic conductivity
pkffact_w	Soil erodibility factor, rock fragment free (susceptibility of soil erosion to water movement)
pNO10_w	Percent soil passing sieve no. 10 (coarse)
pNO200_w	Percent soil passing sieve no. 200 (fine)
pOM_w	Organic matter content (% by weight)
pph_w	Soil pH
psand_w	Percent sand (Mineral particles 0.05 to 0.10mm in equivalent diameter as a weight percentage of the less than 2 mm fraction)
psilt_w	Percent silt (Mineral particles 0.05 to 0.10mm in equivalent diameter as a weight percentage of the less than 2 mm fraction)

Climate Data

Climate data were obtained from the USDA Forest Service Forestry Sciences Laboratory (FSL) online data repository for each FIA plot based on published geographic coordinates. Published coordinates include randomly-altered differences of up to 1.6 km from the actual plot locations to ensure the confidentiality of forest landowners (Wang et al. 2011). Due to strong spatial autocorrelation in most climate variables over this distance in eastern North America, the use of published rather than actual coordinates was deemed adequate. Elevation data were also obtained based on FIA plot published coordinates, noting that past work has shown strong positive correlations between elevations tabulated using published versus actual coordinates (Wang et al. 2011). Digital Elevation Model (DEM) data were obtained from Global 30 Arc-Second Elevation Data Set (GTOPO30) downloaded from the URL http://eros.usgs.gov/#/Find_Data/Products_and_Data_Available/GTOPO30. The unique plot id, elevation, latitude, and longitude were sent to the FSL's custom climate data request server at URL http://forest.moscowfsl.wsu.edu/climate/customData/ (Rocky Mountain Research Station 2012).

The FSL climate server returns contemporary and predicted climate variables for the years 2030, 2060, and 2090 for multiple climate change scenarios (Rehfeldt et al. 2006). Thirty-year averages are denoted as "contemporary" because they represent conditions from 1961 to 1990, a period that coincides with the conditions under which many extant forests in North America developed (Rehfeldt 2006). Examples of variables included in climate data downloads include monthly means for maximum, minimum, and average temperatures and precipitation. Downloads also include derived annual climate variables that quantify parameters relevant to tree growth, such as growing season precipitation (Table 5). To characterize geographic variation, the Moscow FSL climate data are interpolated across the region using thin-plate splines (Crookston 2012a; Hutchinson 1998a; Hutchinson 1998b; Rehfeldt 2006). Thin plate splines are also used to downscale General Circulation Model (GCM) predictions for future climate data from a coarse spatial resolution. The future climate data for the years 2030, 2060, 2090 are based upon updating the contemporary spline surfaces with projections from three General Circulation Models (GCMs) (Crookston et al. 2010; Hansen et al. 2001). GCMs have relatively coarse resolutions ranging between 1° and 4° of latitude and 1° and 4° longitude. Downscaling from the coarse GCM grids to point locations uses the weighted average of the monthly change for the GCM cell centers to update contemporary spline surfaces (Crookston et al. 2010). The climate data from FSL is based on three different GCMs and several A and B climate change scenarios developed by the Intergovernmental Panel on Climate Change (IPCC) (Crookston 2010) (Table 3).

Table 3. GCMs and scenarios used in Climate-FVS available for download from Moscow FSL

Short name	GCM group name and scenario identification
CGCM3_A1B	Canadian Center for Climate Modeling and Analysis, scenario A1B
CGCM3_A2	Canadian Center for Climate Modeling and Analysis, scenario A2
CGCM3_B1	Canadian Center for Climate Modeling and Analysis, scenario B1
GFDLCM21_A2	Geophysical Fluid Dynamics Laboratory, scenario A2
GFDLCM21_B1	Geophysical Fluid Dynamics Laboratory, scenario B1
HADCM3_A2	Hadley Center/World Data Center, scenario A2
HADCM3_B2	Hadley Center/World Data Center, scenario B2

The IPCC (2000) developed long-term emission scenarios that have been widely used in the analysis of possible climate change impacts and options to mitigate climate change. Four families of storylines were developed to describe the relationships between

the driving forces of emissions and their evolution and to add context for the scenario quantification. Each storyline represents different demographic, social, economic, technological and environmental developments. All scenarios based on one storyline are known as a scenario "family." All together 40 differentscenarios have been developed by the IPCC, all equally valid and with no assigned probabilities of occurrence. The four main storylines are the A1, A2, B1, B2 (IPCC 2007).

There are three groups within the A1 family, they characterize alternative developments of energy technologies: A1FI (fossil fuel intensive), A1B (balanced), and A1T (predominantly non-fossil fuel). They describe a future world of very rapid economic growth, global population that peaks in mid-century and declines after, and the rapid introduction of new and more efficient technologies. The A2 family describes a very heterogeneous world focusing on self-reliance and preservation of local identities. The B1 storyline describes a convergent world with the same global population scenario as A1 but with rapid changes in economic structures toward a service and information economy with reductions in material intensity and the introduction of clean and resource-efficient technologies. The B2 storyline describes a world in which the emphasis is on local solutions to economic, social, and environmental sustainability (IPCC 2007). These different IPCC scenarios are used to supply assumptions for various applications of GCMs. In Table 3 the scenario with the highest carbon dioxide-equivalent (CO_2-eq) emission levels is the A2 storyline at 1250ppm. The CO_2-eq emissions of the other scenarios studied here are shown in Table 4.

Table 4. Special report on Emission Scenarios storylines CO_2-eq concentrations in 2100

SRES storylines	**CO_2-eq (ppm)**
B1	600
B2	800
A1B	850
A2	1250

Data Screening

Data compilation was carried out, mainly in the fall 2011 and spring 2012 semesters. FIA data were downloaded for the states listed in Table 1 from the FIA DataMart (Forest Inventory and Analysis 2012). Corresponding SSURGO soils data for each state were downloaded from the GeoSpatial Data Gateway (USDA NRCS 2011). Tabular soils data, including map unit identifiers, were spatially joined to the FIA field plot tabular data by authorized Forest Service staff without revealing the actual plot coordinates to unauthorized personnel. Using the FIA published plot location information; plots were screened for obvious errors in published geographic coordinates, for example, when visual inspection showed that a plot identified as being measured in one state had published coordinates more than a few km outside of the state's geographic boundary. This was accomplished by mapping all plots within a state, one state at a time, so plots outside the boundaries could be easily identified.

Only plots measured under the national annual inventory design were used. Further to avoid duplication of remeasured FIA plots, only the data cataloged in the most recently recorded inventory year for each state were selected. The plot design in the annual inventory design consists of four 24-foot circular fixed-radius subplots for trees $\geq$ 5 inches diameter at breast height (DBH), and four 6.8-foot fixed-radius microplots for seedlings and trees $\geq$ 1 and < 5 inches DBH. The central plot is subplot 1, with subplots 2, 3, and 4 located 120.0 feet from plot center to plot center, at azimuths of 360°, 120°, and 240°, respectively. Microplot centers are positioned 12 feet east of each subplot center (Figure 2) (Forest Inventory and Analysis 2011). Data from optional 58.9-foot fixed radius macroplots were not included in this study.

Various other factors describing the land cover, land-use, and recent disturbance history were used in selecting suitable data for the analysis. For example, one factor of interest was the current land cover condition of the FIA plot. Only plots classified as "forest land" were used. Forest land is land with at least 10 percent cover (or equivalent stocking) by live trees of any size, including land that formerly had such tree cover and that will be naturally or artificially regenerated (Forest Inventory and Analysis 2011). Naturally regenerated plots with no recent disturbances observed were considered suitable, as were

those having a naturally occurring disturbance such as fire, or presence of insect activity. All plots with other disturbance codes such as fertilization, thinning, or harvesting were excluded so as to reduce the effects in the database of direct human impacts on natural forests. Artificially regenerated plots were retained regardless of the status of any disturbance or treatment codes. An assumption made in the selection of artificially-regenerated forest plots is that conditions suitable for plantation management in the contemporary climate should also be considered suitable for plantation management under future climate scenarios.

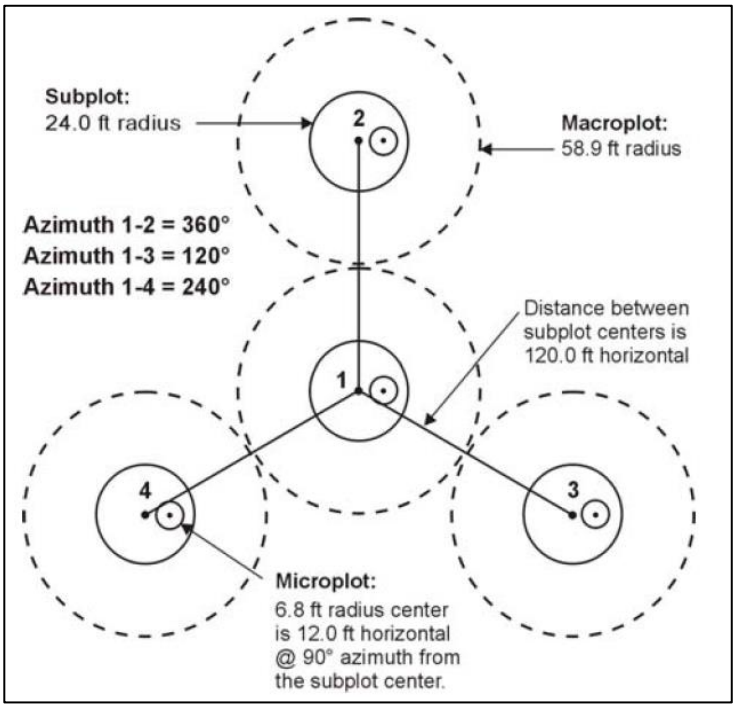


Figure 2. Annual FIA Inventory Plot Design (Design code 1)

Following data screening, presence or absence, on each plot, of the five target pine species were coded using binary 0/1 variables. Species importance value (IV) was also calculated for each plot as:

$$IV = \left(\frac{\frac{Species\ TPA}{Total\ TPA} + \frac{Species\ Basal\ Area}{Total\ Basal\ Area}}{2} \right) * 100 \qquad (1)$$

where

TPA denotes trees per acre.

Tabulated vegetation data were subsequently merged with the tabular soils and climate attributes so that all variables of interest were compiled in a common database for further analysis. Data screening and compilation were carried out using the R language and software environment (version 2.14.0) for statistical and graphical computing (The R Project 2011). Mapping and geospatial analyses were conducted using the Earth Systems Research Institute (ESRI) ArcGIS version 10 suite of software and R (ESRI 2012; The R Project 2011).

Random Forests (Data Mining)

Data mining is the process of analyzing data having many different linkages and combinations of variables, and interpreting and summarizing the data to obtain meaningful information. New developments in data-mining software allow users to analyze data collected over many different dimensions, categorize, and identify relationships among many variables. Correlations may exist across dozens of fields in large, complex databases (Sharma et al. 2012; Yin et al. 2011). The R implementation of Random Forest was used to conduct data mining analysis here (Liaw and Wiener 2002).

Decision and regression tree-based methods in Random Forests involve a data-partitioning algorithm that recursively splits observations into groups to produce trees with branches and nodes. Each node represents a split of the data into two non-overlapping groups, and each branch represents one of two data subsets constructed after splitting. Trees consist of many nodes and branches that ultimately split the data into different groups, such as present or absent, each of them disjoint from the others. A random forest is a collection of trees with each tree dependent on the values of a random vector of predictors sampled independently and with the same distribution for all trees within the forest (Rehfeldt et al. 2006). Modeling was performed using the Random Forests procedure implemented in R, where the algorithm builds a set of independent regression trees from an input data set (Liaw et al. 2012).

The Random Forest algorithm first extracts a bootstrap sample – with replacement and having the same number of the observations as the original data set – to construct a decision tree. The remaining observations, called the "out-of-bag sample," are used to compute classification errors from the tree. The "in-bag" sample consists of the observations that were randomly selected to be part of the tree (Breiman 2001). At each node of the tree, a random sample of the predictor variables is selected. Of these candidate variables the one that minimizes classification error is chosen. Nodes are further split until no more improvement can be achieved (Breiman 2001).

Both the classification and regression modeling systems of Random Forest use threshold values to make classification tree decisions based on a training dataset. A new bootstrap training set is drawn, with replacement, from the original training set, then a tree is grown based on the new training set using random feature selection (Breiman 2001). For our models 200 trees were developed for each plot record being analyzed. It is the aggregation of these 200 trees that will become the predicted importance value or presence-absence decision that is reported by the model for each plot.

Climate-Soils-Vegetation Model Relationships

In this analysis a climate-soils-vegetation modeling approach was used to determine which soil characteristics coincide with certain climate conditions where a species was observed in a recent FIA inventory. By looking at the soil properties and contemporary climate relationships, predictions can be made and species range shifts can be identified. The contemporary climate will serve to develop climate-vegetation relationships for subsequent use in predictions under future climate scenarios. Once the current vegetation-climate-soils relationships have been determined, patterns will be updated with the future climate predictions to ascertain how the changing climate could affect the species distributions. The new ranges will represent areas where, in the future there will be suitable conditions for the species of interest to grow. Table 5 lists the climate variables used in this analysis, this table was adapted from Table 1 in Rehfeldt et al. (2006).

The climate-soils-vegetation modeling approach was conducted by building two models for each species: one for the abundance measure of importance value, and one for the categorical approach of presence-absence. For the categorical approach of presence-

absence a Balanced Random Forest (BRF) will be used. In a BRF the number of observations of a minority class is what determines the maximum number of plots from the majority class that are used in the bootstrap samples (Chen et al. 2004). For example in the case of loblolly pine models, the minority class had 14,309 observations that are available to create the BRF, while pond pine models only has 317 observations available to create the BRF (Table 6). The majority class for the loblolly pine analysis has 56,125 observations and the pond pine majority class consists of 70,117 observations. While loblolly pine has the largest minority class it still only comprises 20 % of the total data. Pond pine, the species with smallest minority class, comprises < 0.5% of the data.

Classification algorithms aim to minimize the overall error rate, but in the case of imbalanced data when the rare, or in this case present, category is the subject of interest these algorithms do not perform well (Bose 2012). There are several ways to approach the problem of extremely imbalanced data. The BRF technique is one example which consists of down-sampling the majority class to create a more balanced dataset. Chen et al. (2004) cautions that while down-sampling the majority class is effective in improving model performance, it can lead to loss of information due to a large part of the majority class not being used. Chen et al. (2004) also reported that they found a BRF to be more efficient at improving model performance for prediction of "rare" categories than other techniques and equally as effective as Weighted Random Forest (WRF). A WRF puts more weight on the minority class, penalizing the model more heavily for misclassifying the minority class. A drawback of the WRF is that by placing more weight on the minority class, or over-sampling, the model may become more vulnerable to noise (mis-labeled class) than a BRF (Chen et al. 2004).

Table 5. Climate variables used as independent variables in random forests regression analyses. For climate interactions a "x" indicates a multiplication function was used to create the interaction variable, while a "_" denotes a division of the variables to create the interaction term.

Acronym	Definition
AMI	Annual moisture index, dd5/map
d100	Julian date the sum of degree-days >5°C reaches 100
dd0	Degree-days <0°C
dd5	Degree-days >5°C
fday	Julian date of the first freezing date of autumn
ffp	Length of frost-free period (days)
gsdd5	Degree-days >5°C accumulating within the frost-free period
gsp	Growing season precipitation, April-September (mm)
map	Mean annual precipitation (mm)
mat	Mean annual temperature (°C)
mmax	Maximum temperature in the warmest month (°C)
mmin	Minimum temperature in the coldest month (°C)
mmindd0	Minimum degree-days <0
mtcm	Mean temperature in the coldest month (°C)
mtwm	Mean temperature in the warmest month (°C)
PRATIO	Ratio of summer precipitation to total precipitation, gsp/map
sday	Julian date of the last freezing date of spring
SMI	Summer moisture index, gsdd5/gsp
smrpb	Summer precipitation balance: (July+August+September)/(April+May+June)
smrsprpb	Summer/Spring precipitation balance: (July+August)/(April+May)
TDIFF	Summer-winter temperature differential, mtwm - mtcm

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.

From: (Rehfeldt et al. 2006)

Table 6. The total number of plots used in the analysis and the breakdown of how many plots contained each of the target species.

Species	Number of plots
Shortleaf pine	4,436
Slash pine	2,753
Pond pine	317
Longleaf pine	1,447
Loblolly pine	14,309
All plots in analysis	70,434

Climate-FVS

Climate-FVS is a recent addition to the FVS modeling system that augments the predictors to take into account the effects of climate on mortality, growth, site productivity, and regeneration for forest in the western U.S. (Crookston et al. 2010). Climate-FVS uses climate variables to predict the viability of a species at a point in space. Species viability scores are based on predictions of presence or absence by Random Forests classification trees based on numeric scores between 0.0 and 1.0 generated by the Random Forest model. A viability score of 0.0 means zero trees in the random forest classified the species as "present" at a given place. A score of 1.0 means 100% of the trees classified the species as "present" (Crookston 2010). More generally, the classification scores used in Climate-FVS to measure species viability are known as votes (Breiman 2001). In Climate-FVS species with viability scores <0.2 are classified as non-viable for the given climate conditions (Crookston et al. 2010).

Results

Model Development and Training

Two models were developed for each species for a total of ten random forest models. The presence-absence (PA) models were developed using a categorical BRF technique. Numbers of FIA plots of which a species was observed as "present" constitute the minority sample sizes in BRF (Table 6). Predicted importance value models (IV) were developed using a regression based random forest algorithm using the full dataset rather than a balanced approach (Liaw and Wiener 2002; R Development Core Team 2011).

It was determined that 200 trees would be sufficient for model development through investigation of error rates of models containing 200 trees and 500 trees. Using slash pine as an example it was noted that the mean of squared residuals (MSE) for the IV model is high when the number of trees is small (Figure 3). As the number of trees increases, MSE decreases until it reaches an inflection point and levels-off. The plots in Figure 3 confirm that little additional decrease in MSE will be realized by increasing the number of trees above 200. So the choice of 200 trees used here was deemed to be adequate. Table 7 shows the numerical output and confirms that the MSE of the example models containing 200 and 500 trees are approximately 91. The percent of variability explained by both models are roughly 43%, so increasing the number of trees from 200 to 500 did not greatly improve the predictive ability of the model. For the slash pine example PA model in Figure 3 the black line represents the out of bag (OOB) estimate of error rate. For both classes (present and absent) the red line represents the classification error for plots where the species did not occur, and the green line represents the classification error for the plots where the species did occur. The PA model behaves similarly to the IV model in that the error rates are the highest when only a small number of trees are used and the error rates decreased as more trees are used, reaching a stable lower limit. In Table 7 the OOB, absent classification, and present classification error rates of the PA models containing 200 and 500 trees are approximately the same. This leads to the conclusion that like the example IV model the addition of 300 trees does not significantly improve PA model performance. The use of BRF typically leads to classification error rates for presence and absence that are roughly equal in magnitude (Chen et al. 2004).

Table 7. Error rates of IV and PA models for slash pine containing 200 and 500 trees

	Mean of squared residuals	Percent of variability Explained	OOB estimate of error rate	Absent classification error	Present classification error
IV 200 trees	91.3121	43.27%			
IV 500 trees	90.9782	43.48%			
PA 200 trees			6.32%	6.37%	5.05%
PA 500 trees			6.26%	6.30%	5.23%

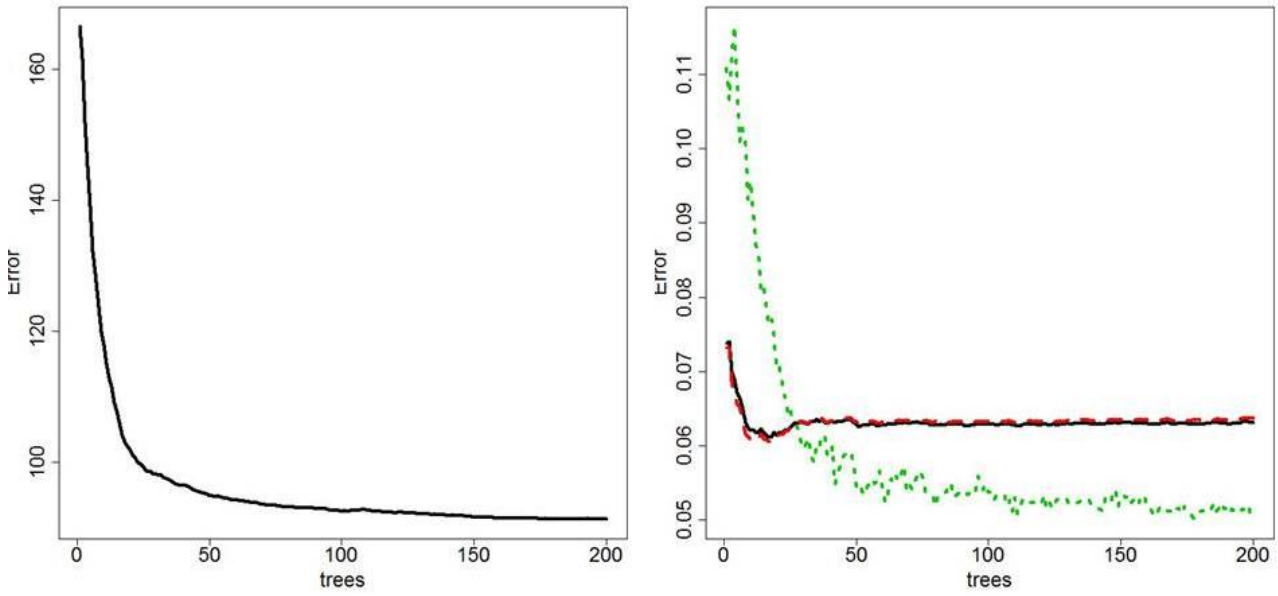


Figure 3. Mean-squared error plots for slash pine IV random forest model (left) and PA random forest model (right) containing 500 trees. The black line represents the out of bag (OOB) estimate of error rate for the model. The red line represents the classification error for plots where the species did not occur, and the green line represents the classification error for the plots where the species did occur.

For the classification model the results of the trees were aggregated into vote fractions ranging from zero to one. If 120 of 200 trees classified a plot as "present" for the species the vote score of 0.6 would be recorded. Under model defaults the class receiving the majority vote count is the "winning" class. In the case of PA modeling since only two classes are used then 0.5 is the threshold for determining the majority class making the final classification of this plot "present." If a more conservative or liberal predictive ability is desired then the threshold value can manually be raised or lowered. Figure 4 is an example of how one classification tree in a random forest might appear. The top node indicates splitting the data into two groups, one with values for MAPxMTCM < 4910 (see Table 2 and Table 5 for variable definitions). Data split to the branch left of a node satisfy the node's stated condition. In the case of MAPxMTCM < 4910, plots having both cold winters and low annual precipitation values will be split to the left branch. Others will be split to the right branch. If a record did return a true statement for the first step, MAPxMTCM < 4910, then it would follow the left node to the next branch; which is MTCM_MAP < 0.001088. If this statement is also returned as true then the algorithm

branches left to a terminal node that classifies the plot as absent for the species of interest. If the statement for the second branch would have been false then the algorithm takes the right node instead, and then smrsprpb is evaluated and classified in the same true/false or yes/no decision making process. This continues until the algorithm reaches a terminal node within the tree and can make a final decision of present or absent.

The threshold values within each classification tree for the different variables are based off of the training data set. The splits and threshold values within the classification tree, or nodes, are what the algorithm uses to classify the presence or absence of this example based on true/false decision. This process is repeated again creating a different tree with different variables until 200 unique trees are created and 200 decisions of presence or absence have been reached. The aggregated vote count is then calculated and based on either the default value of 0.5 or a user determined threshold the plot is given the final classification of present or absent.

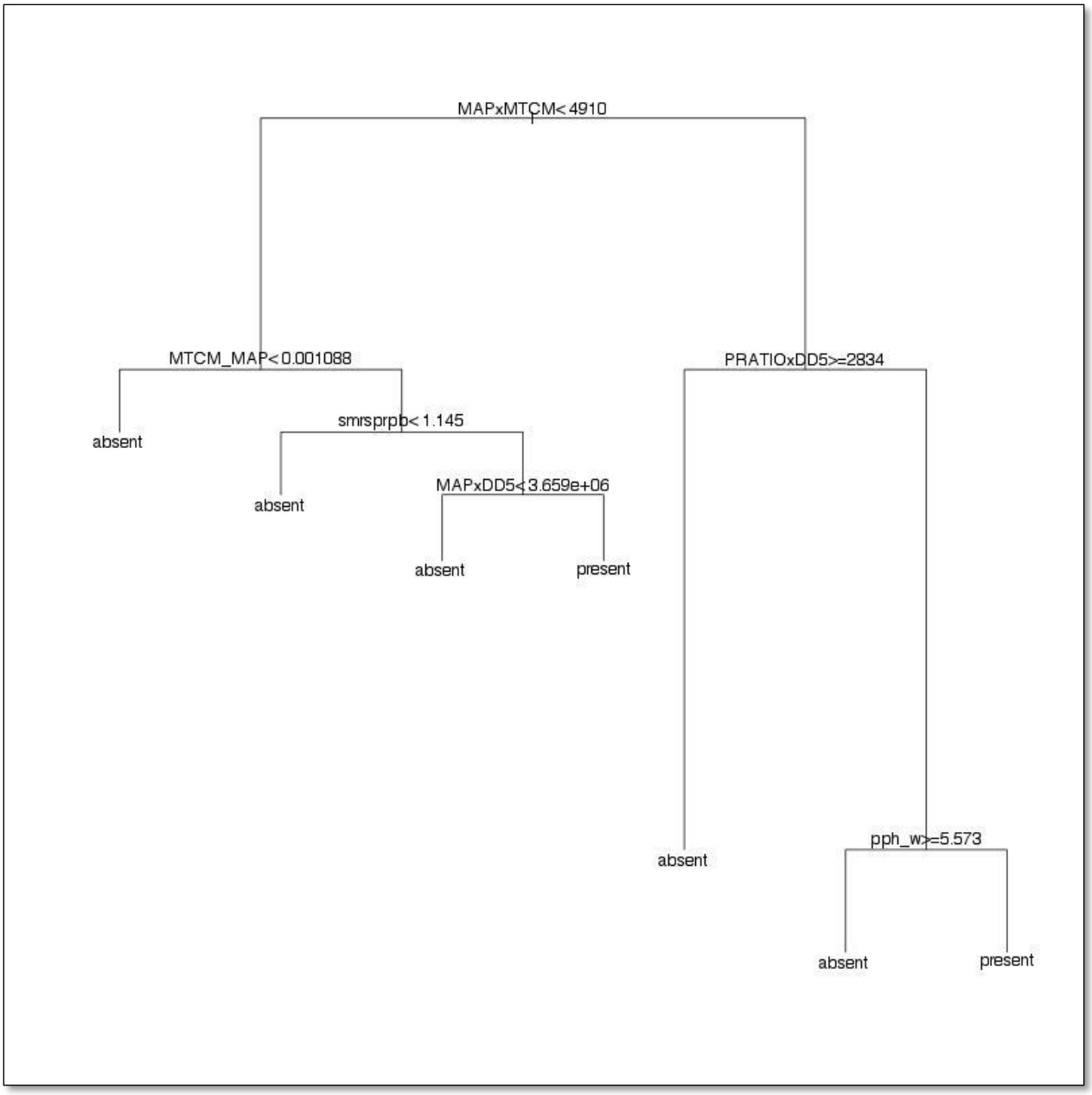


Figure 4. Decision tree example for loblolly pine presence/absence classification.

The IV models that explained the highest percentage of variability was loblolly pine with 43.31% and slash pine at 43.27%, Table 8. Loblolly pine also has the highest mean squared residual 360.65. The IV model that had the lowest percentage of variability explained was longleaf pine at 13.76%, Table 8. Longleaf pine is a species with a low plot count compared to other target species, the only species with a lower plot count was pond pine (Table 6).

Table 8. Random Forest IV model accuracy.

Species	Mean of squared residuals	% Variability Explained	Current Average IV
Shortleaf Pine	30.73	23.85	17.37
Slash Pine	91.31	43.27	54.05
Pond Pine	6.70	17.13	30.29
Longleaf Pine	39.99	13.76	35.35
Loblolly Pine	360.65	43.31	51.23

Slash pine not only produced a IV model with a high percentage of variability explained but it also produced a PA model with the lowest out of bag (OOB) error rate, 6.32% (*Table 9*) . In the case of the IV models the higher the percentage of variability explained the better the model, however in the PA models a lower OOB error rate results in a better model. For example, the PA model for pond pine has a low OOB rate of 7.31%, however the IV model had one of the lowest percent of variably explained.

In the PA models, the classification error describes the error that occurs when a plot condition is misidentified; for example a condition of "present" classified as "absent" or vice-versa. It is desirable for the classification errors for both presence and absence to be balanced so that it is known that the model does not consistently predict for one condition more or less accurately than the other condition.

Common metrics like true positive rates and false positive rates can be calculated. A true positive rate is also called the hit rate or recall and can be estimated as:

$$tp\ rate = \frac{\text{Positives correctly classified}}{\text{Total positives}} \qquad (2)$$

The false positive rate, also known as the false alarm rate is estimated as:

$$fp\ rate = \frac{\text{Negatives incorrectly classified}}{\text{Total negatives}} \qquad (3)$$

Both of these metrics are used in receiver operating curve graphing procedures.

Table 9. Confusion matrices for Random Forest PA model accuracy.

Species		absent	present	OOB (%)	Classification Error
Shortleaf Pine	**absent**	55313	10685	16.06%	0.1618
	present	627	3809		0.1413
Slash Pine	**absent**	63372	4309	6.32%	0.0637
	present	139	2614		0.0505
Pond Pine	**absent**	64989	5128	7.31%	0.0731
	present	18	299		0.0568
Longleaf Pine	**absent**	60781	8206	11.79%	0.1189
	present	100	1347		0.0691
Loblolly Pine	**absent**	49513	6612	10.76%	0.1178
	present	969	13340		0.0677

An ROC graph helps the user visually compare the relative tradeoff between benefits (true positives) and costs (false positives) Figure 5. A ROC graph plots the proportions of true positive to false positive predictions that would be realized in a training data set given the range of possible prediction thresholds (T) from [0,1] the true positive rate is plotted on the Y axis and the false positive rate is plotted on the X axis. The prediction thresholds for our models (T) were determined using these ROC graphs based on the point in the curves at which the errors were minimized. The ROC curve is a depiction of classifier performance. Area under the ROC curve (AUC) is a value always between 0 and 1 that is a portion of the area of the unit square. Random guessing produces a diagonal line between (0, 0) and (1, 1) which has an area of 0.5, therefor no realistic classifier will have an AUC less than 0.5. The AUC is an equivalent to the Wilcoxon test of ranks because the AUC of a classifier is equivalent to the probability that the classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance (Fawcett 2006). Of the species studied, the classification model for shortleaf pine PA had the lowest AUC = 0.92 (Figure 5). In comparison, the classification for slash pine PA had the highest AUC value (0.98) (Figure 6). ROC curves for longleaf, loblolly, and pond pine can be found in the appendix.

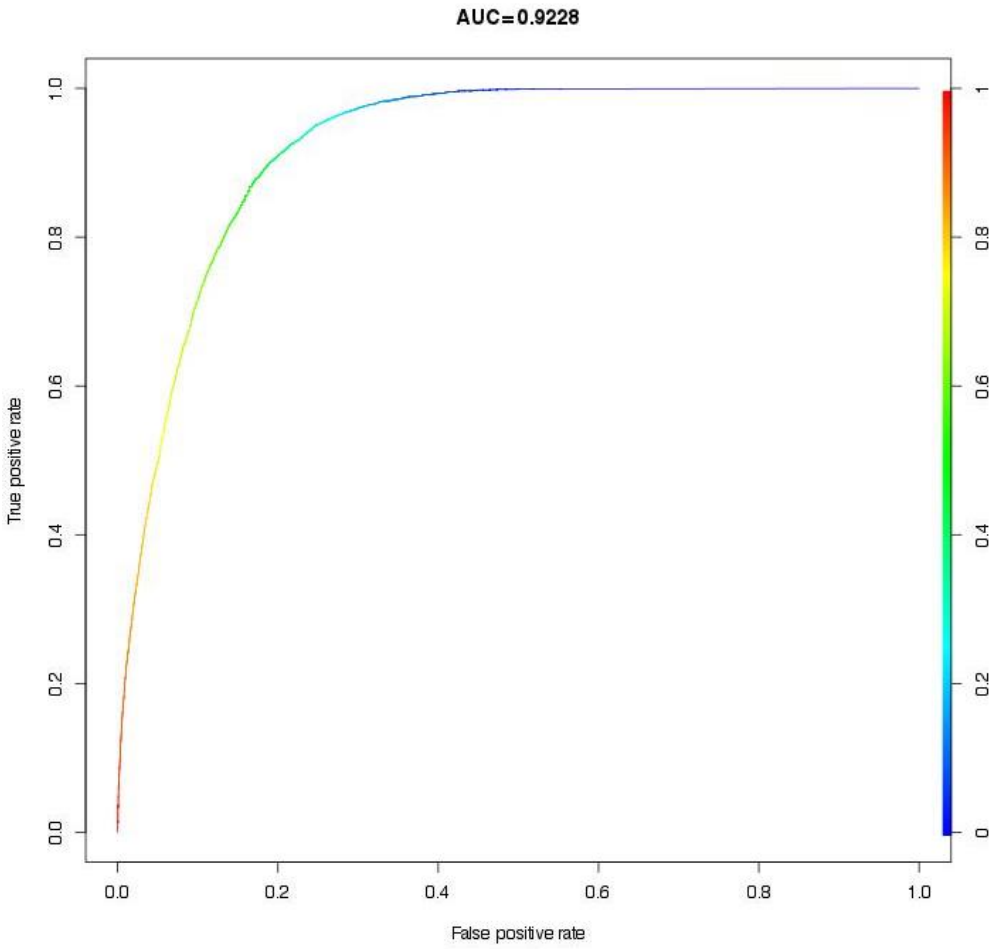


Figure 5. ROC graph for shortleaf pine PA model showing area under the curve, true positive rate (left y-axis), false positive rate (x-axis), and vote count ratio (right y-axis)

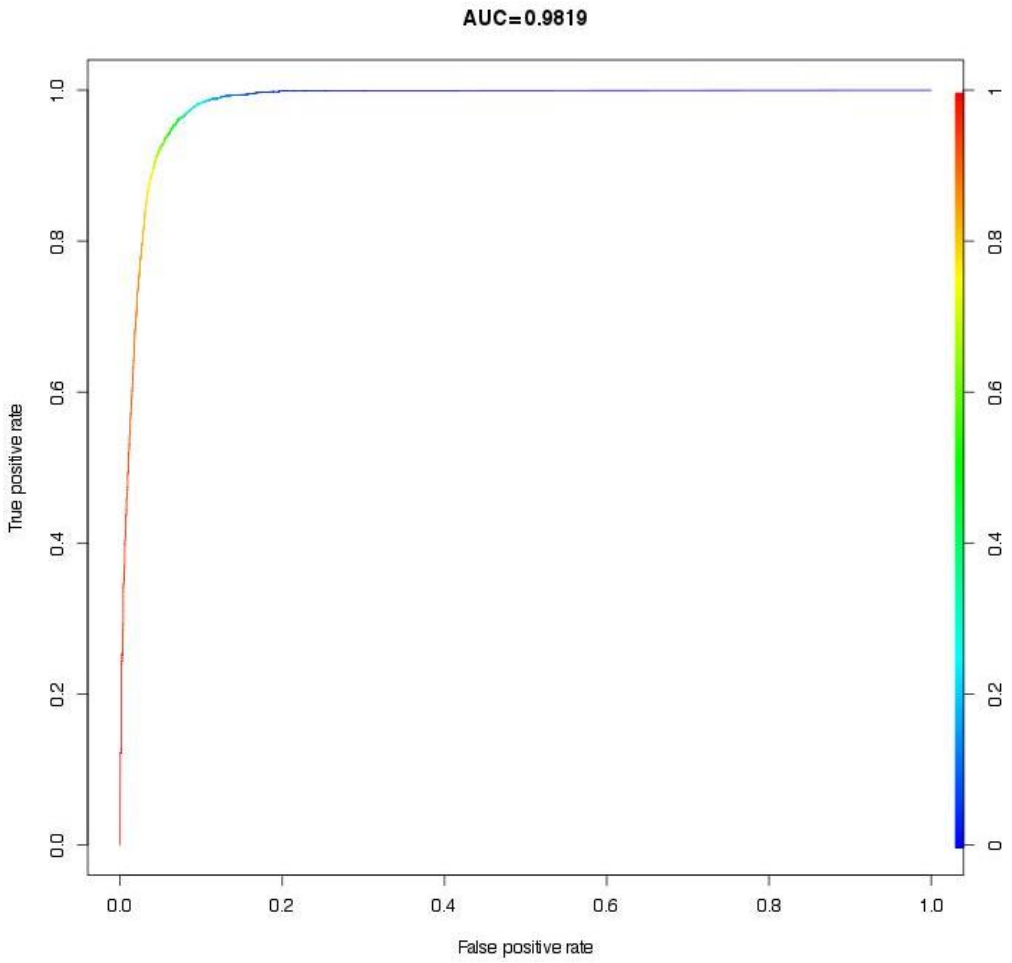


Figure 6. ROC graph for slash pine PA model showing area under the curve, true positive rate (left y-axis), false positive rate (x-axis), and vote count ratio (right y-axis)

Variable Importance

Once the models were developed, the importance variable plots within randomForest were displayed to see if there were any similarities/dissimilarities between species or across modeling techniques. Variable importance plots are dot-charts of variable importance as measured by a Random Forest (Figure 7-Figure 11) (Liaw and Wiener 2002). For regression IV model importance plots, column A is the mean decrease in accuracy. For the PA model importance plots, column B is the percent increase in MSE (%IncMSE). Both columns A for IV and column B for PA represent the decrease of the predictive ability of the model when a predictor is replaced by random noise. A higher value indicates greater variable importance (Vincenzi et al. 2011).

For the shortleaf pine models the soils variable pNO10_w (percent soil passing sieve no. 10) is the most important variable in both IV and PA modeling techniques, as determined by the %IncMSE and MeanDecreaseAccuracy columns (Figure 7). Shortleaf pine IV models selected soils variables as among the most important in 7 out of the 10 predictors. Shortleaf pine PA models selected climate variables as among the most important in 9 out of 10 predictors (Figure 7). Similarly in Figure 8 and Figure 9 for slash pine and longleaf pine the IV models determined that predominately soils variables were important for model classification while the PA models used predominately climate variables. Pond pine IV and PA models selected predominately climate variables as being among the most important variables (Figure 10). Loblolly pine IV and PA models on the other hand selected predominately soils variables as being among the most important variables (Figure 11). The loblolly pine IV and PA models have seven common variables in the top ten, more than any of the species IV/PA models (Figure 11).

Twenty different predictors ranked among the top 10 most important variables over the five species studied. Of the twenty, eleven described soils attributes and nine were climate variables. Soil bulk density (pbd_w) was ranked as a top 10 predictor for all IV models. Three other most commonly occurring IV models variables were, pawc_ w, pkffact_w, and pph_w (see Table 2 for definitions). Models of species PA showed greater variation in the selection of important predictors, with 24 different climate variables and 9 soils attributes ranking in the top 10 of one or more models. Other most commonly occurring predictor variables for the PA models were dd0, MAPxMTCM , mmin , and MTCM_MAP each appearing in three out of the five species models.

The two most frequently occurring variables across both PA and IV models were soil bulk density (pbd_w) and available water holding capacity (pawc_w) each appearing in the top 10 most important variables for six different models (Table 10). All 11 soils variables are found in the top 10 most important variables for at least 3 IV or PA models while only 8 climate variables are found in at least 3 IV or PA models (Table 10).

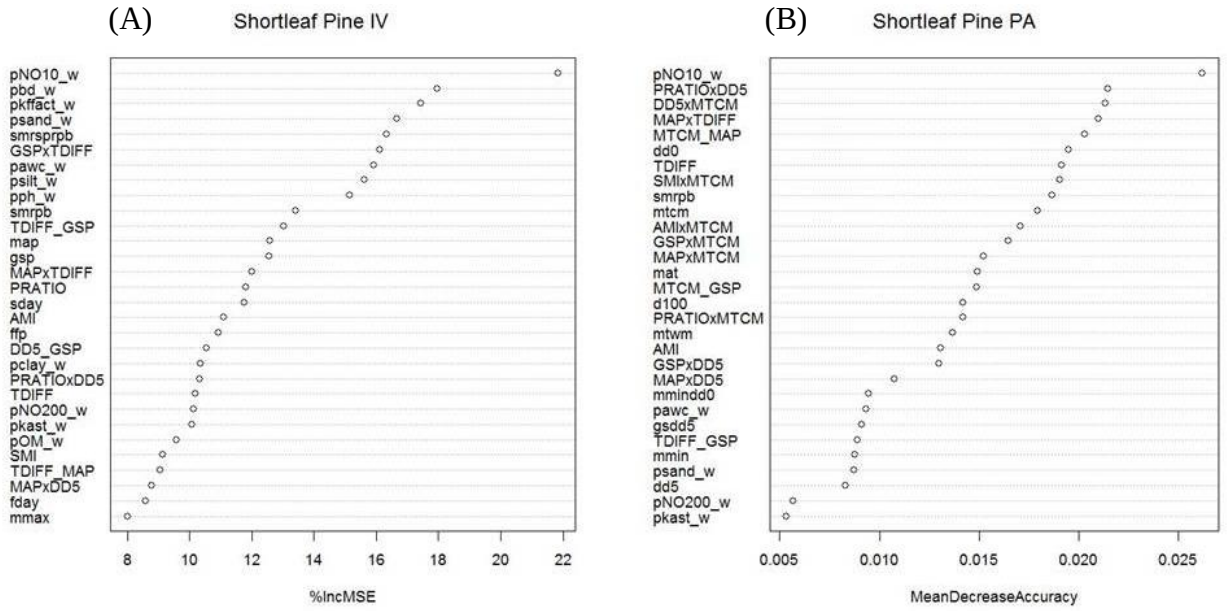


Figure 7. Shortleaf Pine. Variables most important to IV model predictions (A), Variables most important to PA model predictions (B). Variables ending with "_w" indicate soils variables.

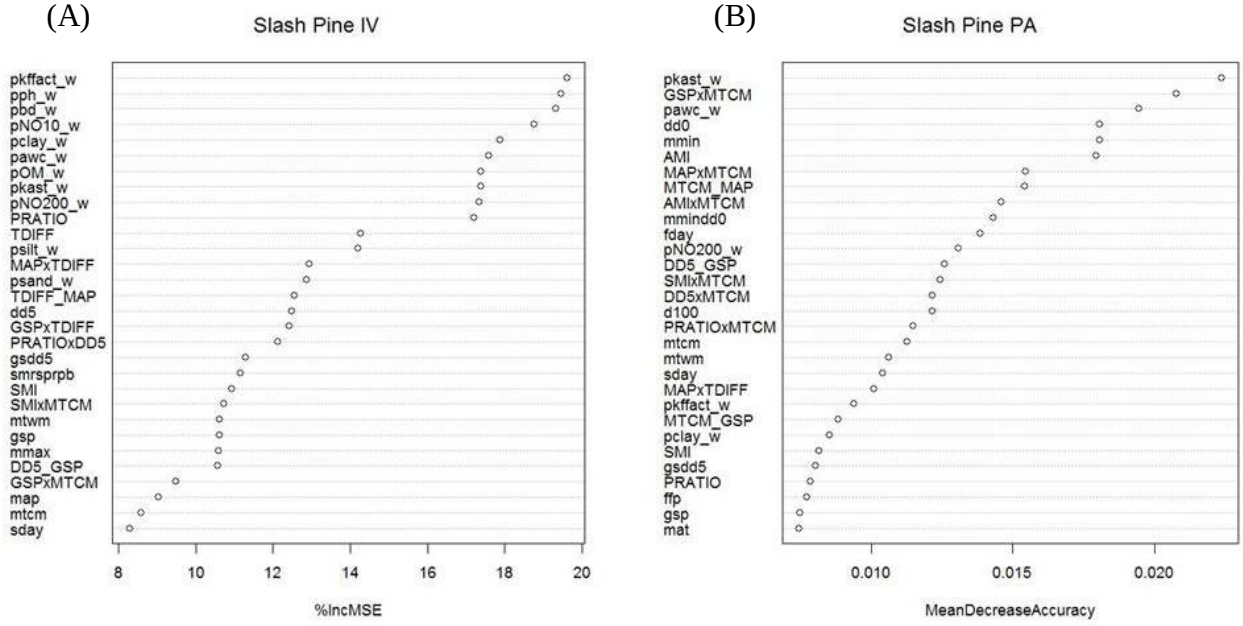


Figure 8. Slash Pine. Variables most important to IV model predictions (A), Variables most important to PA model predictions (B). Variables ending with "_w" indicate soils variables.

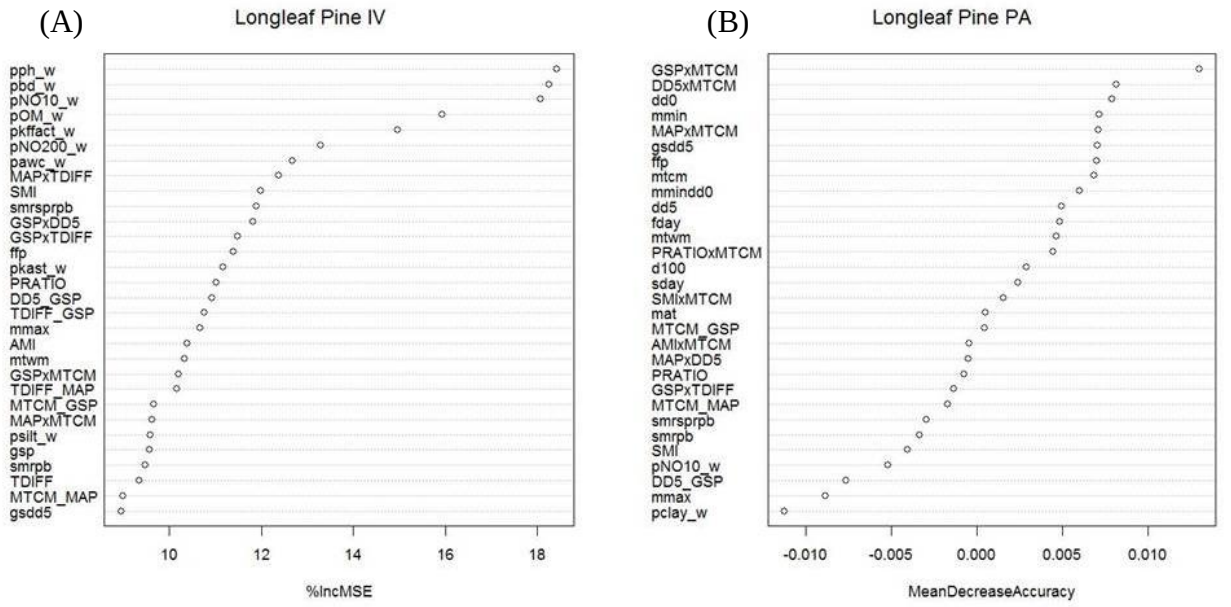


Figure 9. Longleaf Pine. Variables most important to IV model predictions (A), Variables most important to PA model predictions (B). Variables ending with "_w" indicate soils variables.

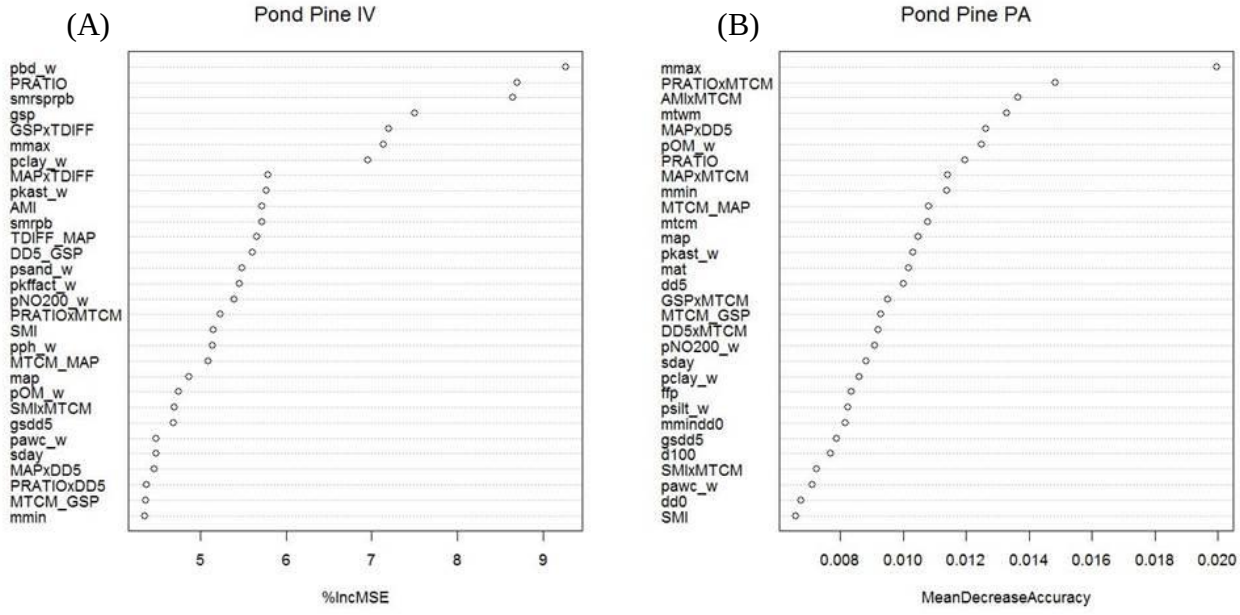


Figure 10. Pond Pine. Variables most important to IV model predictions (A), Variables most important to PA model predictions (B). Variables ending with "_w" indicate soils variables.

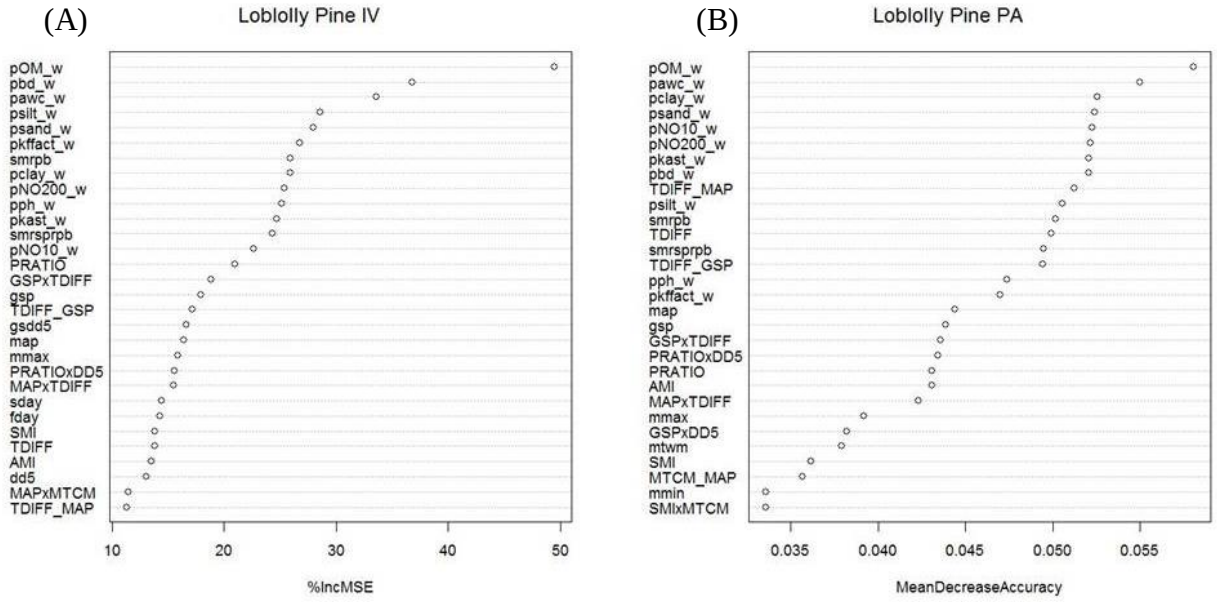


Figure 11. Loblolly pine. Variables most important to IV model predictions (A), Variables most important to PA model predictions (B). Variables ending with "_w" indicate soils variables.

Table 10. Selected climate and soils variables and their rank (1-10) of %incMSE and MeanDecreaseAccuracy by species and IV/PA model. Variables listed are ranked among the top ten most important variables for at least 3 out 10 models.

		Shortleaf Pine		Slash Pine		Longleaf Pine		Pond Pine		Loblolly Pine	
	Variables	IV	PA	IV	PA	IV	PA	IV	PA	IV	PA
Climate	dd0		6		4		3				
	MAPxMTCM				7		5		8		
	MAPxTDIFF		4			8		8			
	mmin				5		4		9		
	MTCM_MAP		5		8				10		
	PRATIO			10				2	7		
	smrpb	10	9							7	
	smrsprpb	5				10		3			
Soils	pawc_w	7		6	3	7				3	2
	pbd_w	2		3		2		1		2	8
	pclay_w			5				7		8	3
	pkast_w			8	1			9			7
	pkffact_w	3		1		5				6	
	pNO10_w	1	1	4		3					5
	pNO200_w			9		6				9	6
	pOM_w			7		4			6	1	1
	pph_w	9		2		1				10	
	psand_w	4								5	4
	psilt_w	8								4	10

Once the model reliability scores, OOB, and percent of variability explained had been explored and the important variables for model classification identified, the models were used to predict for the current range of each species to see how well they matched the observed data. The predict function in Random Forest was used for these calculations (Liaw and Wiener 2002). For IV mapping the methods followed those of Prasad et al. (2007-ongoing) online database called Tree Atlas. Only plots with a modeled importance value of 1% or greater were mapped with the same class intervals used by (Prasad et al. 2007-ongoing) in their maps. The mapping for the modeled PA was done by setting the threshold value of the vote count for determining presence to 0.05.

For shortleaf pine the current FIA plots that have the highest IV are located in the northwestern extent of its range (Figure 12 C). The Random Forest model in Figure 12 D

was able to predict this higher occurrence of IV in the northwestern portion of its range. The modeled PA for shortleaf pine uses T= 0.43 (Figure 12 B). The slash pine IV and PA current maps do not line up with Little's range map; there is a pocket of slash pine in western Louisiana and eastern Texas that was not previously mapped (Figure 8 A, C). This pocket of Slash pine does appear in IV and PA models (Figure 8 B, D). The slash pine in this area outside of the mapped Little's range are predominately categorized as artificial regeneration so this area is probably pine plantations that were planted after the original range map was created. Slash pine PA was modeled at a T = 0.43 (Figure 13) and longleaf pine was modeled at T = 0.42 (Figure 14). Current pond pine PA and IV are widely dispersed and infrequent within the original range map made by Little (Figure 15 A, C). This same pattern is reflected by the modeled IV, however the modeled PA at T=0.49 shows many additional plots as having potential to support pond pine where no pond pine is currently found (Figure 15). Currently loblolly pine has a large range and has high importance values throughout its range (Figure 16 A, C). This trend of high importance values throughout the range can be seen in the IV model map (Figure 16 D).

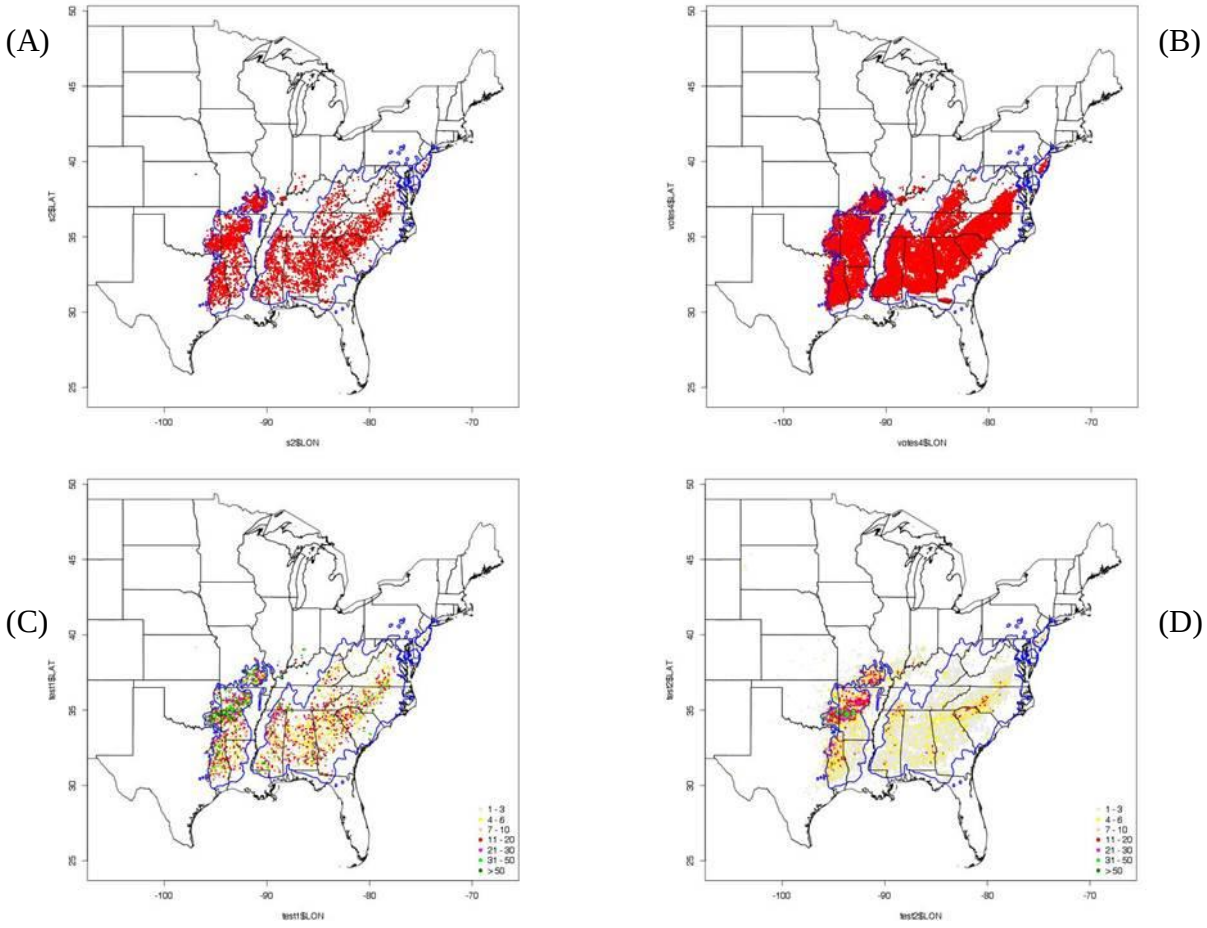


Figure 12. Shortleaf pine maps against Little's range map (1971). FIA plots where shortleaf pine are currently found (A). OOB model predictions for shortleaf pine using a 0.43 threshold (B). Importance values of plots where shortleaf pine are currently found (C). OOB model importance values for shortleaf pine (D).

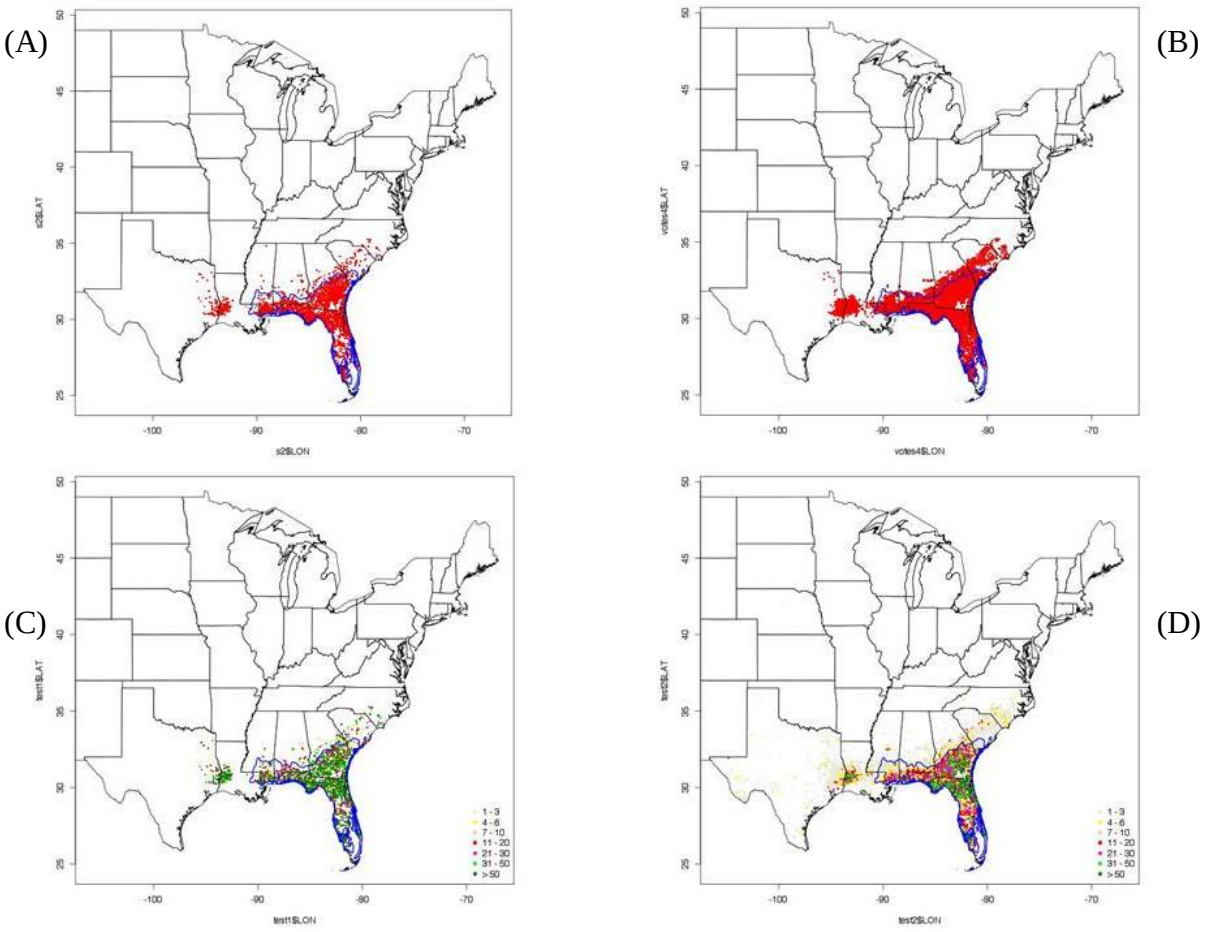


Figure 13. Slash pine maps against Little's range map (1971). (A) FIA plots where Sslash pine are currently found. (B) OOB model predictions for slash pine using a 0.43 threshold. (C) Importance values of plots where slash pine are currently found. (D) OOB model importance values for slash pine.

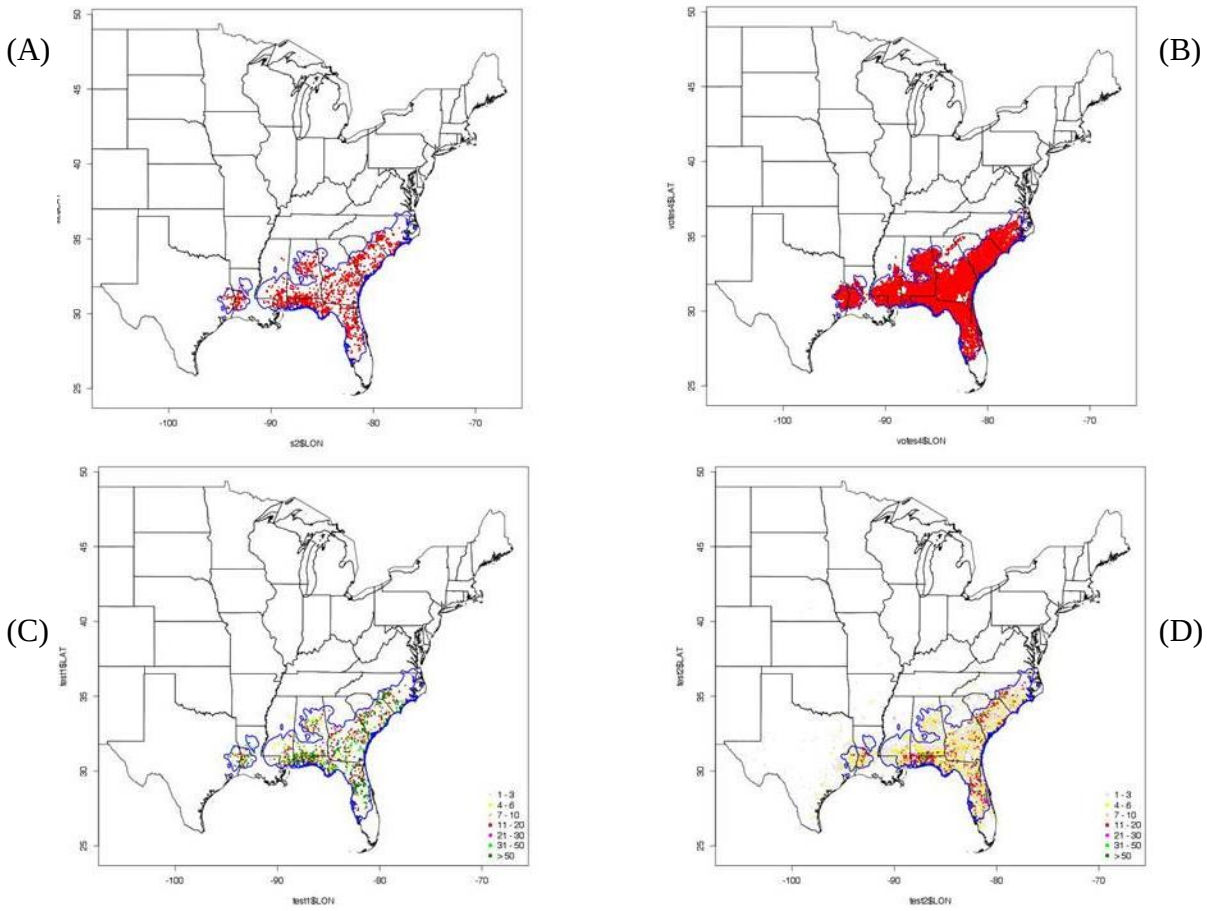


Figure 14. Longleaf pine maps against Little's range map (1971). (A) FIA plots where longleaf pine are currently found. (B) OOB model predictions for longleaf pine using a 0.43 threshold. (C) Importance values of plots where longleaf pine are currently found. (D) OOB model importance values for longleaf pine.

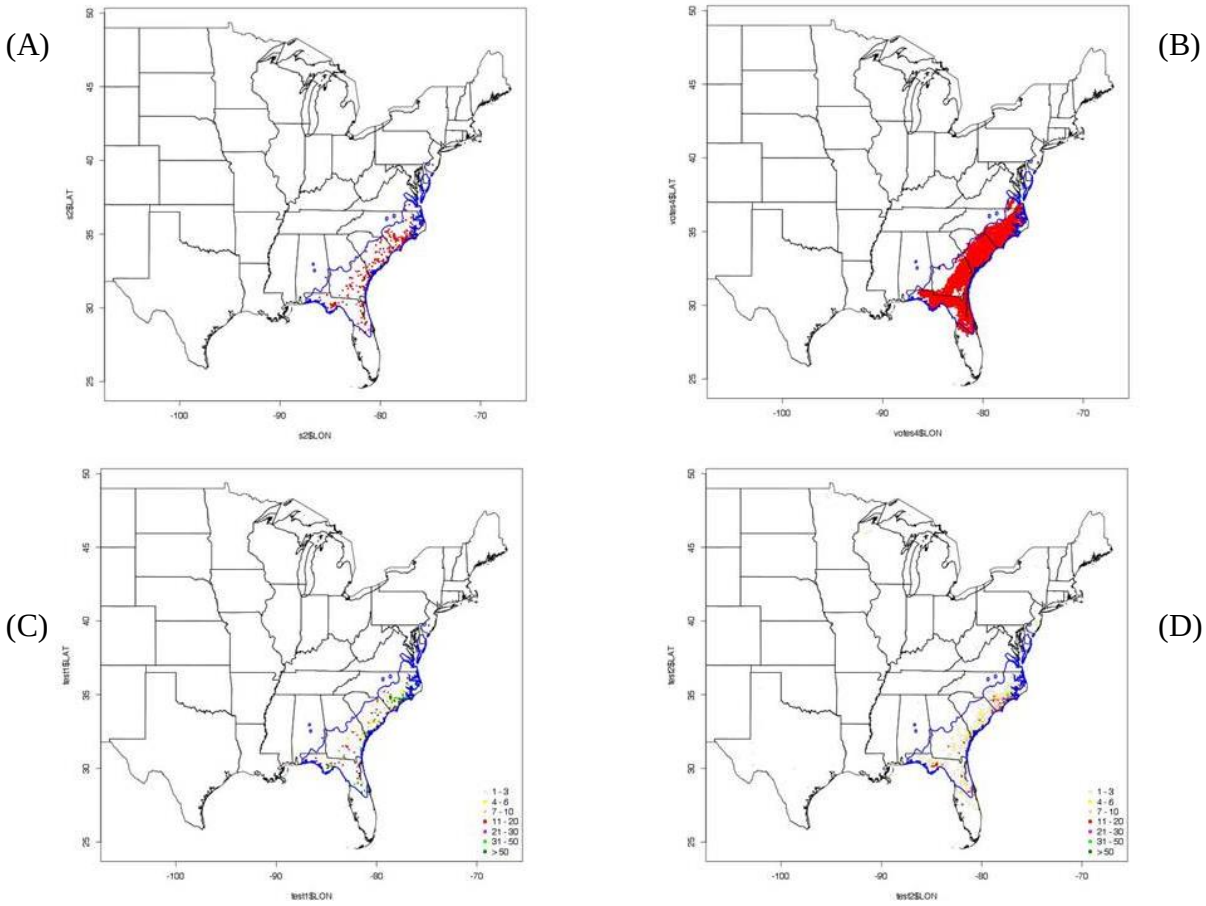


Figure 15. Pond pine maps against Little's range map (1971). (A) FIA plots where pond pine are currently found. (B) OOB model predictions for pond pine using a 0.43 threshold. (C) Importance values of plots where pond pine are currently found. (D) OOB model importance values for pond pine.

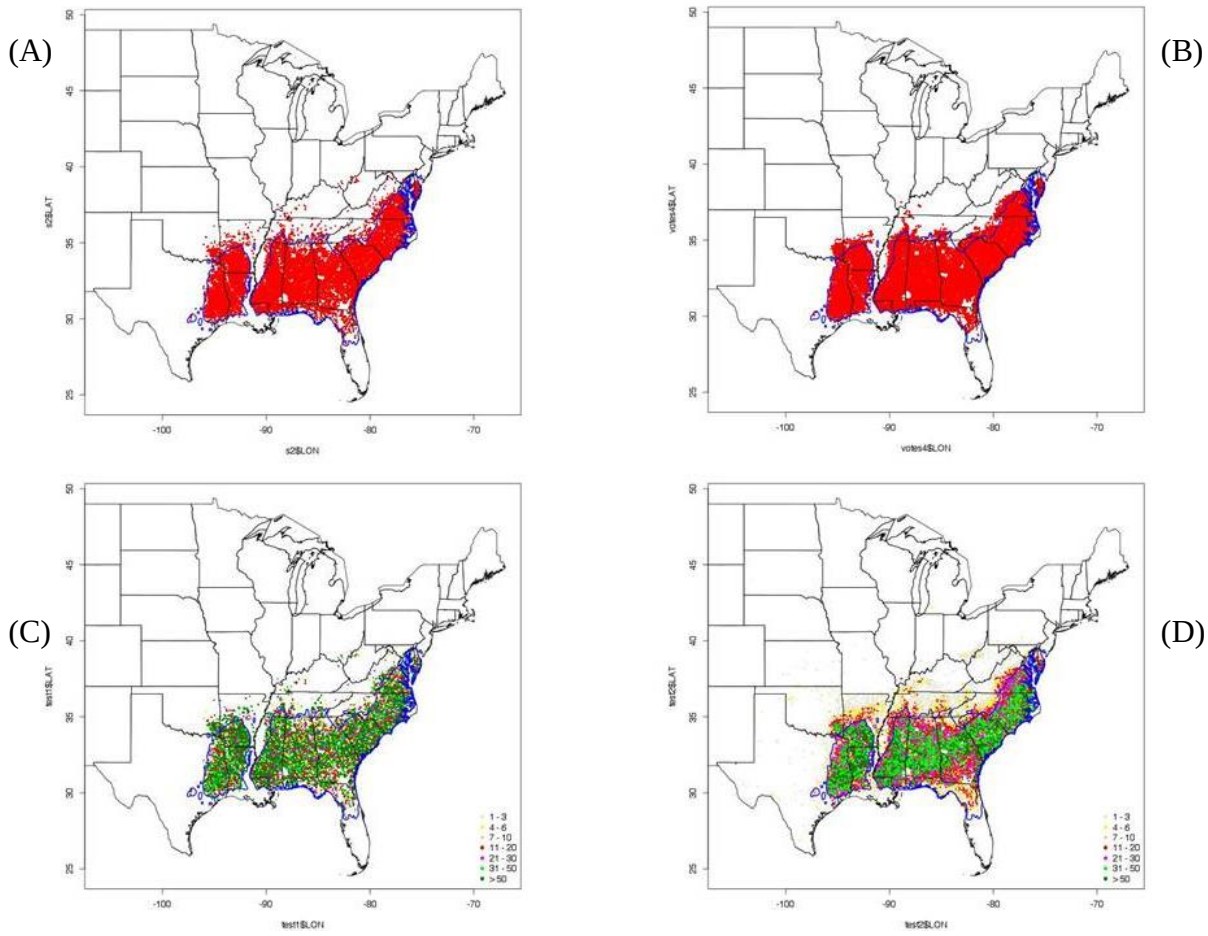


Figure 16. Loblolly pine maps against Little's range map (1971). (A) FIA plots where loblolly pine are currently found. (B) OOB model predictions for loblolly pine using a 0.43 threshold. (C) Importance values of plots where loblolly pine are currently found. (D) OOB model importance values for loblolly pine.

While ROC analysis was used to help determine the optimal threshold (T) for contemporary climate the future predictions T was lowered to 0.05. Lowering the vote count threshold to 0.05 for determining presence can be viewed in a hypothesis test scenario. By setting the threshold of the vote count to 0.05 it is like setting an alpha of 0.05 for a hypothesis test. The null hypothesis being that the target species is present. If 95% or more of classification trees predict the species is absent, this constitutes strong evidence ($p \leq 0.05$) to reject the null hypothesis. This hypothesis puts more emphasis on eliminating false negatives so that management decisions can be based on knowing where the species is highly unlikely to be found.

Model Predictions

After evaluating the ten PA and IV models trained to predict the contemporary range of each species, the models were then used to make predictions for future climate scenarios over the 21st century (Table 3). Soil properties were assumed to remain constant based on contemporary soils maps, and each species IV and PA were predicted for multiple climate change scenarios for decades: 2030, 2060, and 2090. Results were mapped in the same way as the model predictions of the contemporary ranges (Figure 13 - Figure 16). Additional results for the full set of scenarios are shown in the Appendix.

Using longleaf pine as an example, the general trend for IV models is that as time progresses, the areas where abundance is currently high will remain as high abundance areas (Figure 17 - Figure 23). However, plots along the edges of the current ranges have the potential to increase in abundance as the edge of the potential habitat expands to new areas where previously the abundance score was <1%. In the PA modeling the general trend is that as time progresses forward in the model the potential habitat range of longleaf pine is increasing (Figure 24 - Figure 30).

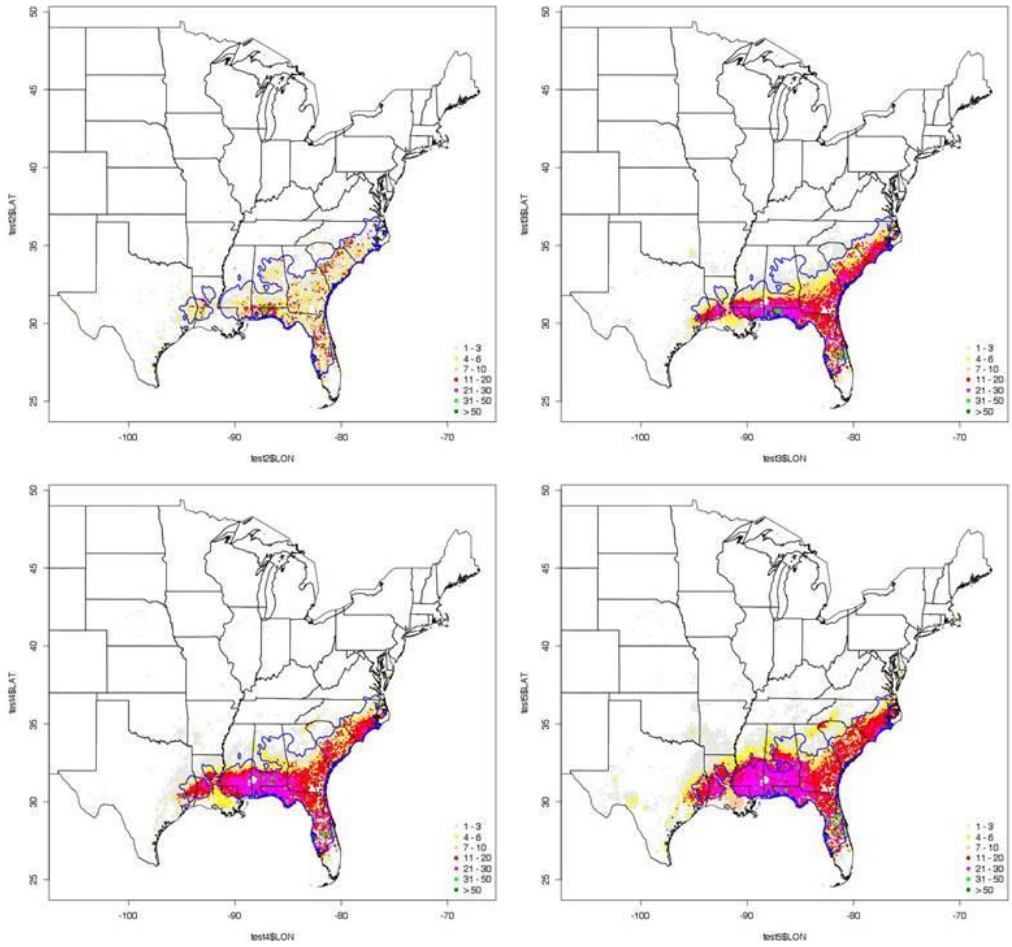
In the lowest CO_2 emission scenario CGCM3 B1 for longleaf pine the predicted habitat range is expanding from the predicted 8,208 plots in the contemporary climate to 21,375 plots in 2090 (Figure 17A, D). In a medium emission scenario CGCM3 A1B the number of plots increases slightly from CGCM3 B1 (Figure 18 D). In 2090 the predicted number of plots raises to 22,253. In the highest CO_2 emission scenario CGCM3 A2 the range of potential habitat is expanding even further (Figure 19 D). In the CGCM3 A2 scenario there are 28,102 plots predicted to have an IV $\geq$ 1, in 2090. The difference between the highest scenario, CGCM3 A2, and lowest scenario, CGCM3 B1 is 6,727 plots.

A similar pattern can be seen in the CGCM3 B1, CGCM3 A1B, and CGCM3 A2 PA models where the highest emission scenario has the greatest number of plots mapped as potential habitat (Figure 24 - Figure 26). The contemporary climate model for PA with a T = 0.42 predicts that 10,847 plots currently have conditions suitable for longleaf pine to be present, while the CGCM3 B1 climate scenario with a T=0.05 predicts that the number of plots to increase to 28,000 in 2090 (Figure 24 A, D) . The CGCM3 A1B scenario predicts that in 2090 there will be 25,982 plots suitable for presence of longleaf pine (Figure 25 D).

The highest emission scenario CGCM3 A2 predicts that in 2090 the number of plots consistent with longleaf pine presence is 30,735 plots (Figure 26 D).

Maps of GCM-derived climate variables identify causal relationships between climate variables and future prediction of the species PA and abundance. Since soils variables were held constant over all time periods and scenarios the climate variables are the only explanatory variables capable of driving future vegetation predictions. For longleaf pine IV, MAPxTDIFF, SMI, smrsprpb were the highest ranking climate predictors (Figure 9 A). MAPxTDIFF is a precipitation and summer-winter temperature differential interaction term, SMI is a summer precipitation index based on the number of degree days above 5° C accumulating within the frost free period divided by the growing season precipitation (April-September), and smrsprpb is the summer /spring precipitation balance (Table 5).

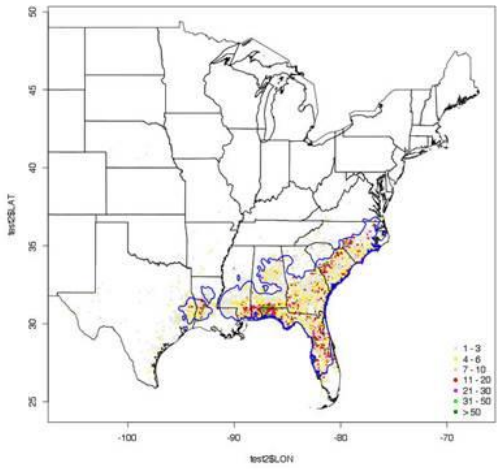
(A) (B)



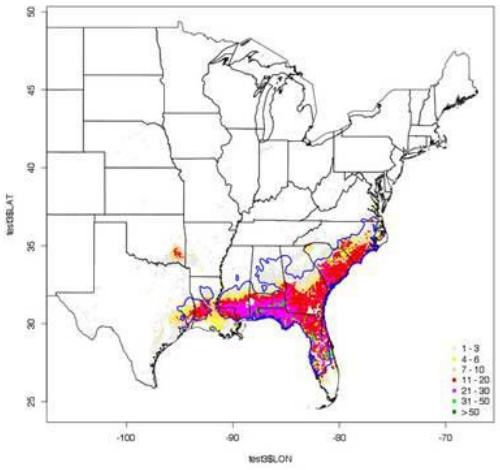
(C) (D)

Figure 17. IV Longleaf pine CGCM3 B1 maps. (A) Longleaf pine OOB base model map, (B) year 2030 predictions for longleaf pine under CGCM3 B1 scenario, (C) year 2060 predictions for longleaf pine under CGCM3 B1 scenario, (D) year 2090 predictions for longleaf pine under CGCM3B1 scenario.

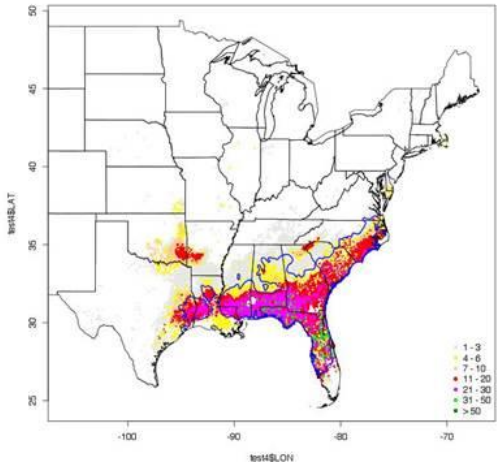
(A)



(B)



(C)



(D)

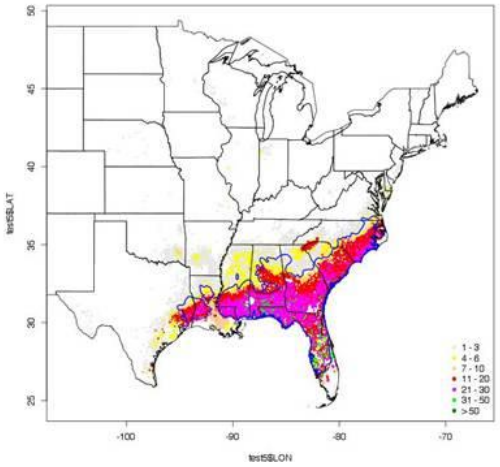
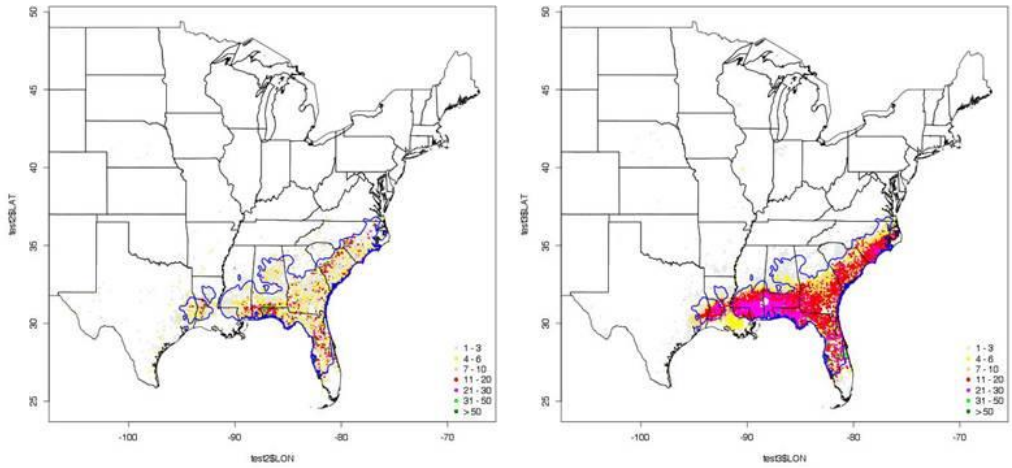


Figure 18. IV Longleaf pine CGCM3 A1B maps. (A) Longleaf pine OOB base model map, (B) year 2030 predictions for longleaf pine under CGCM3 A1B scenario, (C) year 2060 predictions for longleaf pine under CGCM3 A1B scenario, (D) year 2090 predictions for longleaf pine under HADCM3 A2 scenario.

(A) (B)



(C) (D)

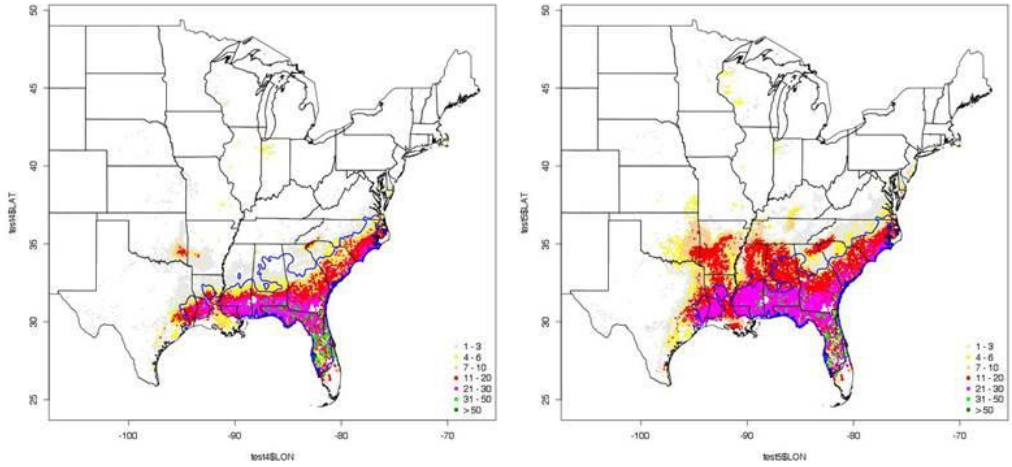
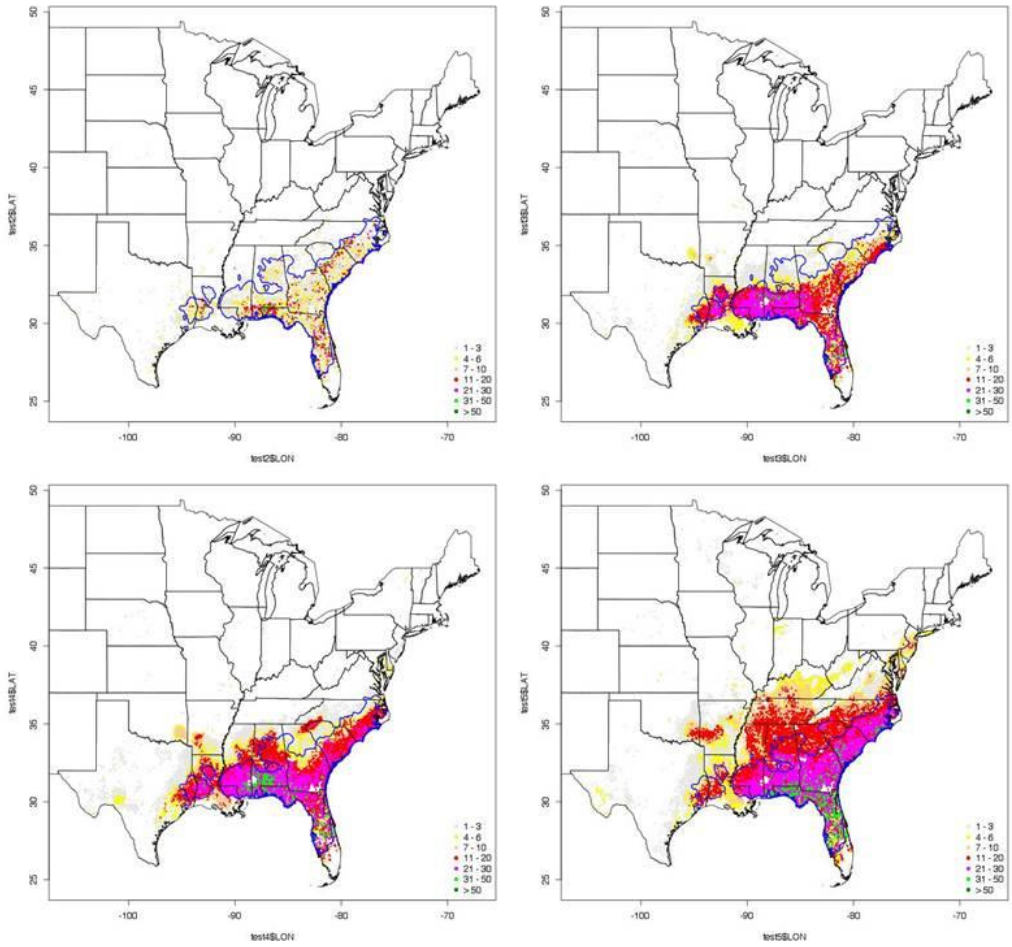


Figure 19. IV Longleaf pine CGCM3 A2 maps. (A) Longleaf pine OOB base model map, (B) year 2030 predictions for Longleaf pine under CGCM3 A2 scenario, (C) year 2060 predictions for Longleaf pine under CGCM3 A2 scenario, (D) year 2090 predictions for Longleaf pine under HADCM3 A2 scenario.

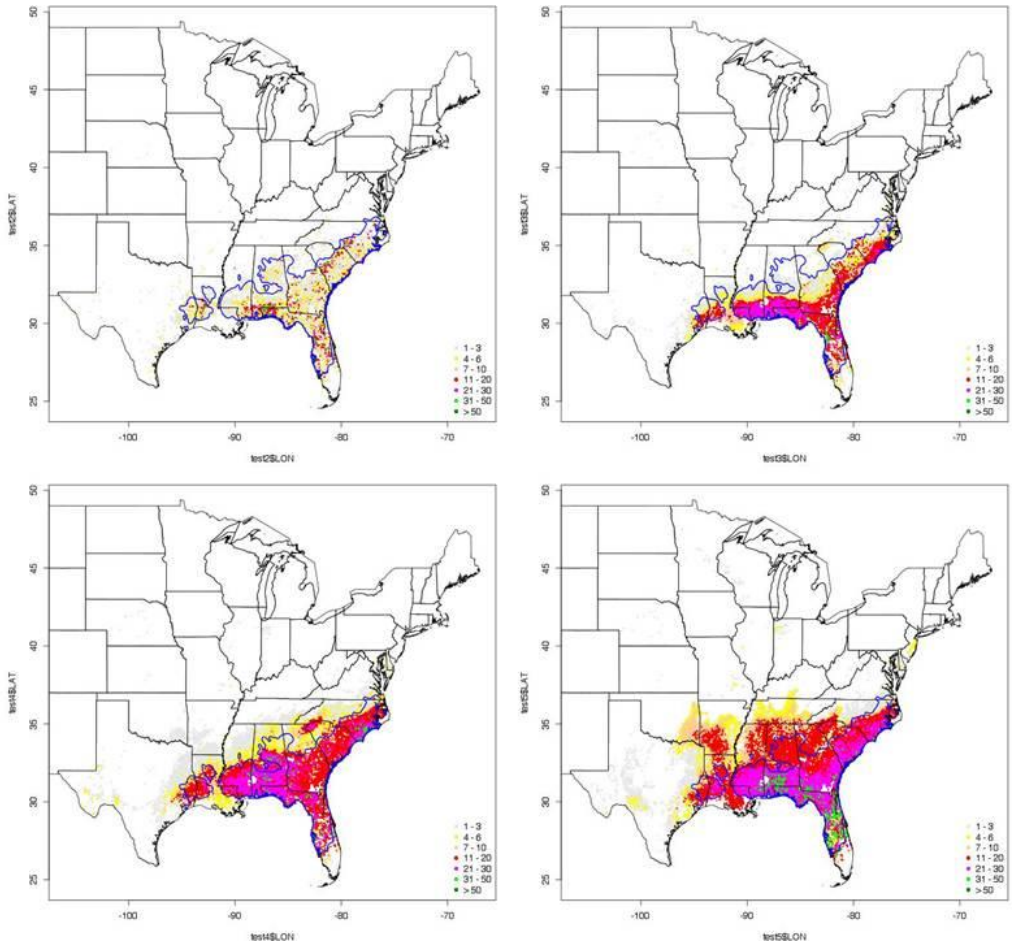
(A) (B)



(C) (D)

Figure 20. IV Longleaf pine HADCM3 A2 maps. (A) Longleaf pine OOB base model map, (B) year 2030 predictions for longleaf pine under HADCM3 A2 scenario, (C) year 2060 predictions for longleaf pine under HADCM3 A2 scenario, (D) year 2090 predictions for longleaf pine under HADCM3 A2 scenario (D)

(A) (B)



(C) (D)

Figure 21. IV Longleaf pine HADCM3 B2 maps. (A) Longleaf pine OOB base model map, (B) year 2030 predictions for longleaf pine under HADCM3 B2 scenario, (C) year 2060 predictions for longleaf pine under HADCM3 B2 scenario, (D) year 2090 predictions for longleaf pine under HADCM3 A2 scenario.

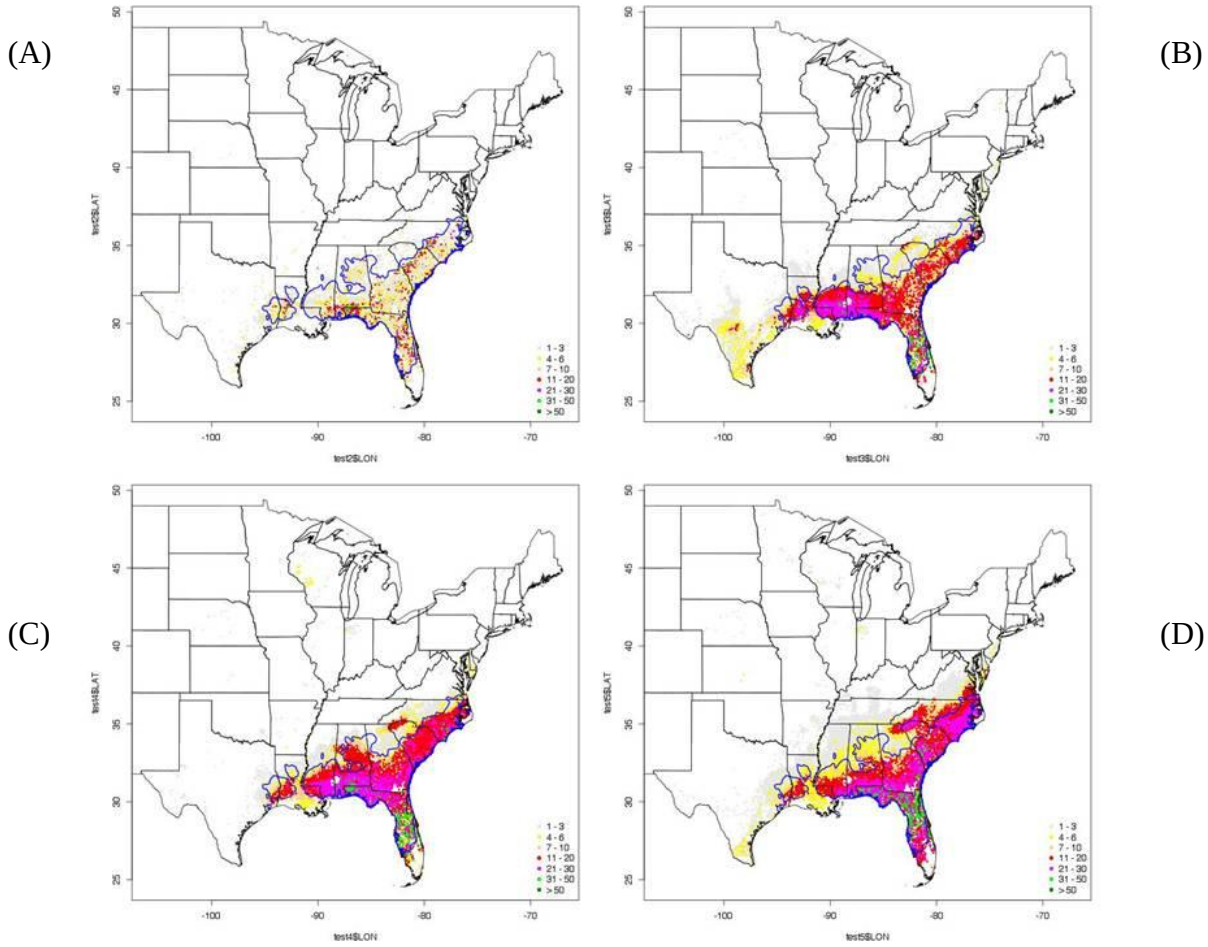
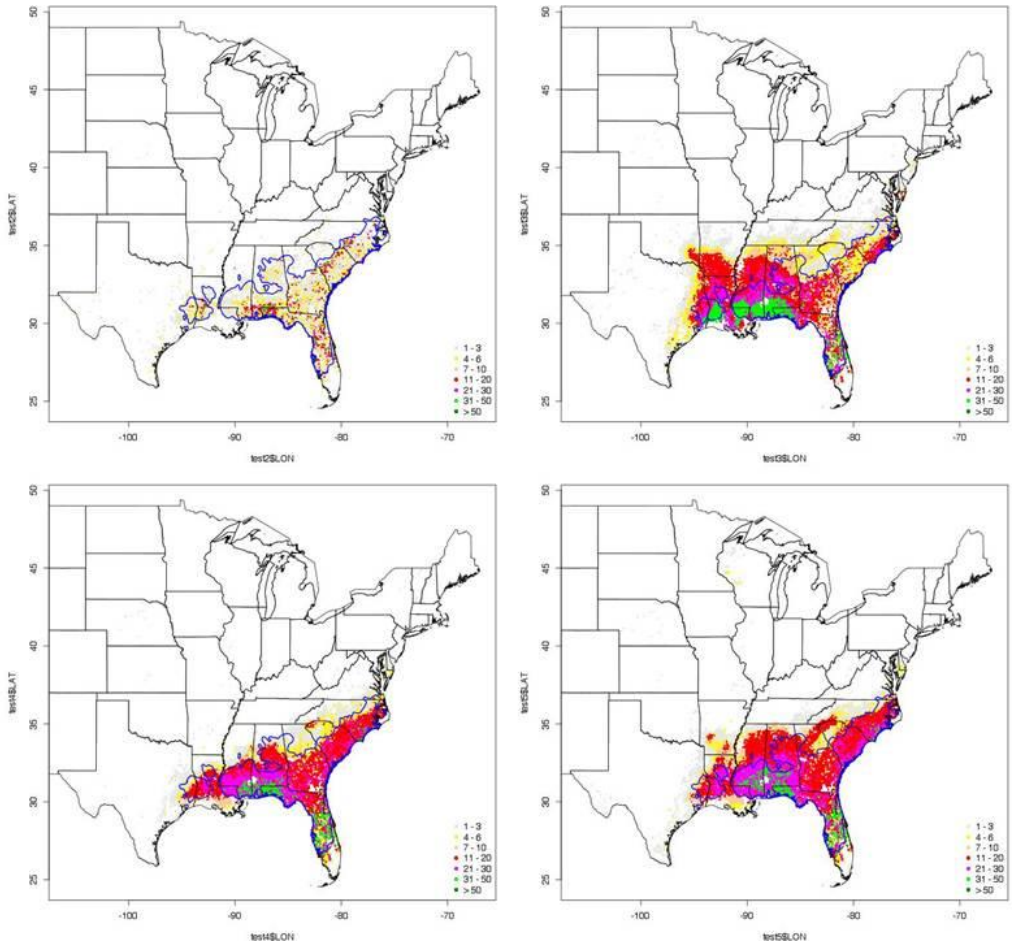


Figure 22. IV Longleaf pine GFDLCM21 A2 maps. (A) Longleaf pine OOB base model map, (B) year 2030 predictions for longleaf pine under GFDLCM21 A2 scenario, (C) year 2060 predictions for longleaf pine under GFDLCM21 A2 scenario, (D) year 2090 predictions for longleaf pine under GFDLCM21 A2 scenario.

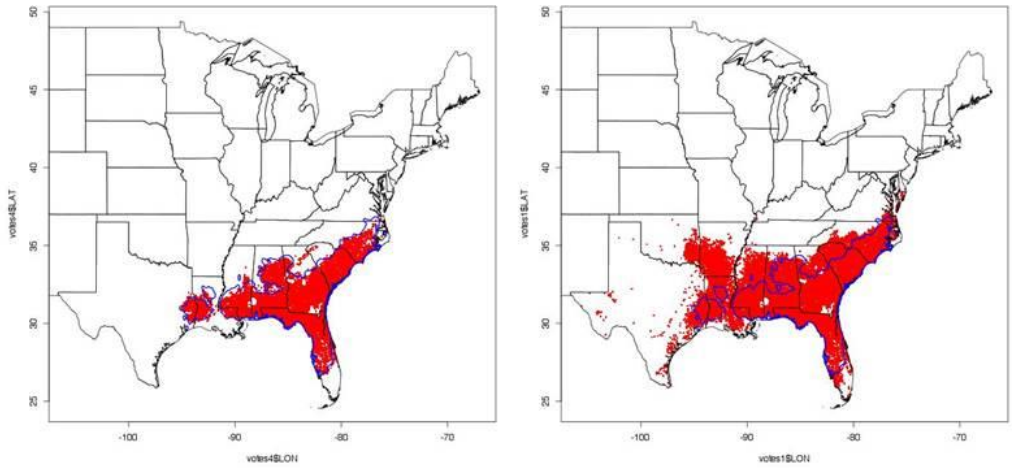
(A) (B)



(C) (D)

Figure 23. IV Longleaf pine GFDLCM21 B1 maps. (A) Longleaf pine OOB base model map, (B) year 2030 predictions for longleaf pine under GFDLCM21 B1 scenario, (C) year 2060 predictions for longleaf pine under GFDLCM21 B1 scenario, (D) year 2090 predictions for longleaf pine under GFDLCM21 B1 scenario.

(A) (B)



(C) (D)

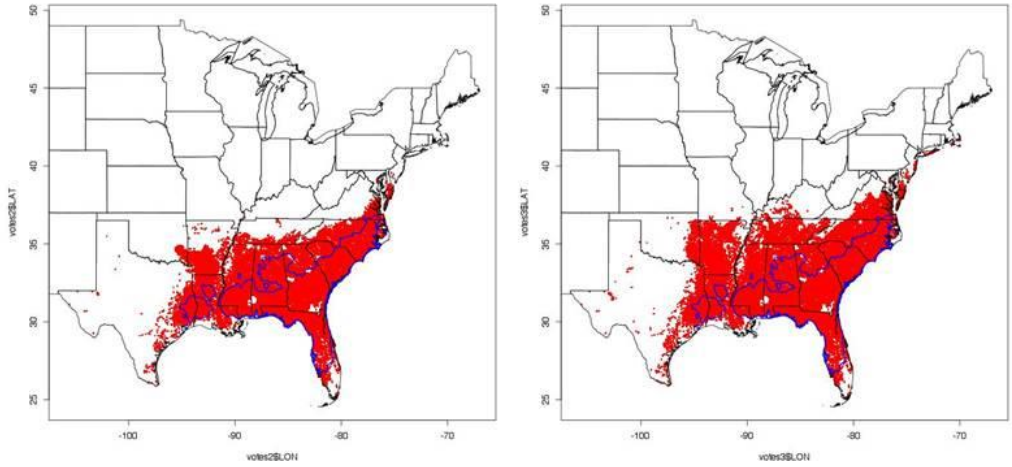
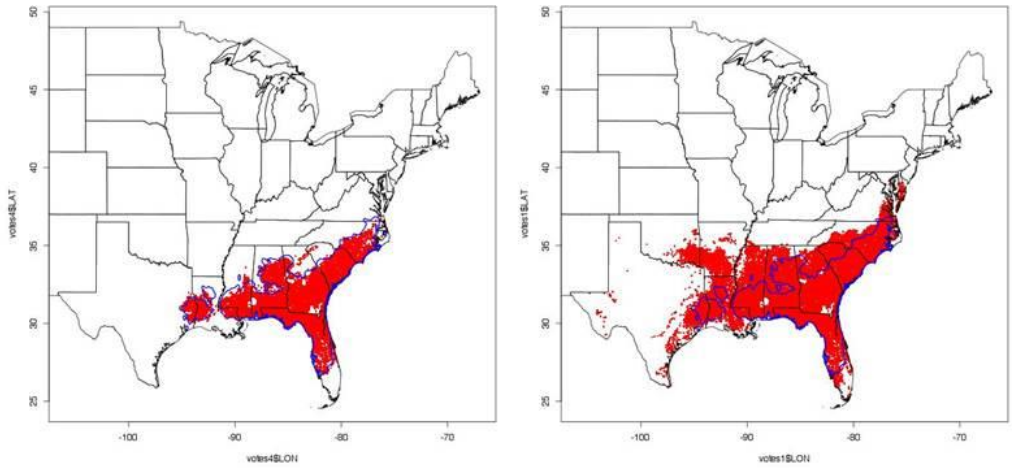


Figure 24. PA Longleaf pine CGCM3 B1 maps. (A) Longleaf pine OOB base model map, (B) year 2030 predictions for longleaf pine under CGCM3 B1 scenario, (C) year 2060 predictions for longleaf pine under CGCM3 B1 scenario, (D) year 2090 predictions for longleaf pine under CGCM3 B1 scenario.

(A) (B)



(C) (D)

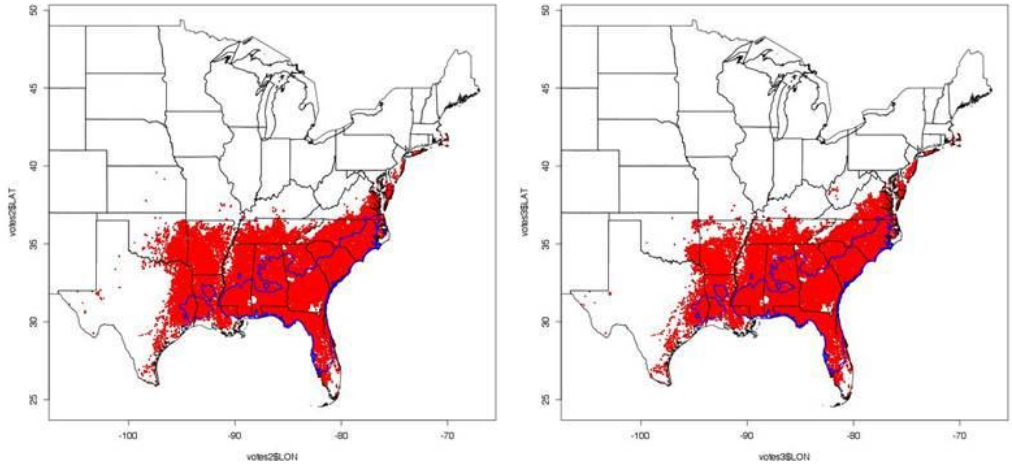
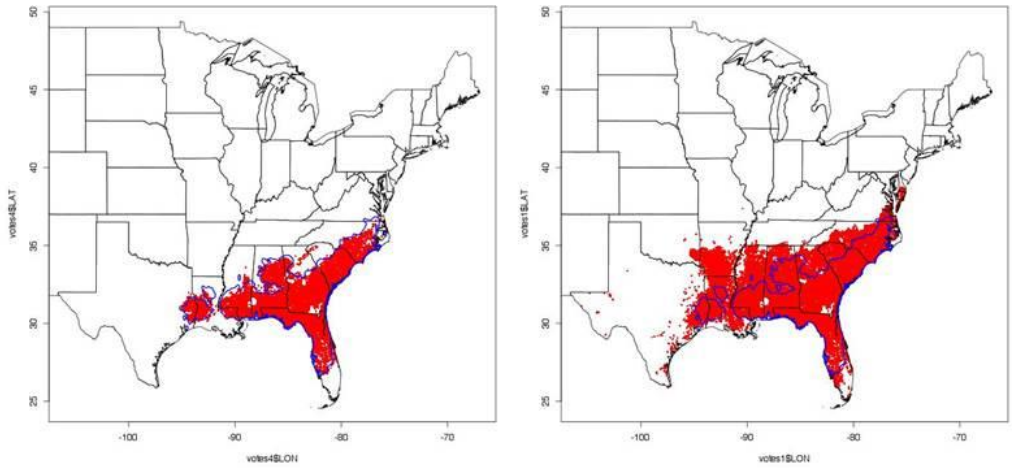


Figure 25. PA Longleaf pine CGCM3 A1B maps. (A) Longleaf pine OOB base model map, (B) year 2030 predictions for longleaf pine under CGCM3 A1B scenario, (C) year 2060 predictions for longleaf pine under CGCM3 A1B scenario, (D) year 2090 predictions for longleaf pine under CGCM3 A1B scenario.

(A) (B)



(C) (D)

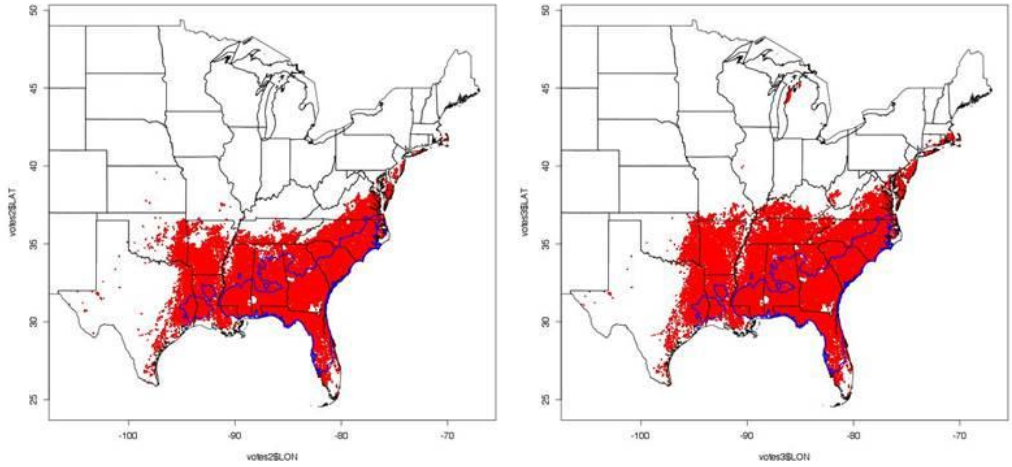


Figure 26. PA Longleaf pine CGCM3 A2 maps. (A) Longleaf pine OOB base model map, (B) year 2030 predictions for longleaf pine under CGCM3 A2 scenario, (C) year 2060 predictions for longleaf pine under CGCM3 A2 scenario, (D) year 2090 predictions for longleaf pine under CGCM3 A2 scenario.

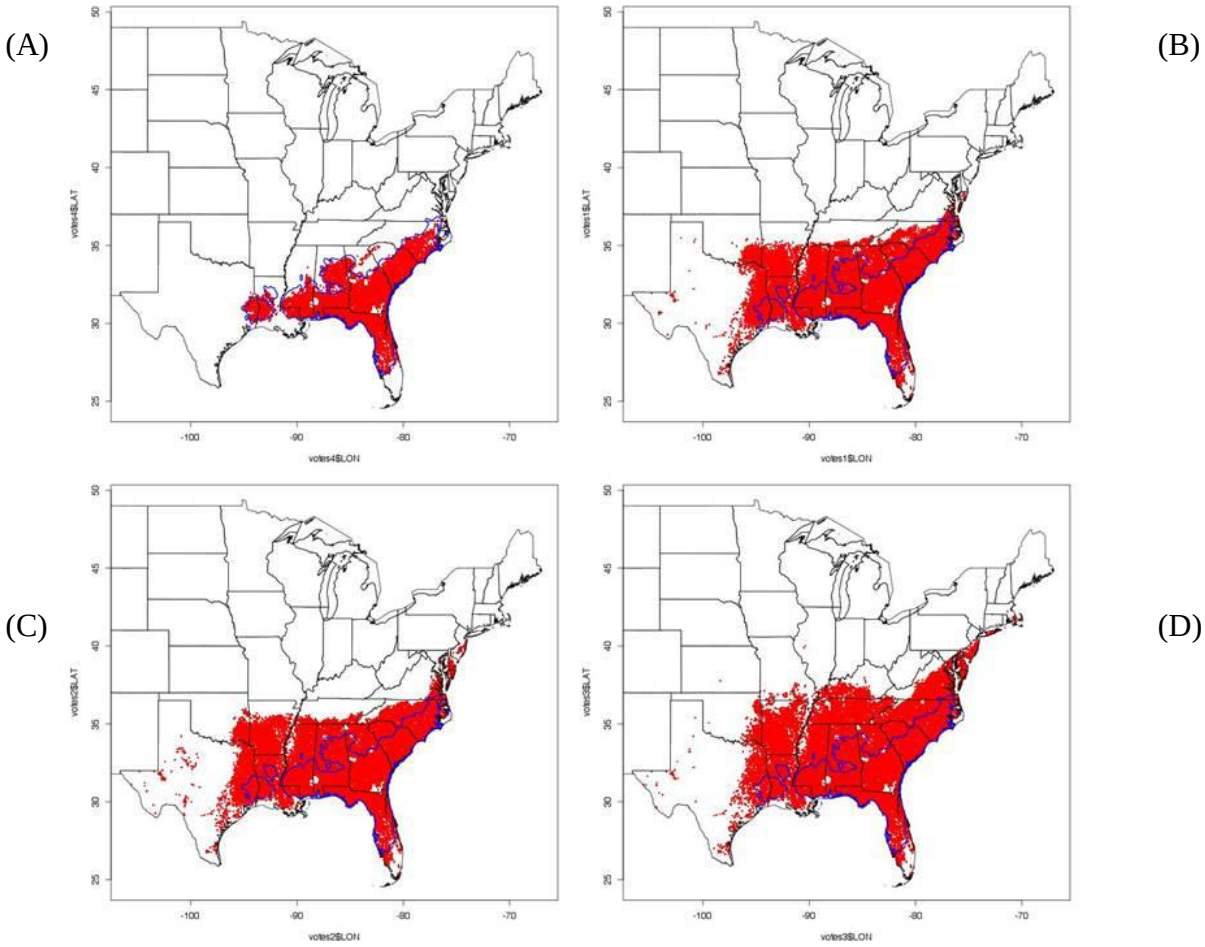
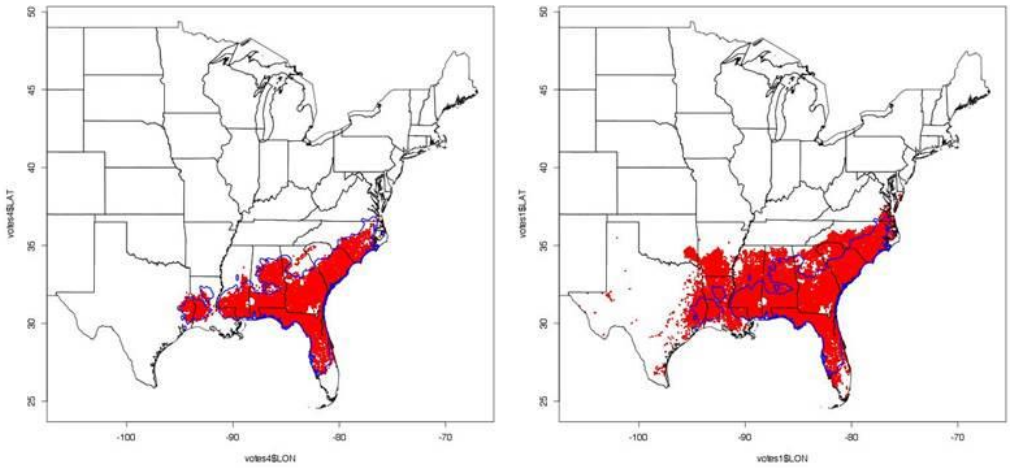


Figure 27. PA Longleaf pine HADCM3 A2 maps. (A) Longleaf pine OOB base model map, (B) year 2030 predictions for longleaf pine under HADCM3 A2 scenario, (C) year 2060 predictions for longleaf pine under HADCM3 A2 scenario, (D) year 2090 predictions for longleaf pine under HADCM3 A2 scenario.

(A) (B)



(C) (D)

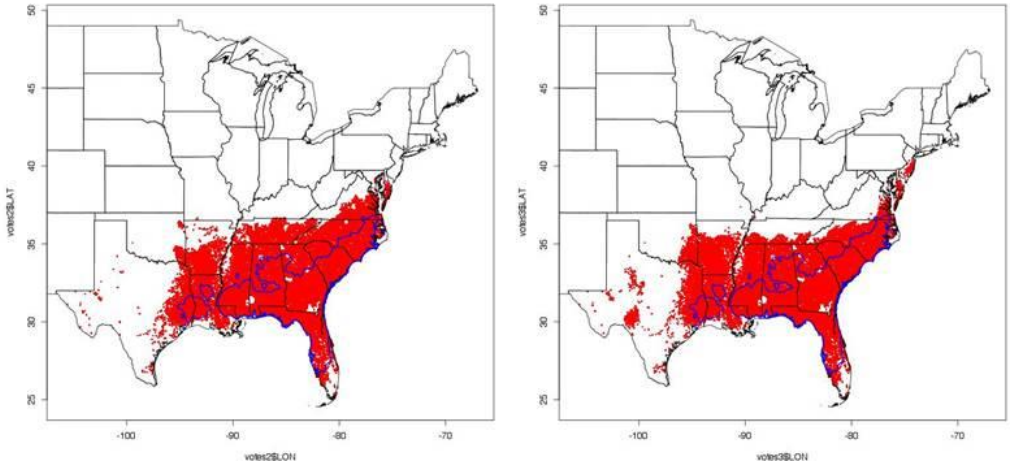
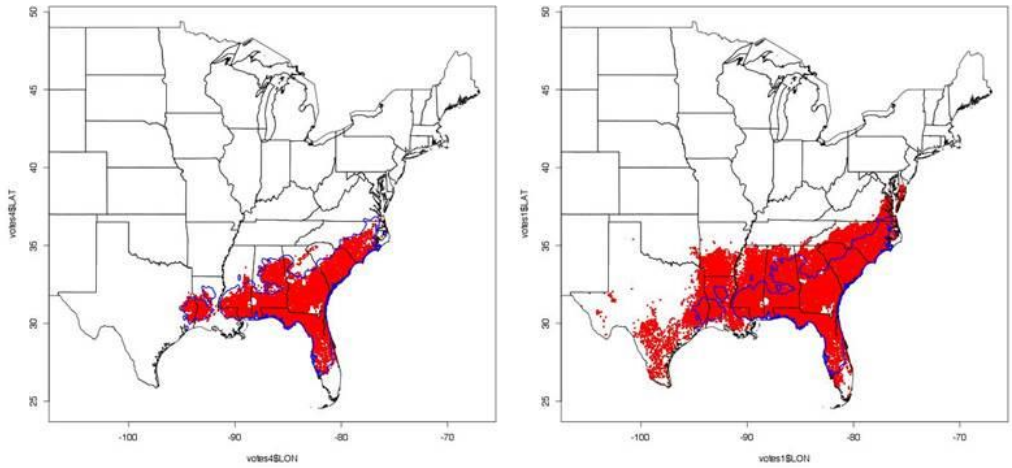


Figure 28. PA Longleaf pine HADCM3 B2 maps. (A) Longleaf pine OOB base model map, (B) year 2030 predictions for longleaf pine under HADCM3 B2 scenario, (C) year 2060 predictions for longleaf pine under HADCM3 B2 scenario, (D) year 2090 predictions for longleaf pine under HADCM3 B2 scenario.

(A) (B)



(C) (D)

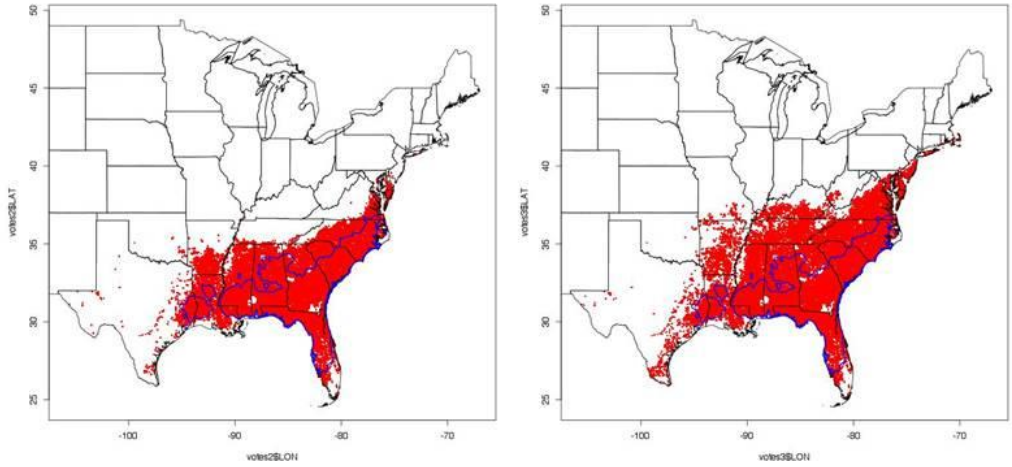
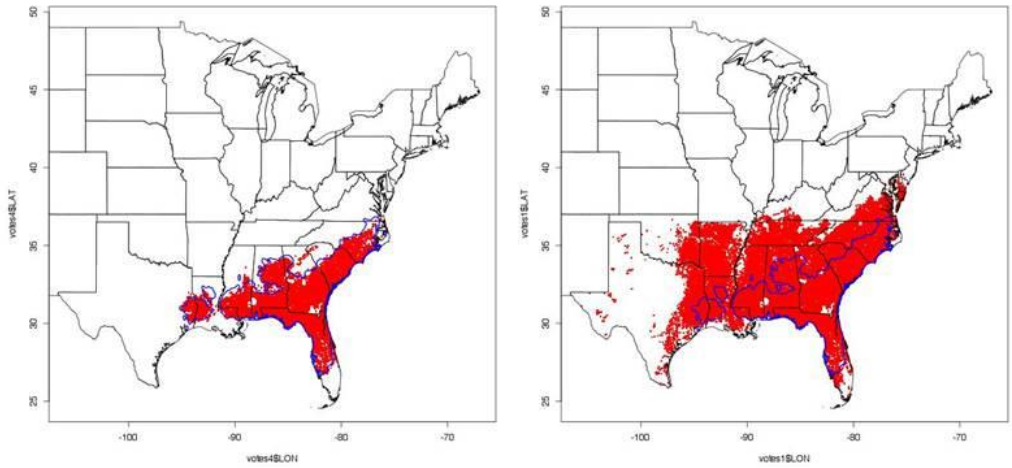


Figure 29. PA Longleaf pine GFDLCM21 A2 maps. (A) Longleaf pine OOB base model map, (B) year 2030 predictions for longleaf pine under GFDLCM21 A2scenario, (C) year 2060 predictions for longleaf pine under GFDLCM21 A2scenario, (D) year 2090 predictions for longleaf pine under GFDLCM21 A2scenario.

(A) (B)



(C) (D)

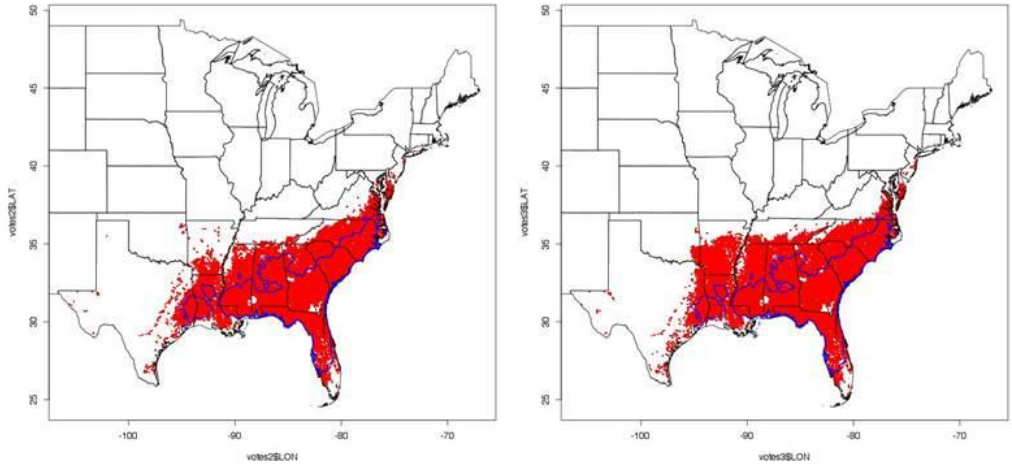


Figure 30. PA Longleaf pine GFDLCM21 B1 maps. (A) Longleaf pine OOB base model map, (B) year 2030 predictions for longleaf pine under GFDLCM21 B1scenario, (C) year 2060 predictions for longleaf pine under GFDLCM21 B1scenario, (D) year 2090 predictions for longleaf pine under GFDLCM21 B1scenario.

(A) (B)

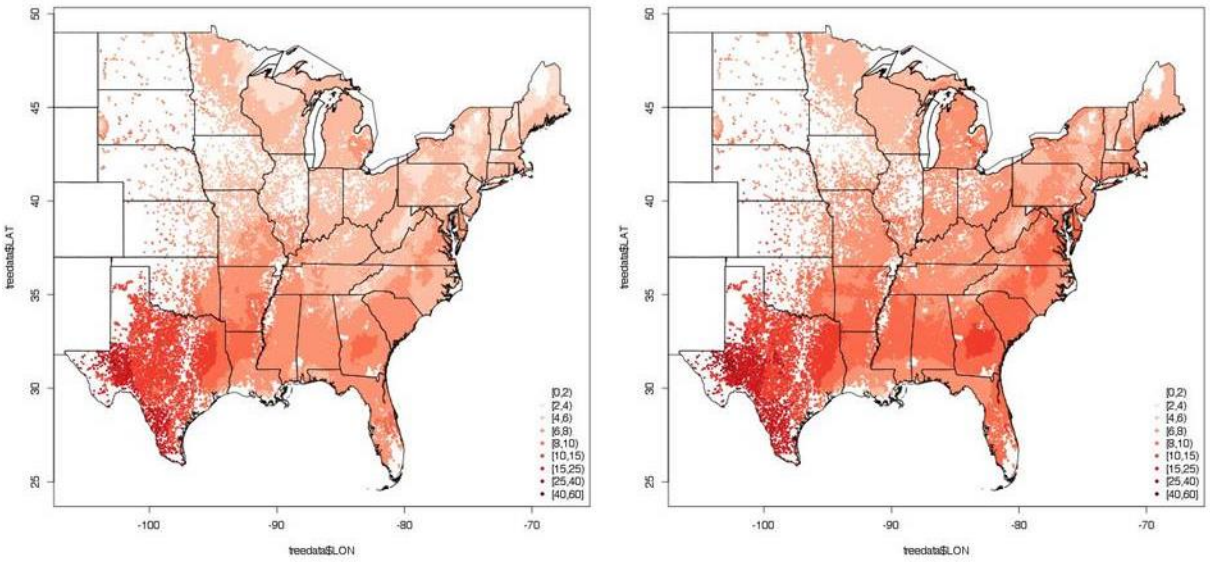


Figure 31. (A) CGCM3 B1 scenario map for SMI in the year 2090 and (B) CGCM3 A2 scenario map for SMI in the year 2090.

(A) (B)

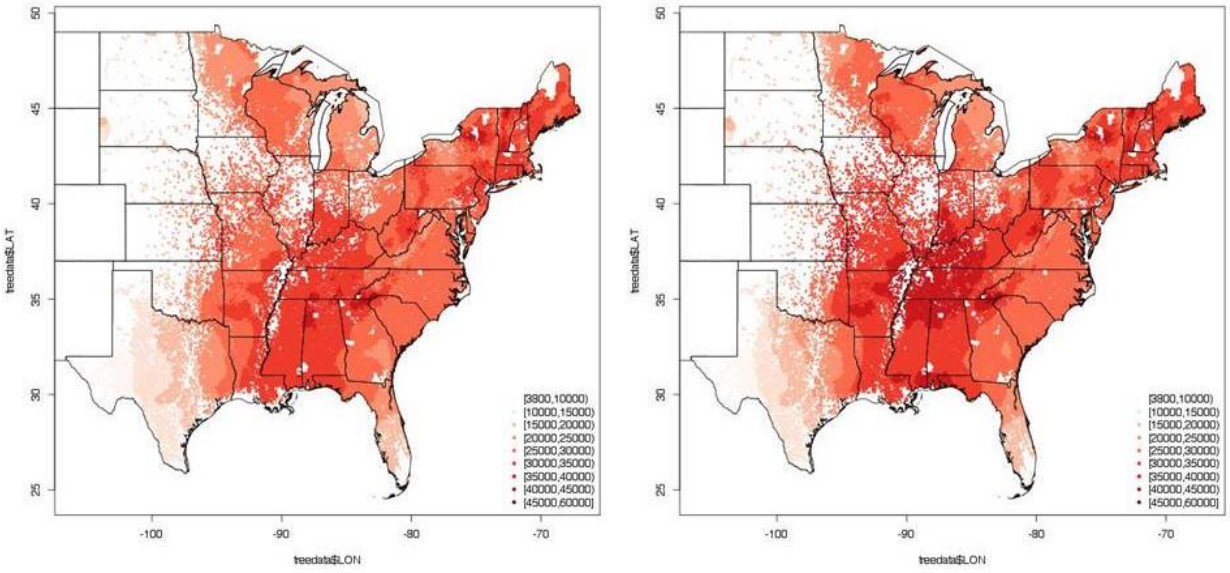


Figure 32. (A) CGCM3 B1 scenario map for MAPxTDIFF in the year 2090 and (B) CGCM3 A2 scenario map for MAPxTDIFF in the year 2090.

(A) (B)

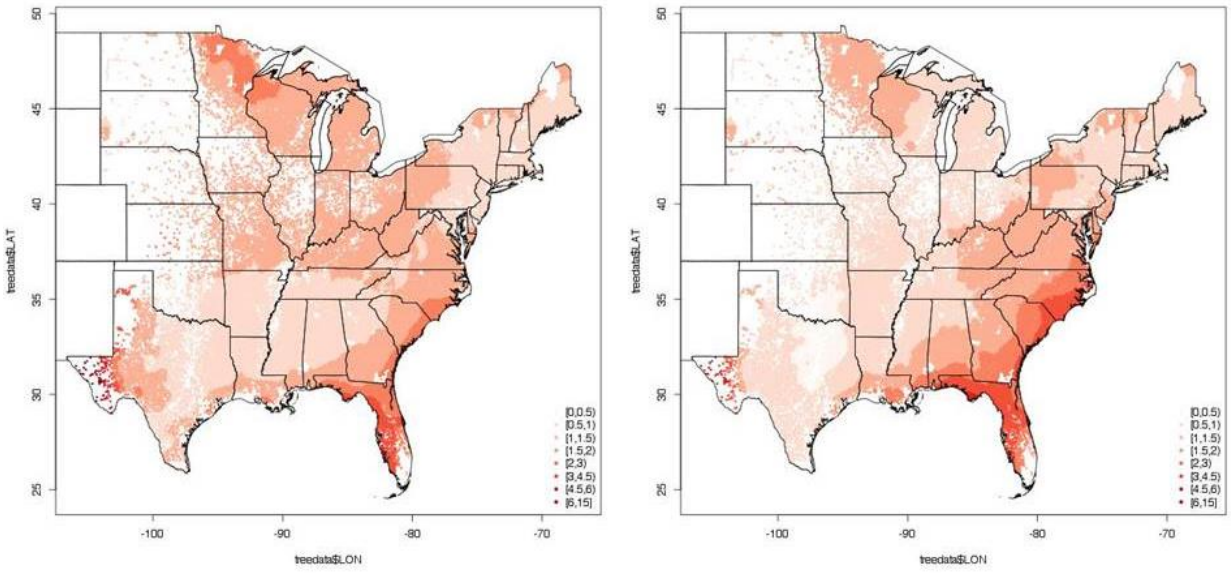


Figure 33. (A) CGCM3 B1 scenario map for smrsprpb in the year 2090 and (B) CGCM3 A2 scenario map for smrsprpb in the year 2090.

The climate maps for the scenarios CGCM3 B1 and CGCM3 A2 display how the different scenarios are predicting the values of each variable to be distributed in the eastern United States in 2090 (Figure 31 - Figure 33). To make inferences about how these climate maps are driving model predictions they must be compared to Figure 17 D and Figure 19 D. For further model investigation of how climate variables may be impacting IV, scatter plots were made (Figure 34 - Figure 36). Longleaf pine in the 2090 scenario for these two climate scenarios, CGCM3 B1 and CGCM3 A2, seem to predict the highest importance values at a median value of MAPxTDIFF, approximately within the range of 15,000-35,000 (Figure 34). While low values, less than three, of smrsprpb predict the highest IV(Figure 36). SMI seems to predict the highest IV for values near or slightly higher than five (Figure 35).

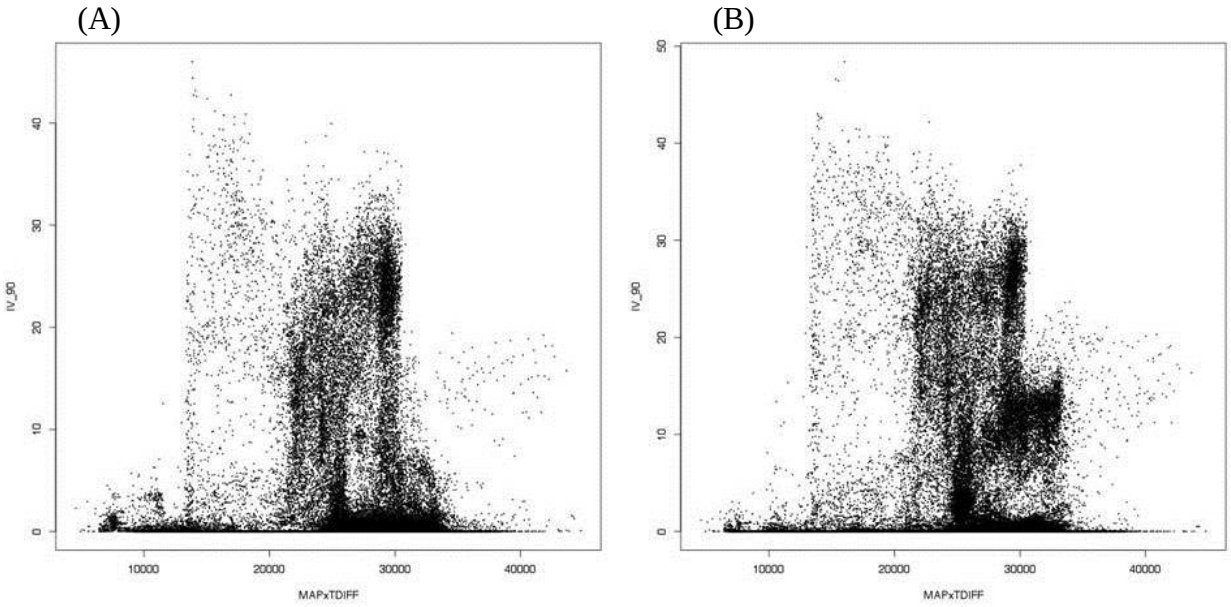


Figure 34. (A) Scatter plot of 2090 predicted longleaf pine IV versus 2090 climate variable MAPxTDIFF for CGCM3 B1 and (B) CGCM3 A2 scenario.

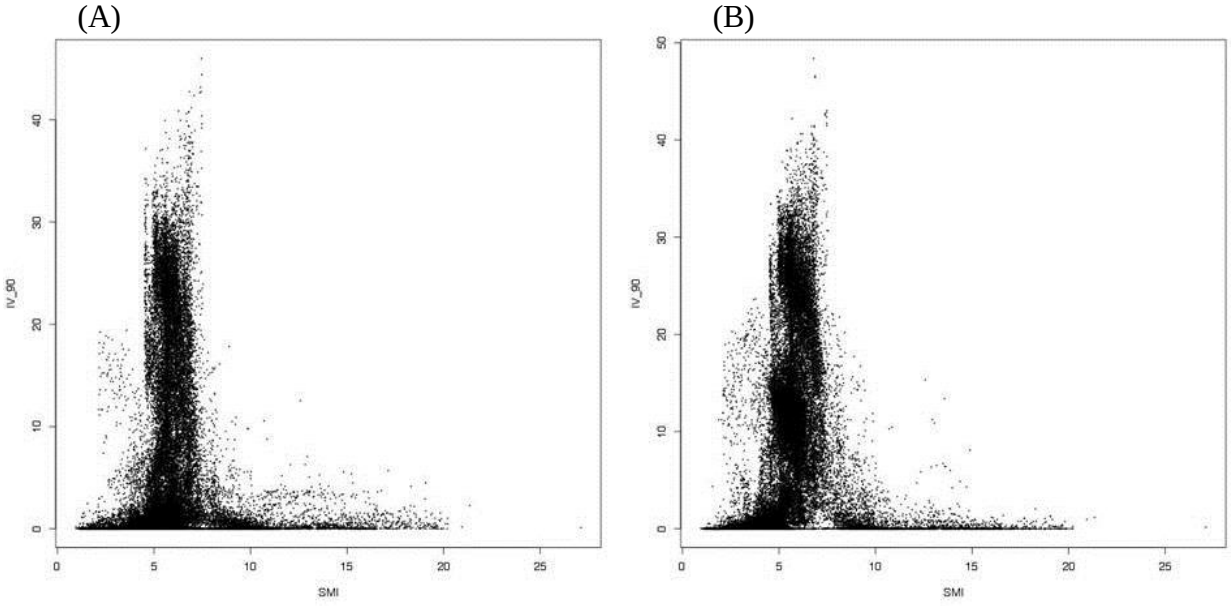


Figure 35. (A) Scatter plot of 2090 predicted longleaf IV versus 2090 climate variable SMI for CGCM3 B1 and (B) CGCM3 A2 scenario.

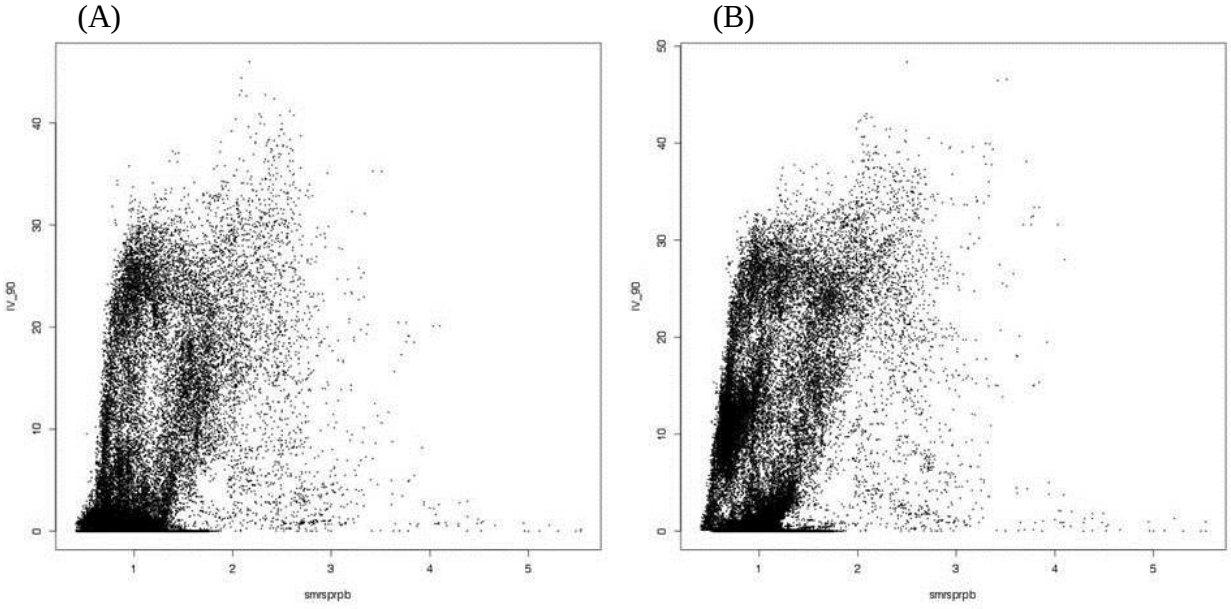


Figure 36. Scatter plot of 2090 predicted Longleaf IV versus 2090 climate variable smrsprpb for CGCM3 B1 (A) and CGCM3 A2 (B) scenario

While the pond pine PA models appeared to predict well using the BRF techniques in the model diagnostics phase Figure 37 appears to have pond pine growing in places that do not seem likely. Figure 10 shows that predominately climate variables are important for model prediction. Pond pine PA is an example of where rarity within the dataset could be causing our models to return false positives. Since the goal of our models was to eliminate false negatives, the model for pond pine is efficient however it might need a higher threshold than 0.05 to make more informative decisions. Animations were made of each species that combine one frame for every threshold value between 0 and 1. These movies allow the user to see the effect of T on the models. These animations can be found in the appendix (Figure A - 109 - Figure A - 113).

A combined modeling approach was tried to see how the two PA and IV models overlapped for pond pine. For this example both the IV and PA models were ran independently then the resulting datasets were merged together based on a plot identifier. When graphing this combined dataset the plots displayed meet both PA and IV criteria. This means that all plots have a PA T=0.05 and an IV > 1%. Figure 39 has a much more

restricted range than either the PA or IV maps individually (Figure 37and Figure 38). PA has the most noticeable changes occurring between the mapping methods but the IV is changing, there are fewer plots in the IV 1-3% class, however other higher classes are not affected. All other species and climate scenarios can be found in the appendix modeled in this same combination approach.

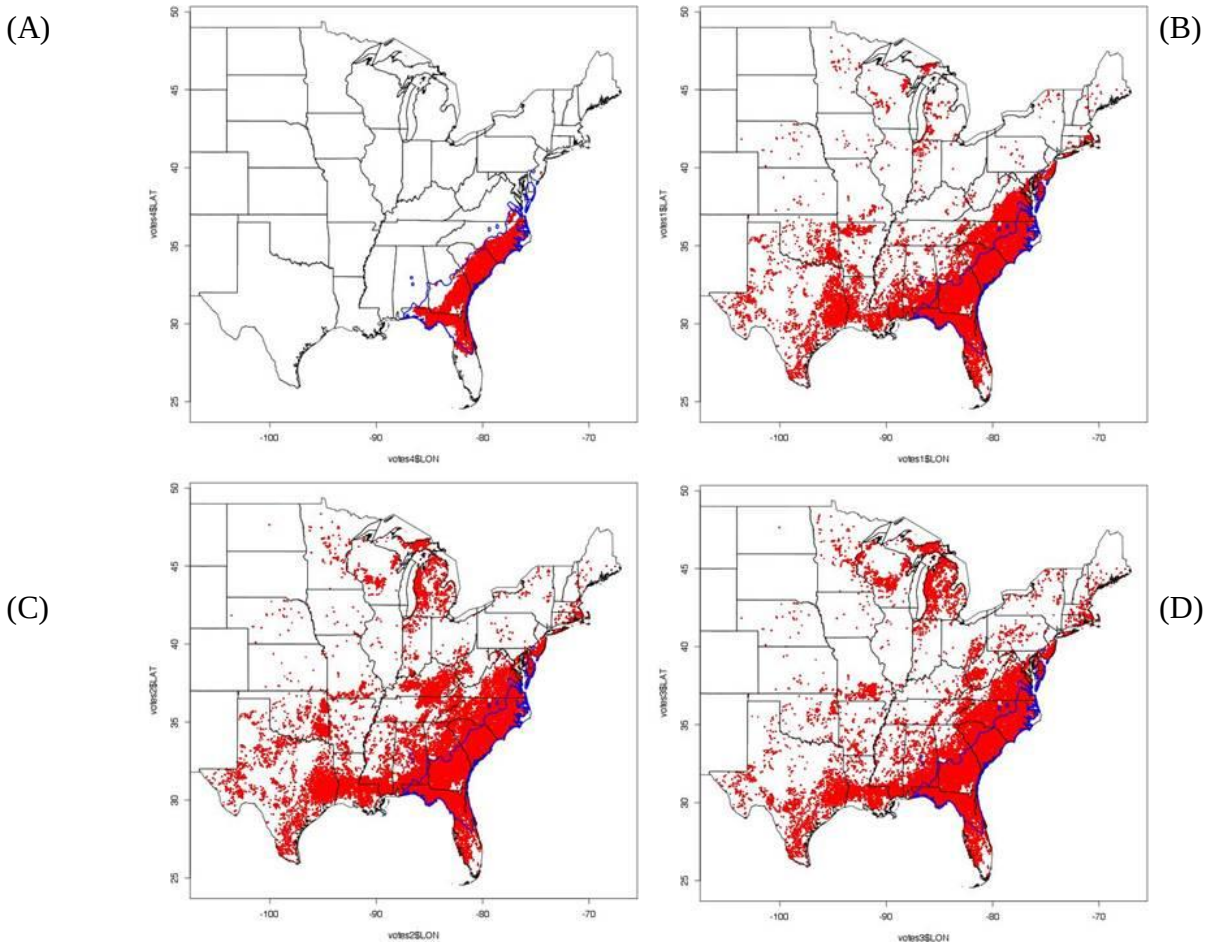
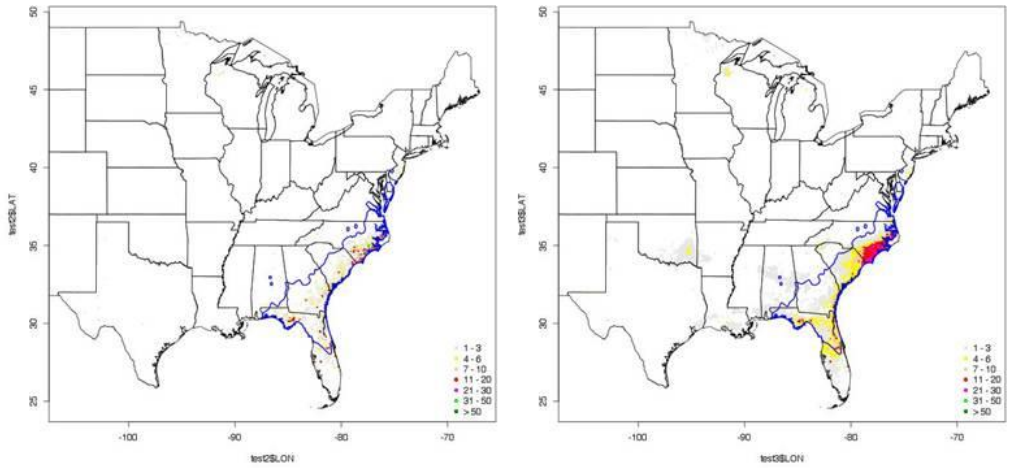


Figure 37. PA Pond pine CGCM3 A1B model maps. (A) Pond pine OOB base model map, (B) year 2030 predictions for pond pine under CGCM3 A1B scenario, (C) year 2060 predictions for pond pine under CGCM3 A1B scenario, (D) year 2090 predictions for pond pine under CGCM3 A1B scenario.

(A) (B)



(C) (D)

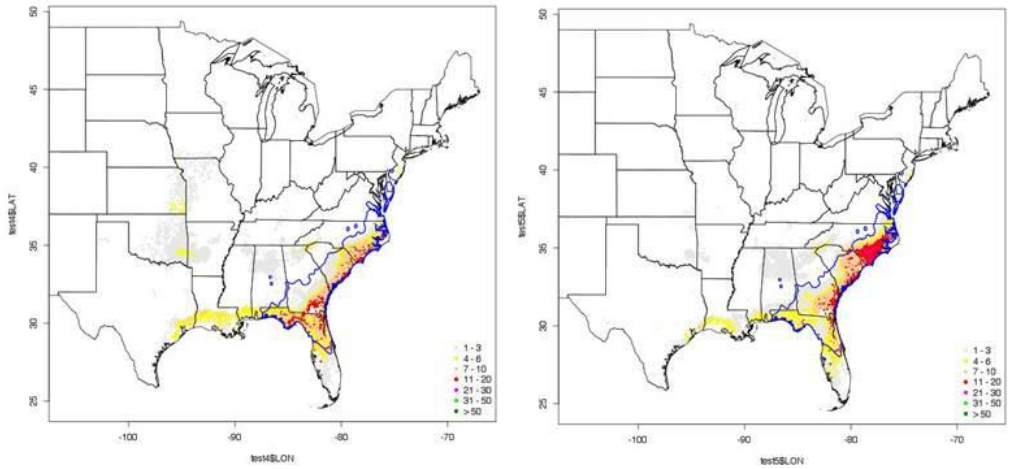


Figure 38. IV Pond pine CGCM3 A1B model maps. (A) Pond pine OOB base model map, (B) year 2030 predictions for pond pine under CGCM3 A1B scenario, (C) year 2060 predictions for pond pine under CGCM3 A1B scenario, (D) year 2090 predictions for pond pine under CGCM3 A1B scenario.

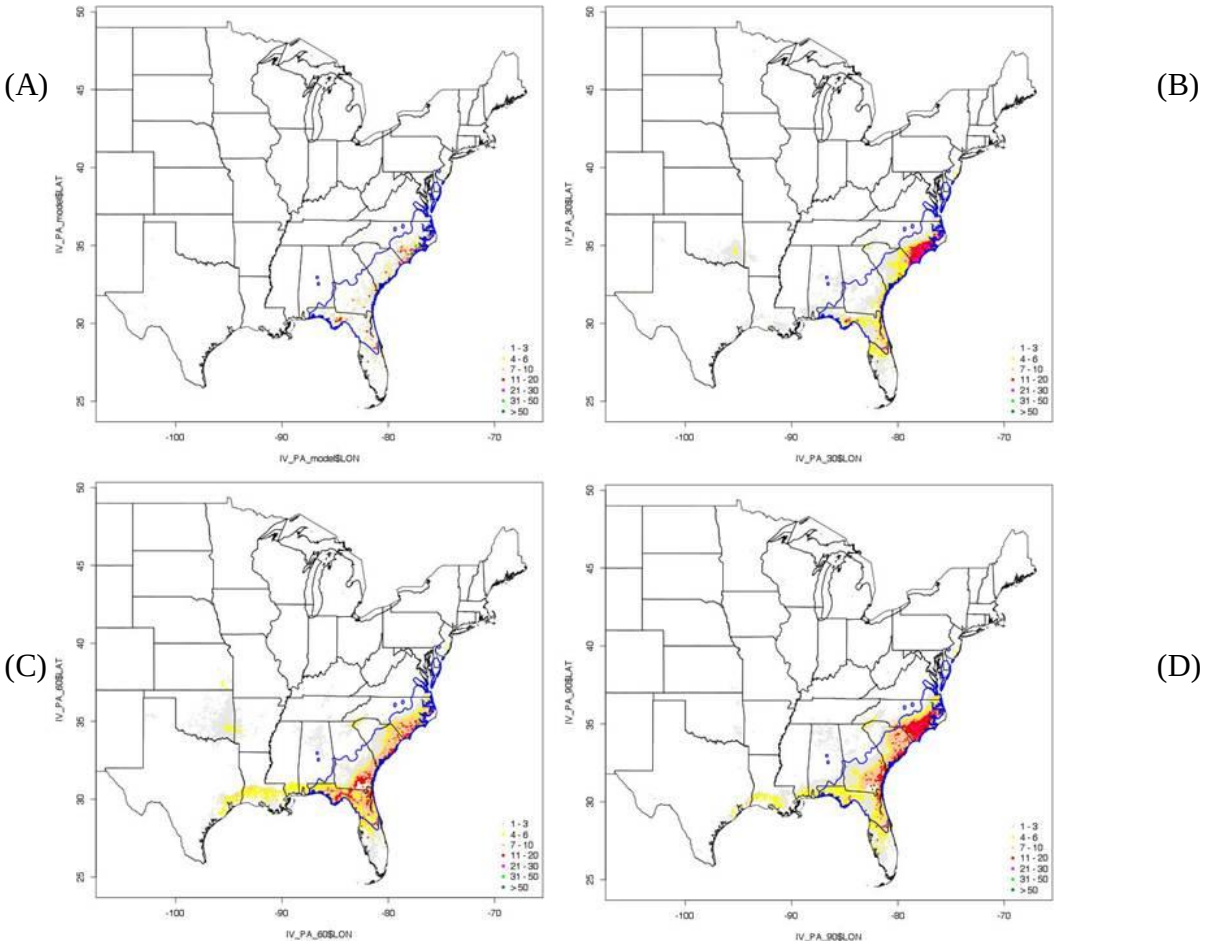


Figure 39. Combined IV and PA Pond pine CGCM3 A1B model maps. (A) Pond pine OOB base model map, (B) year 2030 predictions for pond pine under CGCM3 A1B scenario, (C) year 2060 predictions for pond pine under CGCM3 A1B scenario, (D) year 2090 predictions for pond pine under CGCM3 A1B scenario.

Discussion

In the current literature common data analysis tools among researchers are robust statistical-empirical modeling approach such as random forest (Crookston 2012b; Iverson et al. 2008; Iverson et al. 2004; Rehfeldt et al. 2009; Schultz 1999; Weiskittel et al. 2011b). Relationships between vegetation, edaphic, and climate variables are complex and nonlinear (Austin 2002). Traditional parametric statistical approaches poorly capture the patterns present in the data due to violations in statistical assumptions. Where the literature diverges is how to best manage the data and implement the modeling algorithms to get the most

accurate and meaningful results (Iverson et al. 2008; Iverson et al. 2011; Iverson et al. 2004; Rehfeldt et al. 2006; Rehfeldt et al. 2009).

Some researchers favor using abundance data versus presence-absence data (Iverson et al. 2008; Iverson et al. 2011; Iverson et al. 2004). The argument is that a regression-based approach for abundance (IV) is more powerful than a classification-based approach for presence-absence (PA) when trying to predict future potential habitat ranges. Abundance data interpretations based on the core of a species range can be made rather than a more uncertain range where all boundaries are equally weighted. Iverson et al. (2011) states that the main advantage of using abundance data is that analyses and interpretations can be based on the core of the specie's ranges, rather than the more uncertain range boundaries that are equally weighted in presence-absence data. Iverson et al. (2011) defined the core of a species range as an area of the species distribution where there is a greater certainty of species occurrence. In the specific example cited within their research they found a PA model for a species to predict a 90% loss in the extent of the species habitat but only a 36% loss from its current range when abundance data was used. They attributed the large difference to higher sensitivity of the model being able to distinguish core areas from edge areas (Austin 2002; Iverson et al. 2011).

Other authors have relied more on presence-absence data and bioclimatic models of realized niches for contemporary species distributions (Joyce and Rehfeldt 2012; Rehfeldt et al. 2006; Rehfeldt et al. 2009). These authors use a technique similar to a WRF. They set up their models by developing a sampling protocol which was used to split the vegetation data into 40% presence observations and 60% absence observations with presence observations being weighted by a factor of 2. The climate data was used to create a hypervolume in which each dimension consisted of an estimate of the species climatic limits, mean of all observations of present vegetation data, expanded by ± 0.1 standard deviation. Of the 60% absence data used within the model 40% of the absence observation came from within the same climatic hypervolume, or climate envelope. While the remaining 20% of the absence observations came from outside of this climate envelope (Rehfeldt et al. 2006).

Random Forest was the primary analysis tool for both the IV and PA models previously mentioned. However the models created in this project are slightly different. This project

used the regression based random forest methods with IV abundance data but with a few modifications on data preparation. (Iverson et al. 2008) averaged FIA plots over a gridded area, this project used point data from individual FIA plots. Rehfeldt et al. (2006), Rehfeldt et al. (2009), Joyce and Rehfeldt (2012) developed climate envelopes then using a WRF analyzed their data. A climate envelope was not developed for this projects models instead all plots were used in each models training set and a BRF was used to randomly create a dataset of equal proportions between present and absent observations during the bootstrapping and bagging operations. Another difference in the models presented in this project from the literature is that SSURGO soils data was used. The use of point data instead of gridded data allowed the use of detailed soils data where previously the coarser STATSGO soil data was sufficient for the gridded analysis.

When using classification trees to make predictions, the relative costs of false negative and false positive predictions must be considered. A false positive for these models, for example, would be if a plot was classified as absent when the observed value was present. A false positive, for example, is if the plot was classified as present when the actual observed value was absent. The cost of a false negative in this case could be greater than the cost of a false positive because of the effort, time, and money that might be spent on management plans and conservation efforts to help a species survive. The function that has been used to describe the relationship between a false negative and a false positive is the slope of an iso-performance line in receiver operating characteristics (ROC) space (Fawcett 2006). A ROC graph is a technique that is used for visualizing, organizing, and selecting classifiers based on their performance, and all points that correspond to the slope of an iso-performance line have an equal classification cost.

$$m = \left(\frac{Cost\ of\ a\ false\ negative}{Cost\ of\ false\ positive}\right) \qquad (4)$$

The formula given for the iso-performance line (4) ignores the proportions of presence and absence within the training data because of the use of a BRF. The proportion in this case is equal to one and so it is not necessary to include in the formula.

Work from (Iverson et al. 2011; Rehfeldt et al. 2006; Schwartz et al. 2006) have found any species without a sufficient amount of data are difficult to develop models for while trying to preserve a high confidence in the outputs. Rarity within the dataset could explain poor model reliability for longleaf pine and pond pine (Table 8 and Table 9). The most frequently occurring species in the dataset is loblolly pine, which is frequently found growing in monocultures. This is why loblolly pine experiences a larger average IV than most other species which are more commonly found growing in mixed stands (Table 8). Slash pine is another example of a species that is frequently found in monocultures and has a high average IV.

Some unanswered questions stemming from this research is how the models or data could have been further manipulated to improve results. For example the IV regression models were produced using the full dataset and not by using a balanced design like the PA classification models. Would have developing a random sampling regime to down-sample the 0% IV plots given the model more power to predict for the current range of each species, thus giving a strong predictive power in the future models? A randomized down-sampling technique could be especially helpful for a rare species such as pond pine where out of more than 70,000 observations only approximately 300 had an IV greater than 0% (Table 6).

Another question of interest is how would model development be affected if data from outside of the United States was used? It would not be expected to find any of the five southern pines analyzed here in Canada, however absent / 0% IV is still a data value with associated climatic variables that could be used in model development. How would adding these data points affect the models? Alternatively how would additional presence / IV > 0% data points affect the models? Pine plantations in South America include species of interest to this project such as loblolly pine. How would including additional present observations from warmer moister environments affect the model and would the effects of these observations differ from the effects of the previously mentioned absent observations?

Genetically speaking observations from pine plantations in South America would be different than the typical natural forest FIA plot. Some researchers have analyzed and noted different climatic responses within the same species due to genetic variation and suggest models need reflect these differences (Crookston et al. 2010; Rehfeldt et al. 1999). For the

models presented in this work plantation data and natural forest data from FIA plots were analyzed together and not separately. Future work as an extension of this project would be to make the distinction between natural forest and plantation plots within the FIA data and to refine the models based on specific genetic response results.

References

Austin MP. 2002. Spatial prediction of species distribution: an interface between ecological theory and statistical modelling. *Ecol. Model.* 157:101-118.

Bolin JF. 2007. Seed bank response to wet heat and the vegetation structure of a Virginia pocosin. *J. Torrey Bot. Soc.* 134:80-88.

Bose I. 2012. Preprocessing unbalanced data using support vector machine. *Decision support systems* 53:226-233.

Breiman L. 2001. Random forests. *Machine Learning* 45:5-32.

Burns RM, Honkala BH. 1990. Silvics of North America volume 1: Conifers. In: Agriculture Handbook 654. Washington, DC. USDA Forest Service. 1383 p.

Chen C, Liaw A, Breiman L. 2004. Using Random Forest to Learn Imbalanced Data. UC Berkeley.Technical Report 666. 12 p.

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

Coops NC, Waring RH. 2011. A process-based approach to estimate lodgepole pine (*Pinus contorta* Dougl.) distribution in the Pacific Northwest under climate change. *Climatic Change* 105:313-328.

Crookston NL. 2010. Climate-FVS version 1: Notes for users. Fort Collins, CO. RMRS-Moscow. 10 p.

Crookston NL. 2012a Details on spatial extents, temporal information and data elements http://forest.moscowfsl.wsu.edu/climate/details.php accessed May 30, 2012.

Crookston NL. 2012b Research on forest climate change: potential effects of global warming on forests and plant climate relationships in western North America and Mexico http://forest.moscowfsl.wsu.edu/climate/ accessed May 30, 2012.

Crookston NL, Dixon GE. 2005. The forest vegetation simulator: A review of its structure, content, and applications. *Computers and Electronics in Agriculture* 49:60-80.

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* 260:1198-1211.

Dixon GE. 2002. Essential FVS: A user's guide to the Forest Vegetation Simulator. Fort Collins, CO. USDA Forest Service. 248 p.

ESRI. 2012 ArcGIS Features http://www.esri.com/software/arcgis/features.html accessed May 30, 2012.

Fawcett T. 2006. An introduction to ROC analysis. *Pattern Recognition Letters* 27:861-874.

Forest Inventory and Analysis. 2011. The Forest Inventory and Anaylsis Database: Database description and users manual version 5.1 for phase 2. Arlington, Va. USDA Forest Service. 409 p.

Forest Inventory and Analysis. 2012 FIA DataMart FIADB version 5.1 http://apps.fs.fed.us/fiadb-downloads/datamart.html accessed May 30, 2012.

Hansen AJ, Neilson RR, Dale VH, Flather CH, Iverson LR, Currie DJ, Shafer S, Cook R, Bartlein PJ. 2001. Global change in forests: Responses of species, communities, and biomes. *Bioscience* 51:765-779.

Harding SR, Walters JR. 2002. Processes regulating the population dynamics of red-cockaded woodpecker cavities. *Journal of Wildlife Management* 66:1083-1095.

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

Hutchinson M. 1998a. Interpolation of rainfall data with thin plate smoothing splines - Part I: Two dimensional smoothing of data with short range correlation. *Journal of Geographic Information and Decesion Analysis* 2:139-151.

Hutchinson M. 1998b. Interpolation of rainfall data with thin plate smoothing splines - Part II: Analysis of topographic dependence. *Journal of Geographic Information and Decesion Analysis* 2:152-167.

IPCC. 2000. IPCC special report emissions scenarios: Summary for policymakers. *Special report of IPCC working group III* 21.

IPCC. 2007. Climate change 2007: synthesis report--Pachauri RK RA, eds, ed. Geneva: Intergovernmental Panel on Climate Change. 52 p.

Iverson LR, Prasad AM. 1998. Predicting abundance of 80 tree species following climate change in the eastern United States. *Ecological Monographs* 68:465-485.

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* 254:390-406.

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* 14:1005-1020.

Iverson LR, Schwartz MW, Prasad AM. 2004. How fast and far might tree species migrate in the eastern United States due to climate change? *Glob. Ecol. Biogeogr.* 13:209-219.

Joyce DG, Rehfeldt GE. 2012. Climatic niche, ecological genetics, and impact of climate change on eastern white pine (Pinus strobus L): Guidelines for land managers. *Forest ecology and management*:173-192.

Latta G, Temesgen H, Adams D, Barrett T. 2010. Analysis of potential impacts of climate change on forests of the United States Pacific Northwest. *Forest Ecology and Management* 259:720-729.

Liaw A, Wiener M. 2002. Classification and Regression by randomForest. R News

Liaw A, Wiener M, Breiman L, Cutler A. 2012. Package 'randomForest'. CRAN. p. 29

McNulty S, Sun G, Mohan J, Caldwell P, Prestemon J, Doyle TW, Johnsen K, Liu Y. 2012. Chapter 8 Forests and Climate Change in the Southeast. In: Southeastern Regional Technical Report to the National Climate Assessment--Ingram KT, Dow K, Carter L, eds. 23 p.

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* 38:343-352.

Powell GL, White TL. 1994. Cone and seed yields from slash pine seed orchards. *Southern journal of applied forestry* 18:122-127.

Prasad AM, Iverson LR, Matthews SN, Peters MP. 2007-ongoing. A climate change atlas for 124 forest tree species of the eastern united states. In: Northern Research Station. Delaware, Ohio. USDA Forest Service. http://www.nrs.fs.fed.us/atlas/tree

R Development Core Team. 2011. R: A language and environment for statistical computing. -- Computing RFfS, ed. Vienna, Austria

Reams GA, Roesch FA, Cost ND. 1999. Annual Forest Inventory: Cornerstone of Sustainability in the South. *Journal of Forestry* 97:21-26.

Rehfeldt GE. 2006. A spline model of climate for the Western United States. Fort Collins, CO. U.S. Dept. of Agriculture, Forest Service, Rocky Mountain Research Station. 28 p.

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* 167:1123-1150.

Rehfeldt GE, Ferguson DE, Crookston NL. 2009. Aspen, climate, and sudden decline in western USA. *Forest ecology and management* 258:2353-2364.

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* 69:375-407.

Rocky Mountain Research Station. 2012 Custom climate data request http://forest.moscowfsl.wsu.edu/climate/customData/ accessed May 30, 2012.

Sabatia CO, Lynch TB, Will RE. 2008. Tree biomass equations for naturally regenerated shortleaf pine. *Southern journal of applied forestry* 32:163-167.

Santos MJ, Greenberg JA, Ustin SL. 2010. Using hyperspectral remote sensing to detect and quantify southeastern pine senescence effects in red-cockaded woodpecker (Picoides borealis) habitat. *Remote Sens. Environ.* 114:1242-1250.

Schultz RP. 1999. Loblolly - the pine for the twenty-first century. *New forests* 17:71-88.

Schwartz MW, Iverson LR, Prasad AM, Matthews SN, O'Connor RJ. 2006. Predicting extinctions as a result of climate change. *Ecology* 87:1611-1615.

Sharma S, Osei-Bryson KM, Kasper GM. 2012. Evaluation of an integrated Knowledge Discovery and Data Mining process model. *Expert Syst. Appl.* 39:11335-11348.

Smith WB, Miles PD, Perry CH, Pugh SA. 2010. Forest Resources of the United States, 2007. 349 p.

Stage AR. 1973. Prognosis model for stand development. Ogden, Utah. USDA Forest Service. 32 p

The R Project. 2011 What is R? http://www.r-project.org/ accessed May 30, 2012.

USDA. 1994. State Soil Geographic (STATSGO) Database: Data Use and Information. Micellaneous Publication Number 1492. Washington, DC. Natural Resources Conservation Service. 113 p.

USDA. 1995. Soil Survey Geographic (SSURGO) Data Base: Data Use and Information. Miscellaneous Publication Number 1527. Washington, DC. Natural Resources Conservation Service. 110 p.

USDA Forest Service. 2005. Forest Inventory and Analysis: Sampling and Plot Design. 2 p.

USDA NRCS. 2011 GeoSpatialData http://datagateway.nrcs.usda.gov/GDGOrder.aspx?order=QuickState accessed May 30, 2012.

Vincenzi S, Zucchetta M, Franzoi P, Pellizzato M, Pranovi F, De Leo GA, Torricelli P. 2011. Application of a Random Forest algorithm to predict spatial distribution of the potential yield of Ruditapes philippinarum in the Venice lagoon, Italy. *Ecol. Model.* 222:1471-1478.

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* 3:114-123.

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

Weiskittel AR, Crookston NL, Radtke PJ. 2011b. Linking climate, gross primary productivity, and site index across forests of the western United States. *Canadian Journal of Forest Research-Revue Canadienne De Recherche Forestiere* 41:1710-1721.

Yin YF, Gong GH, Han L. 2011. Theory and techniques of data mining in CGF behavior modeling. *Sci. China-Inf. Sci.* 54:717-731.

Appendix

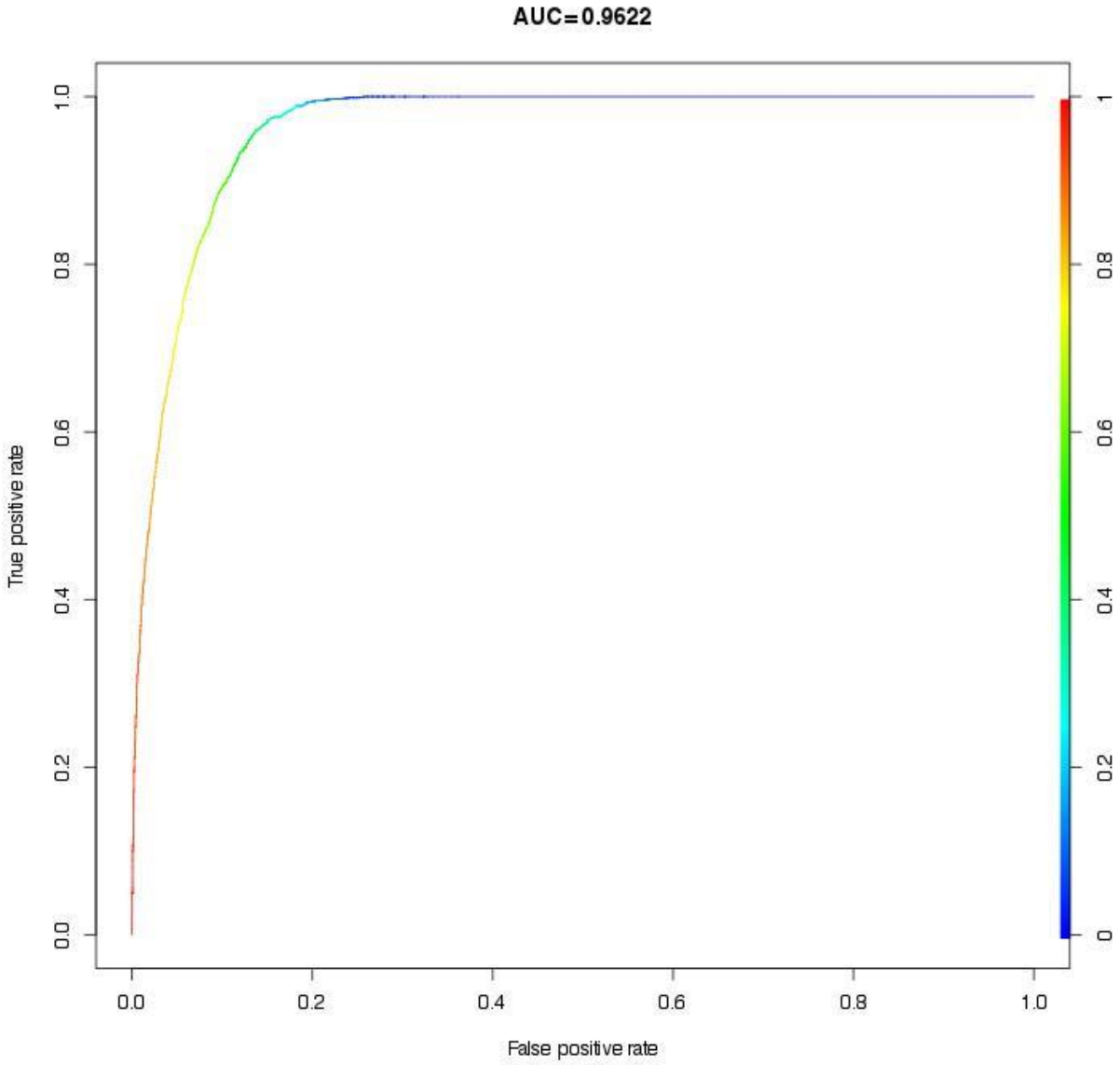


Figure A - 1. ROC graph for longleaf pine PA model showing AUC and where different vote threshold values lie along the curve.

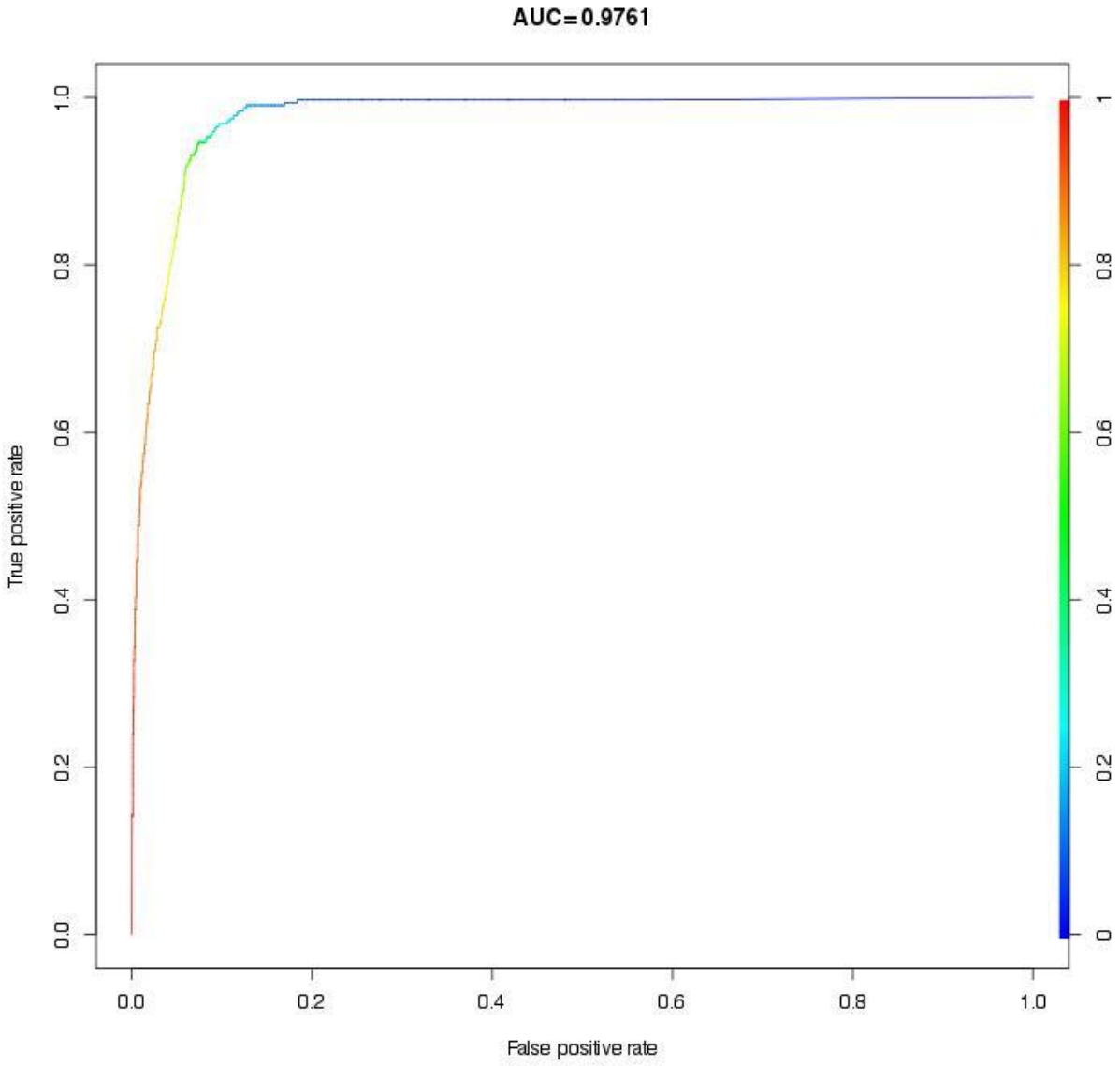


Figure A - 2. ROC graph for pond pine PA model showing AUC and where different vote threshold values lie along the curve.

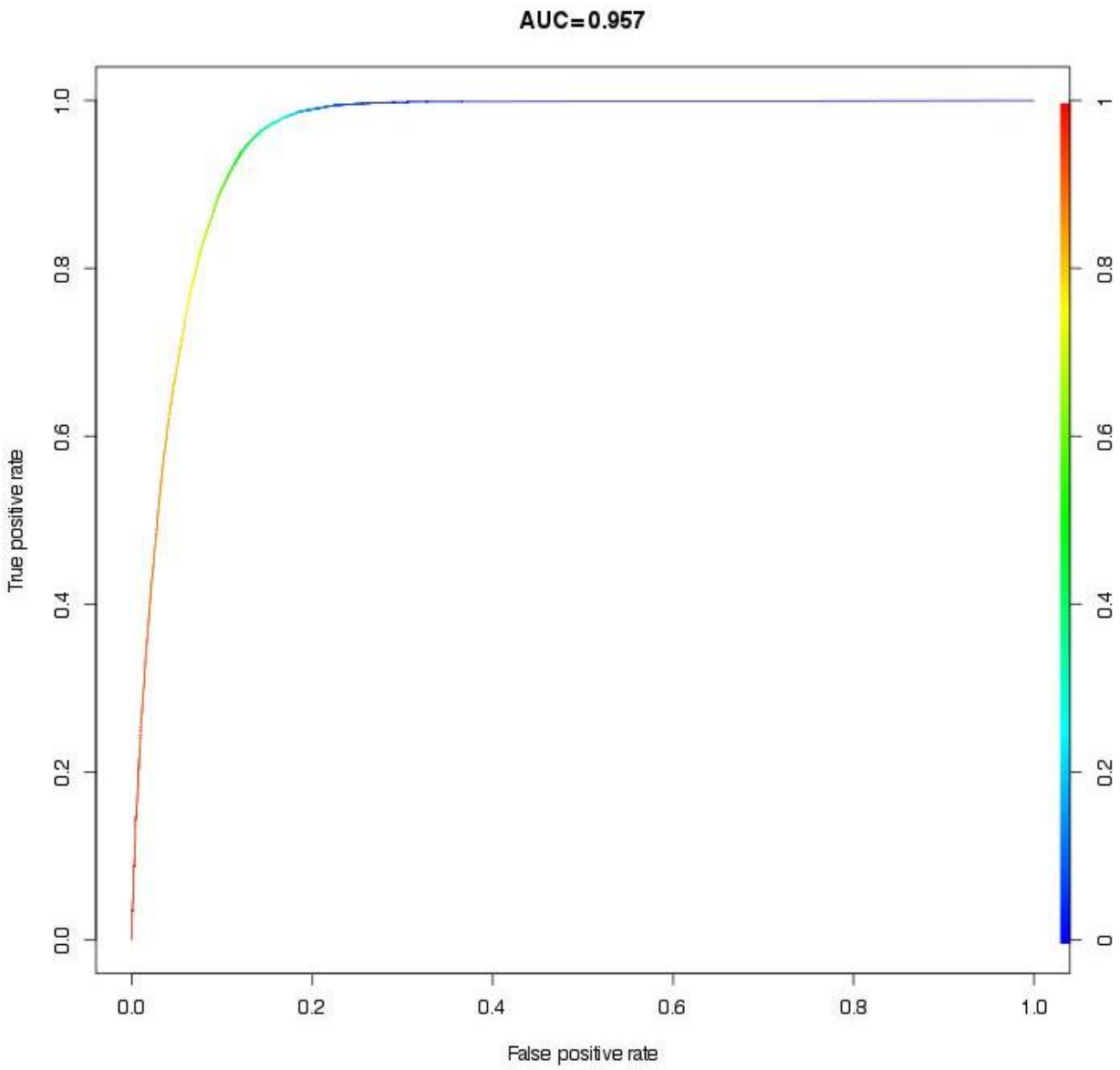


Figure A - 3. ROC graph for loblolly pine PA model showing AUC and where different vote threshold values lie along the curve.

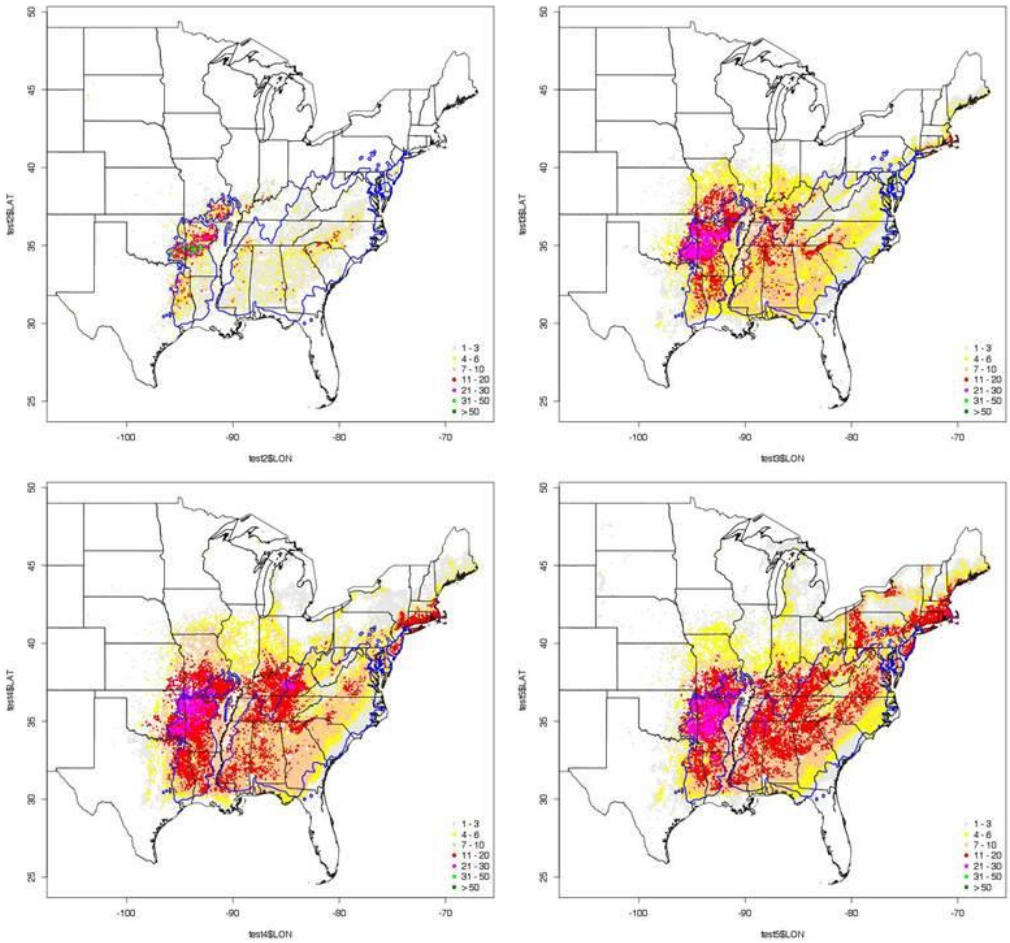


Figure A - 4. IV Shortleaf pine CGCM3 A1B models (top-left), 2030 (top-right), 2060 (bottom-left), 2090 (bottom-right)

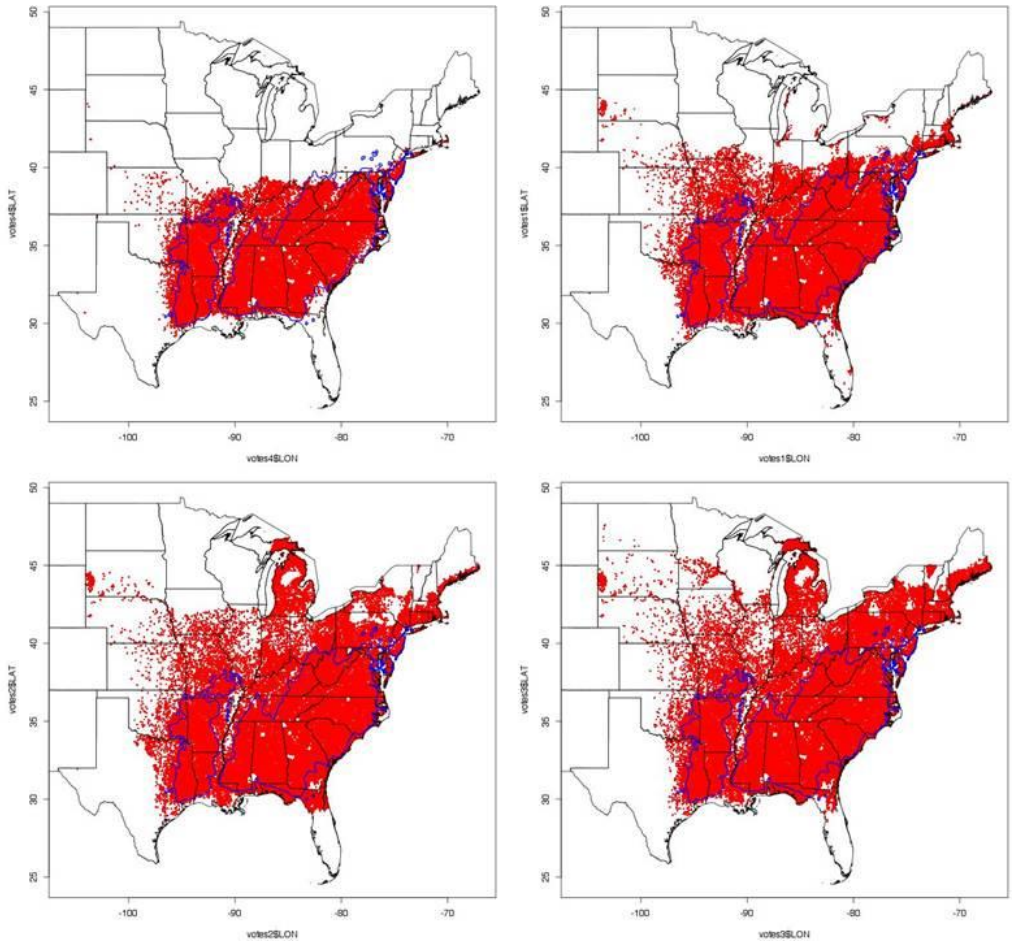


Figure A - 5. PA Shortleaf pine CGCM3 A1B models (top-left), 2030 (top-right), 2060 (bottom-left), 2090 (bottom-right)

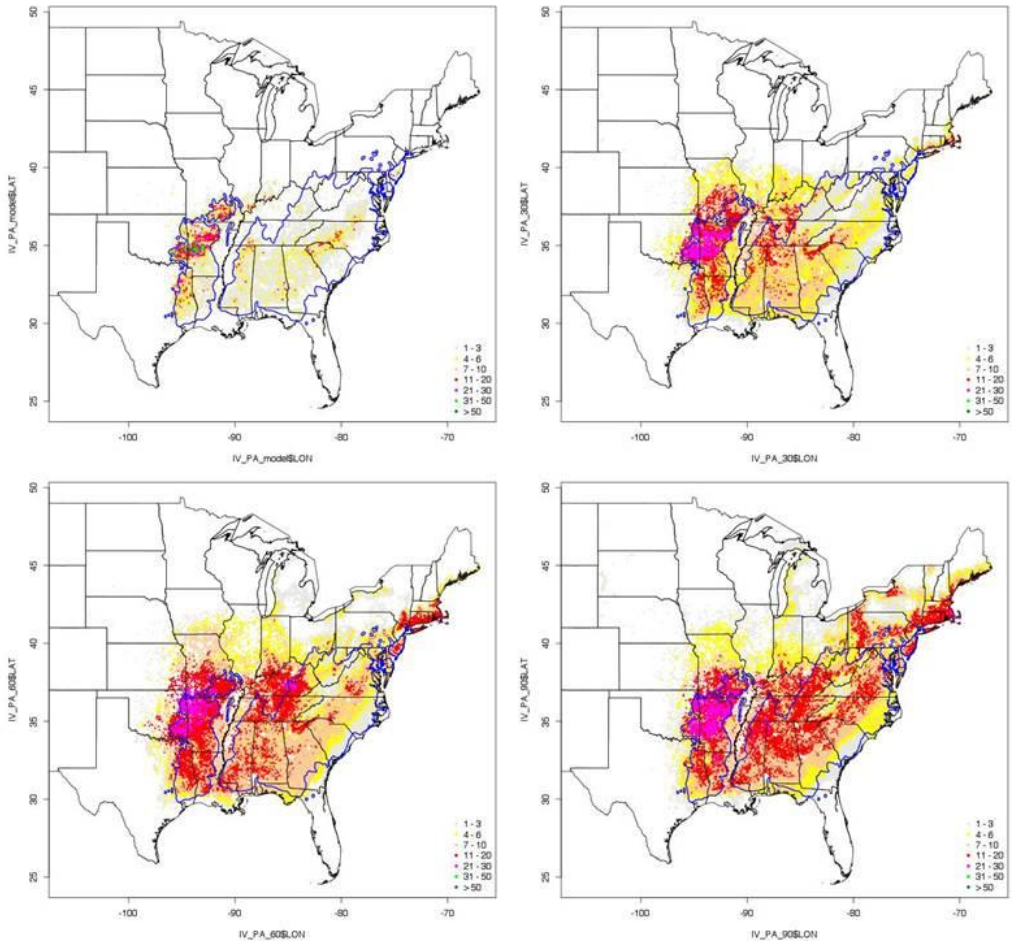


Figure A - 6. Combined IV and PA Shortleaf pine CGCM3 A1B models (top-left), 2030 (top-right), 2060 (bottom-left), 2090 (bottom-right)

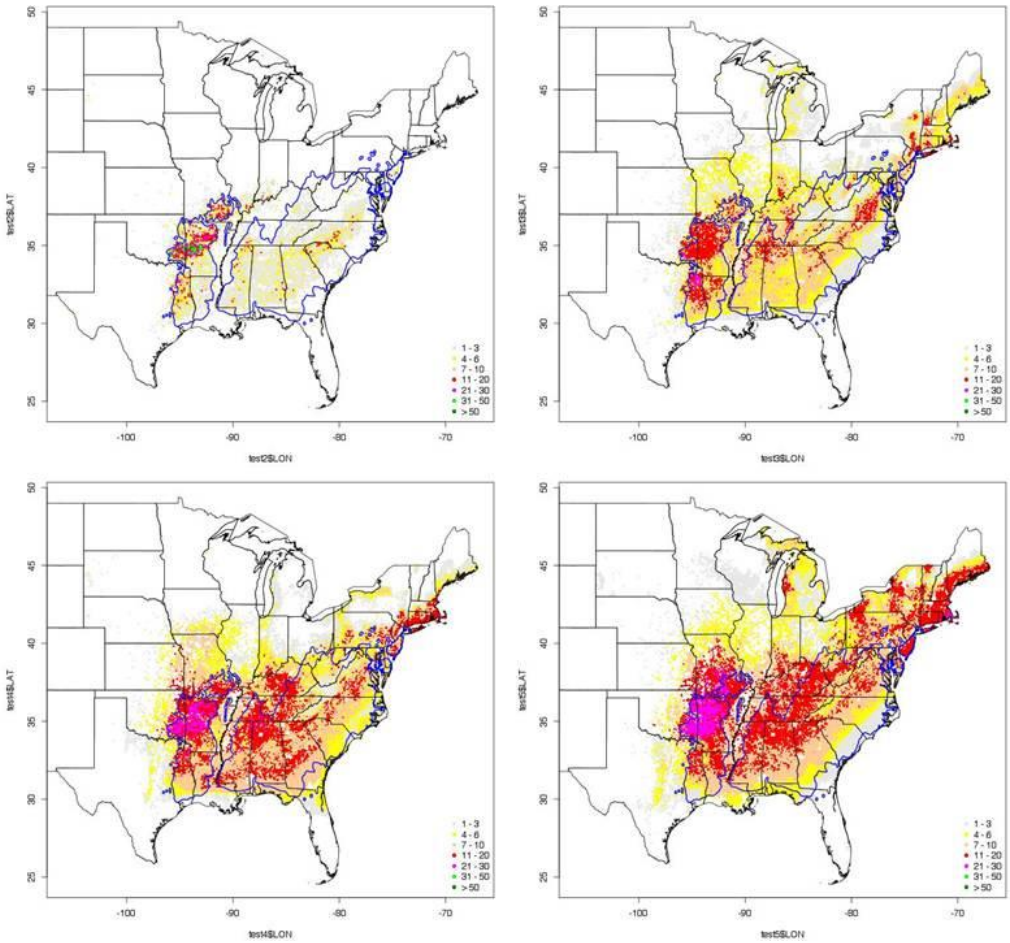


Figure A - 7. IV Shortleaf pine CGCM3 A2 models (top-left), 2030 (top-right), 2060 (bottom-left), 2090 (bottom-right)

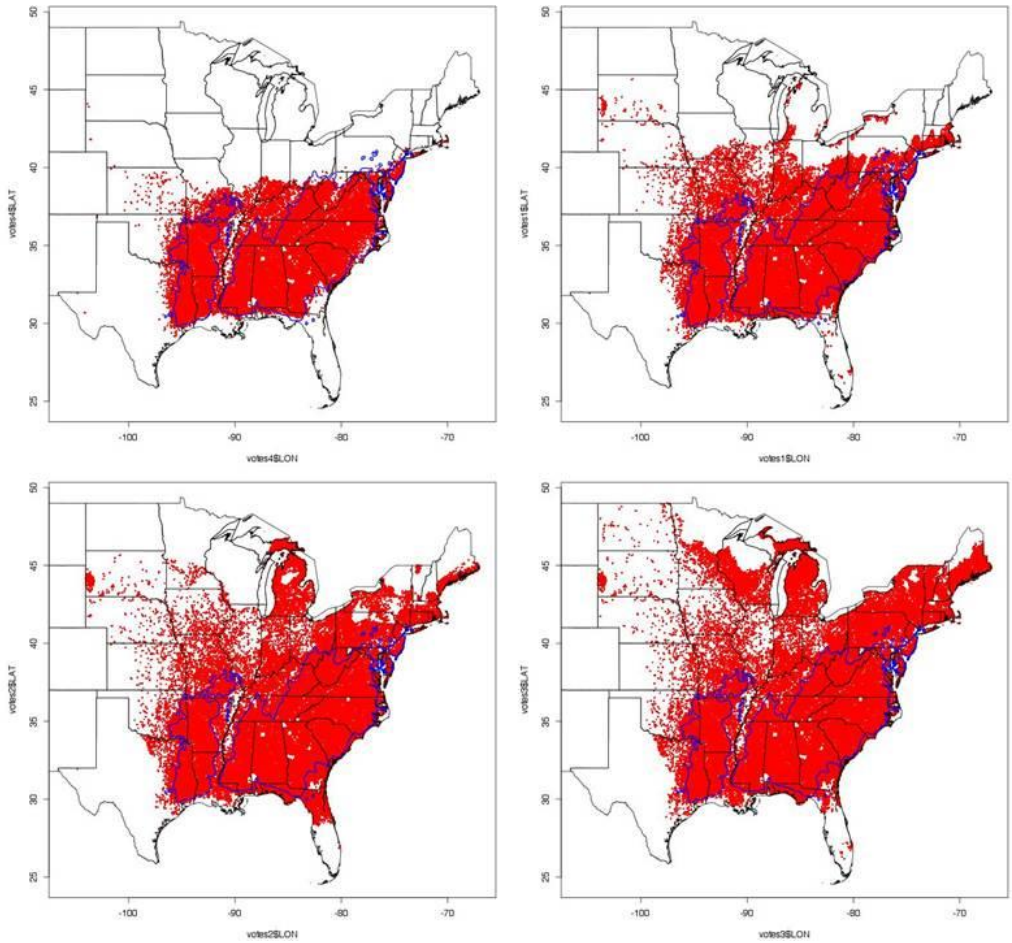


Figure A - 8. PA Shortleaf pine CGCM3 A2 models (top-left), 2030 (top-right), 2060 (bottom-left), 2090 (bottom-right)

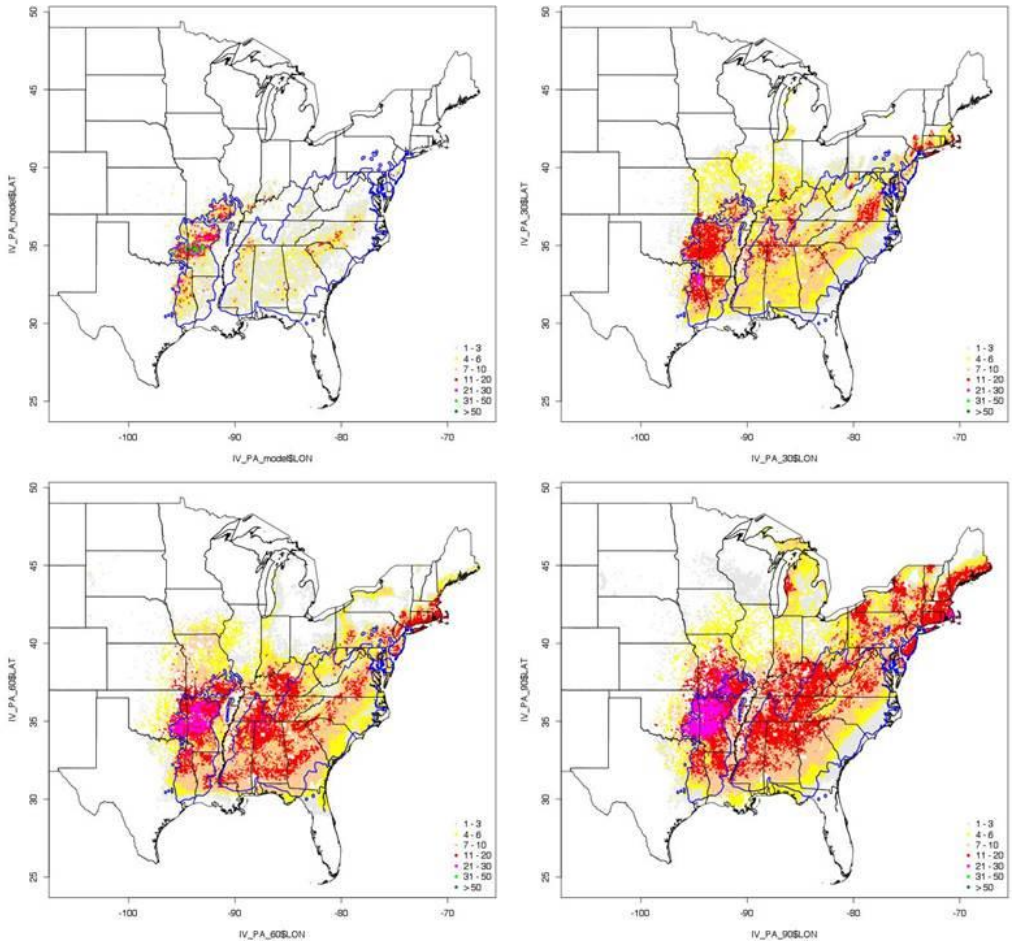


Figure A - 9. Combined IV and PA Shortleaf pine CGCM3 A2 models (top-left), 2030 (top-right), 2060 (bottom-left), 2090 (bottom-right)

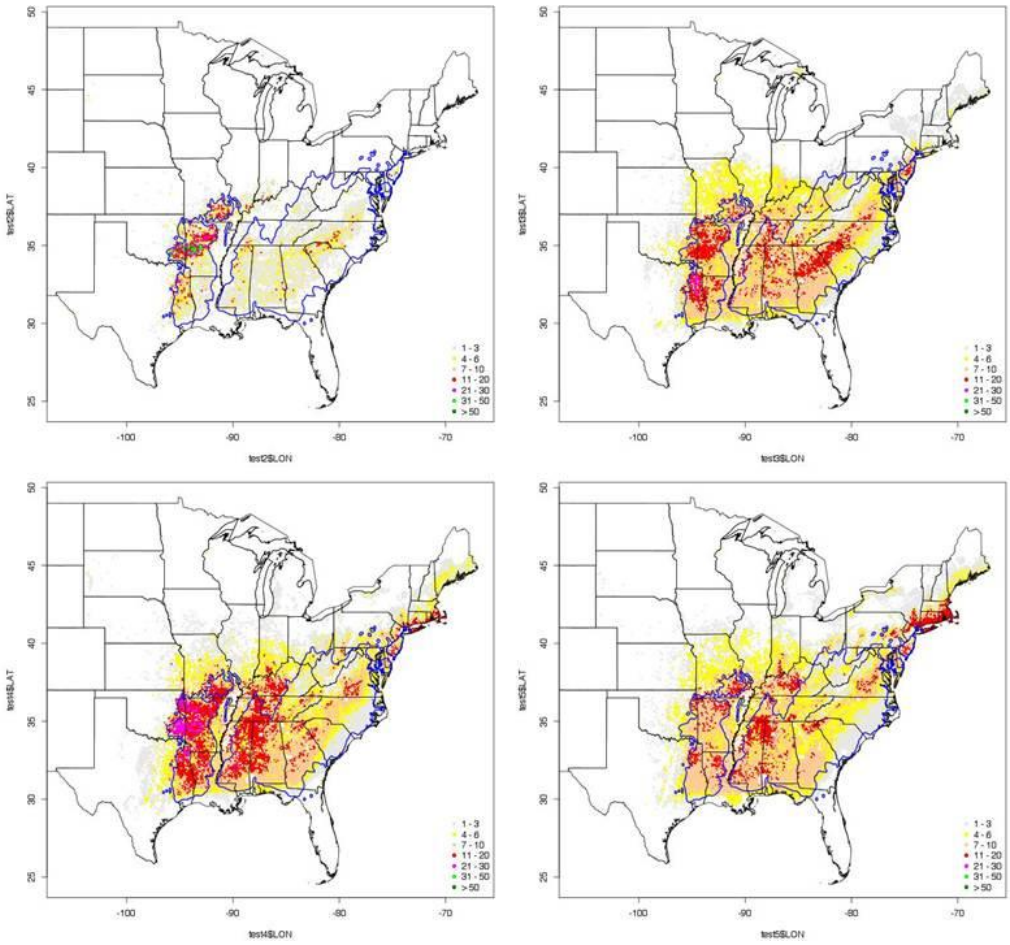


Figure A - 10. IV Shortleaf pine CGCM3 B1 models (top-left), 2030 (top-right), 2060 (bottom-left), 2090 (bottom-right)

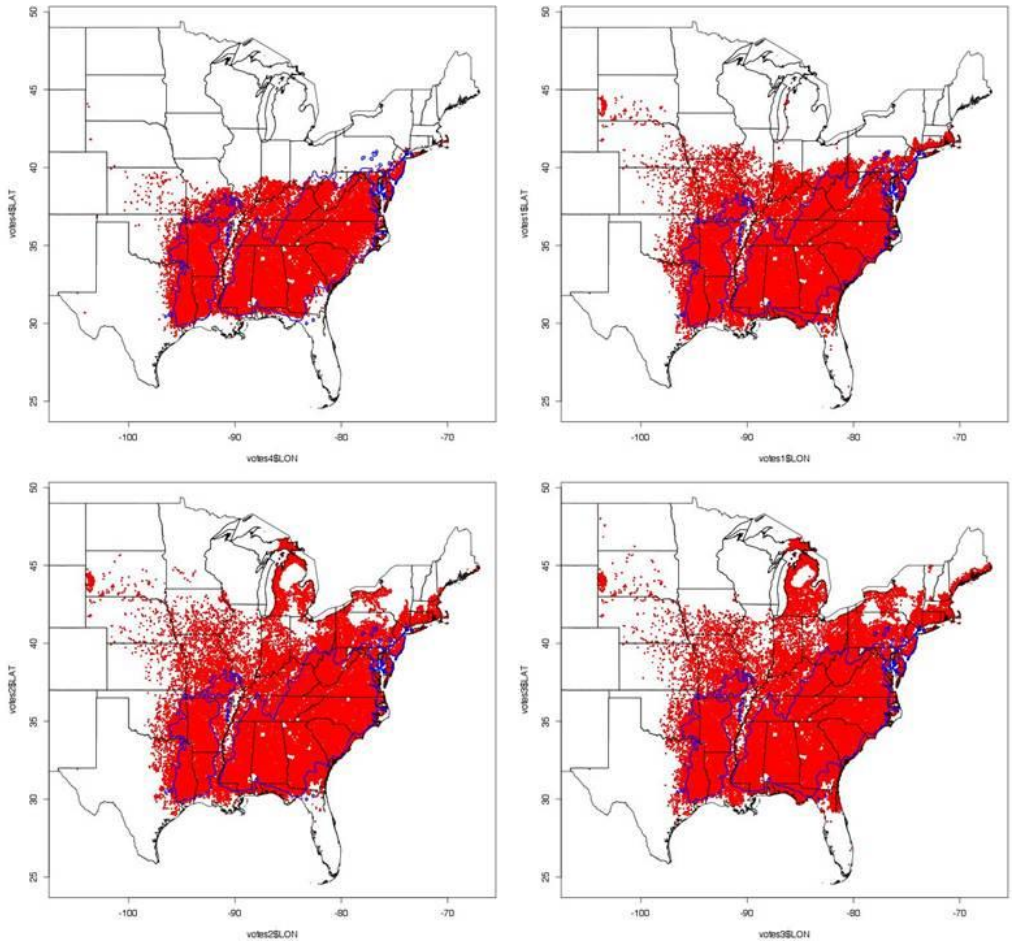


Figure A - 11. PA Shortleaf pine CGCM3 B1 models (top-left), 2030 (top-right), 2060 (bottom-left), 2090 (bottom-right)

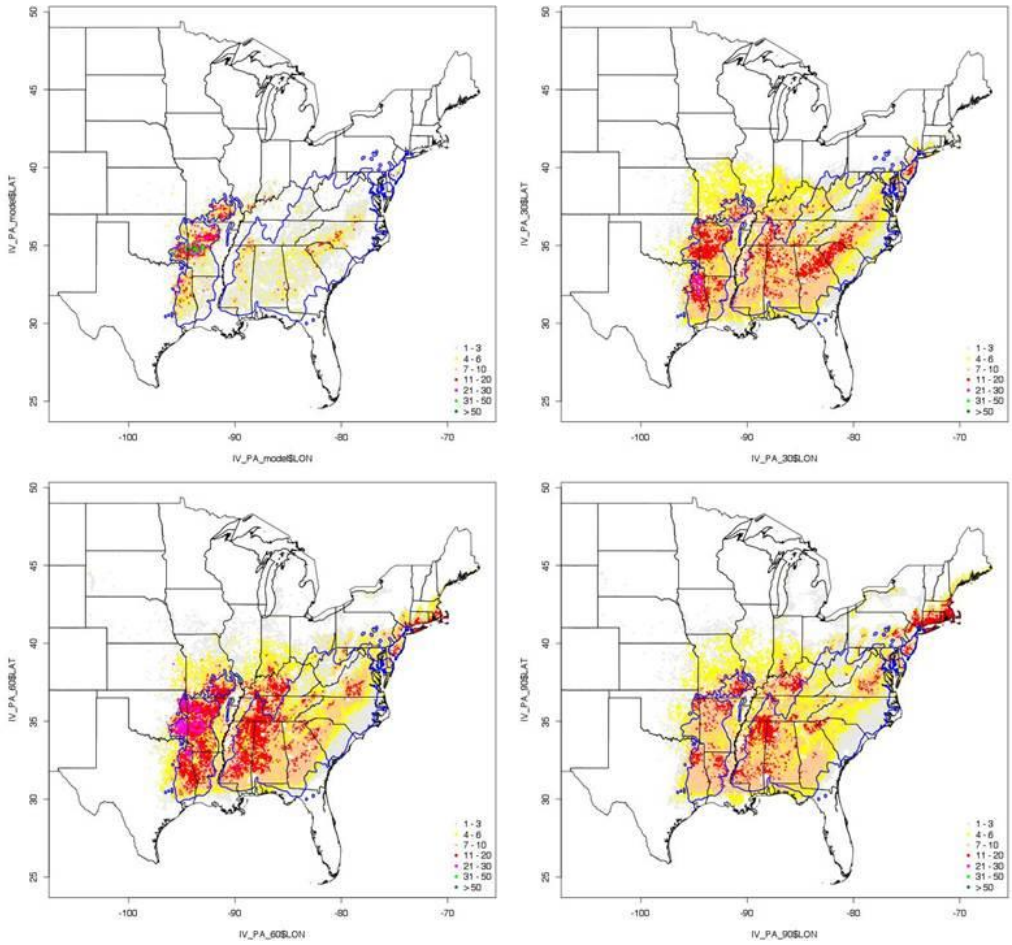


Figure A - 12. Combined IV and PA Shortleaf pine CGCM3 B1 models (top-left), 2030 (top-right), 2060 (bottom-left), 2090 (bottom-right)

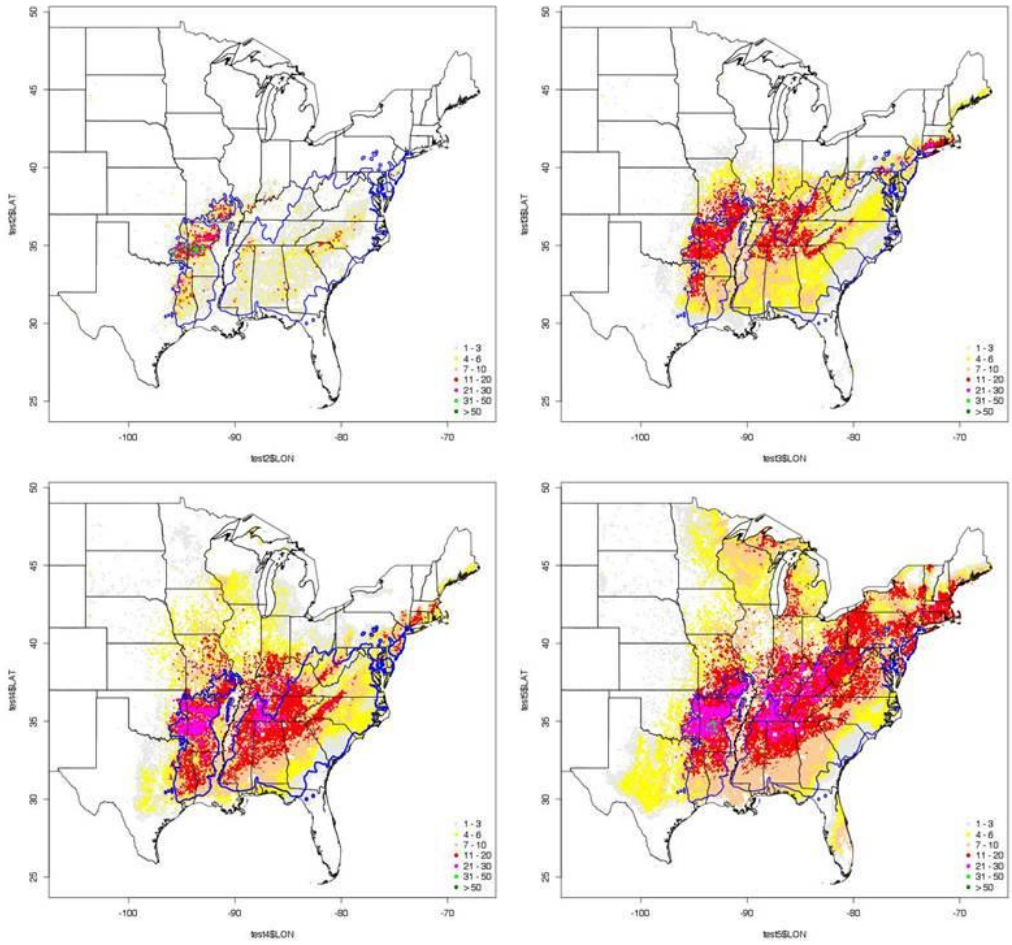


Figure A - 13. IV Shortleaf pine GFDLCM21 A2 models (top-left), 2030 (top-right), 2060 (bottom-left), 2090 (bottom-right)

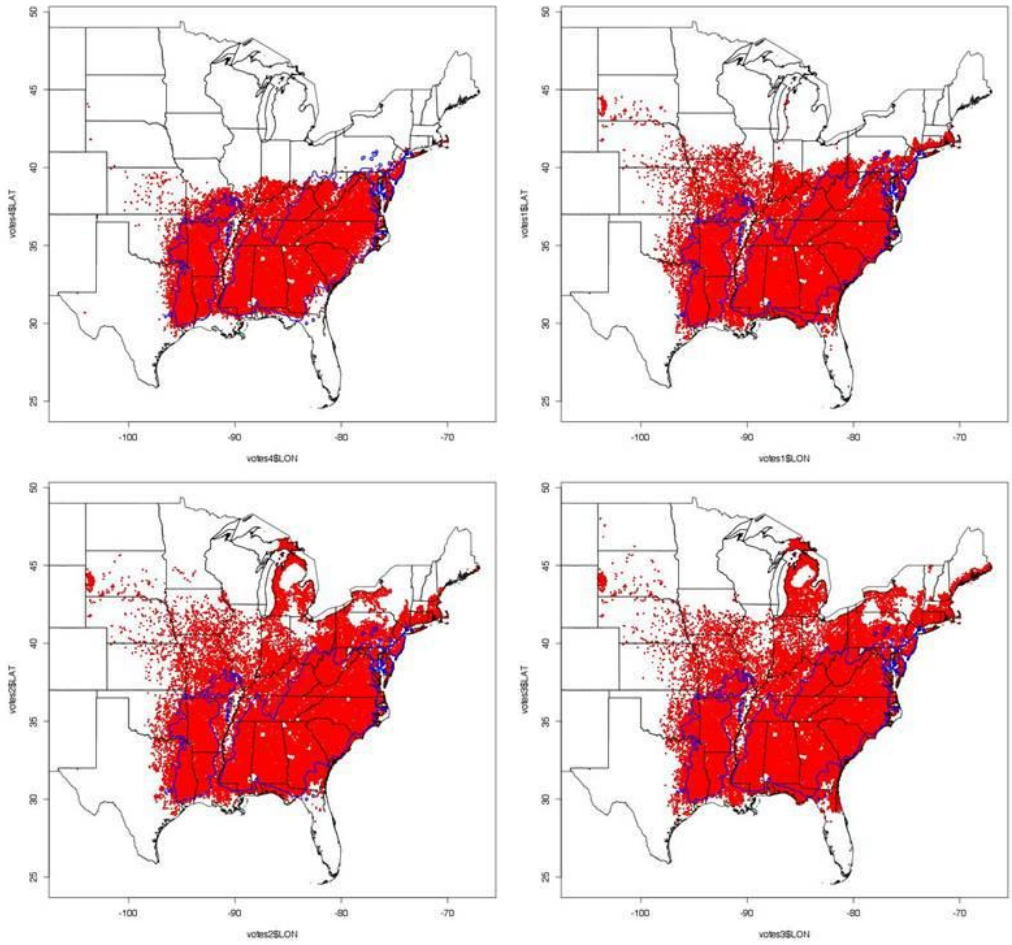


Figure A - 14. PA Shortleaf pine GFDLCM21 A2 models (top-left), 2030 (top-right), 2060 (bottom-left), 2090 (bottom-right)

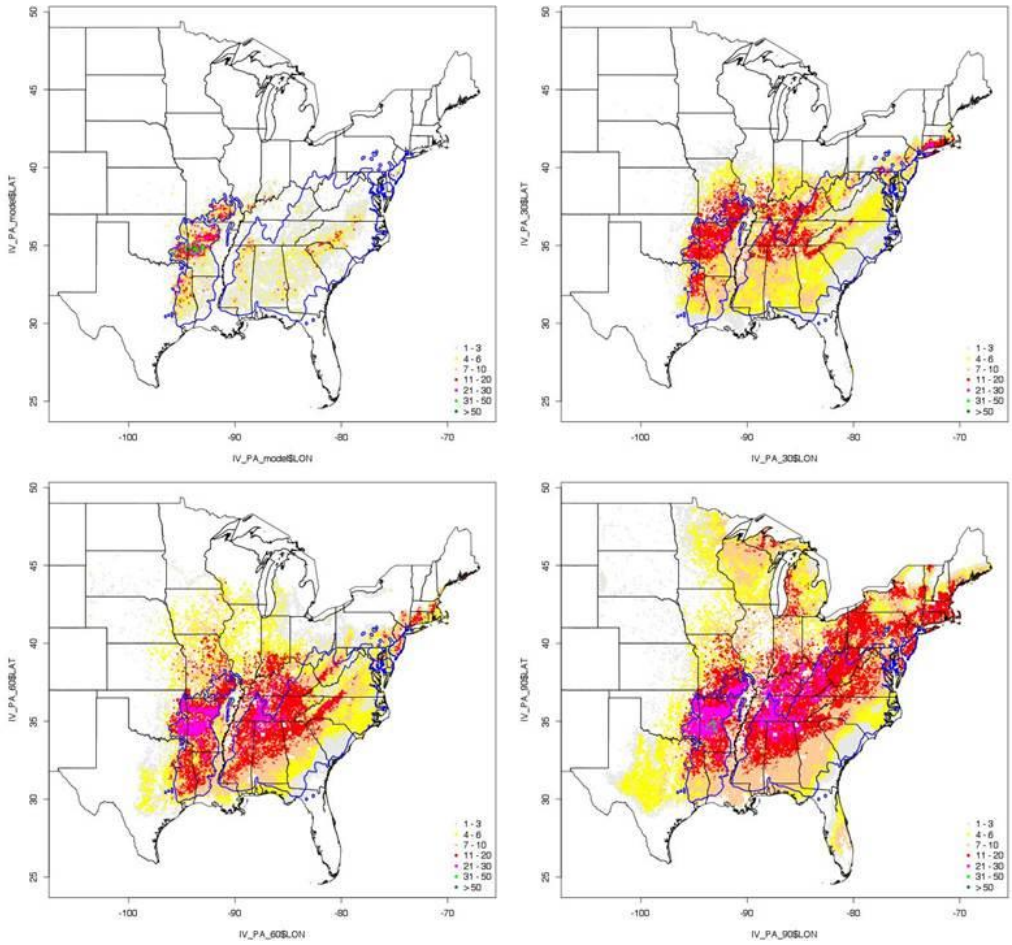


Figure A - 15. Combined IV and PA Shortleaf pine GFDLCM21 A2 models (top-left), 2030 (top-right), 2060 (bottom-left), 2090 (bottom-right)

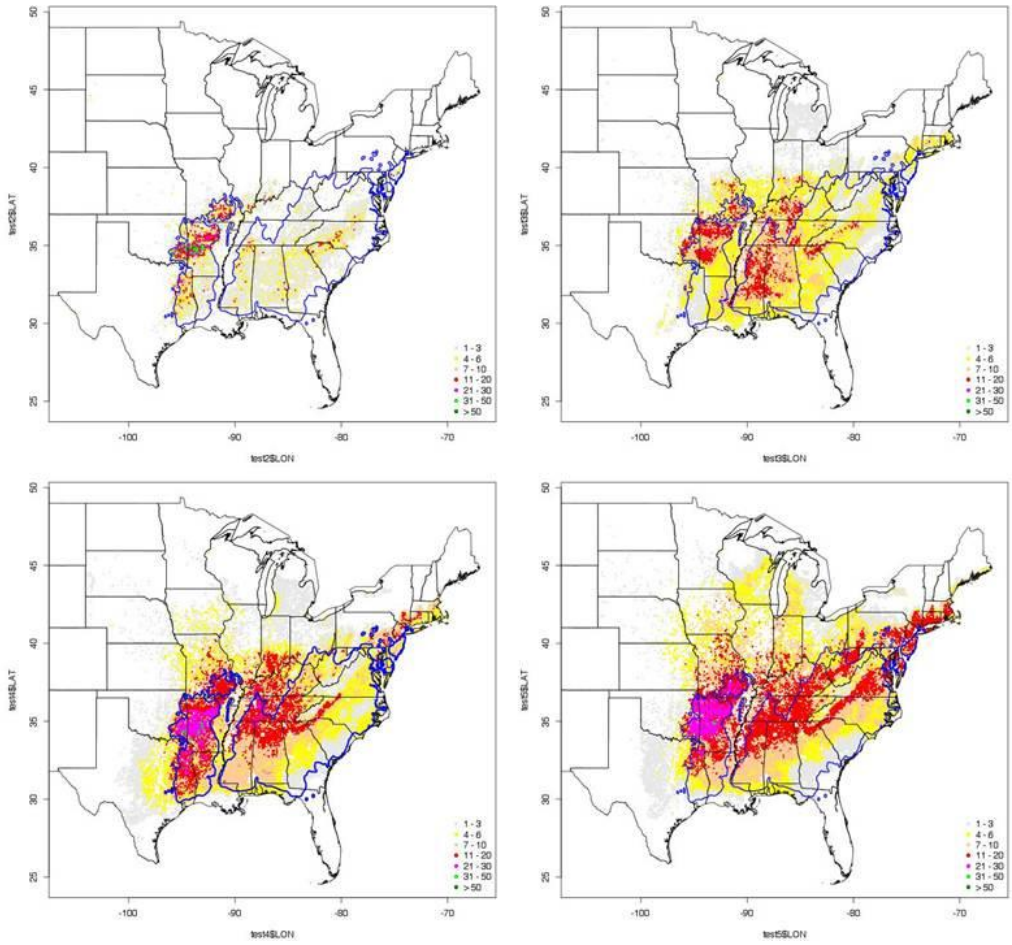


Figure A - 16. IV Shortleaf pine GFDLCM21 B1 models (top-left), 2030 (top-right), 2060 (bottom-left), 2090 (bottom-right)

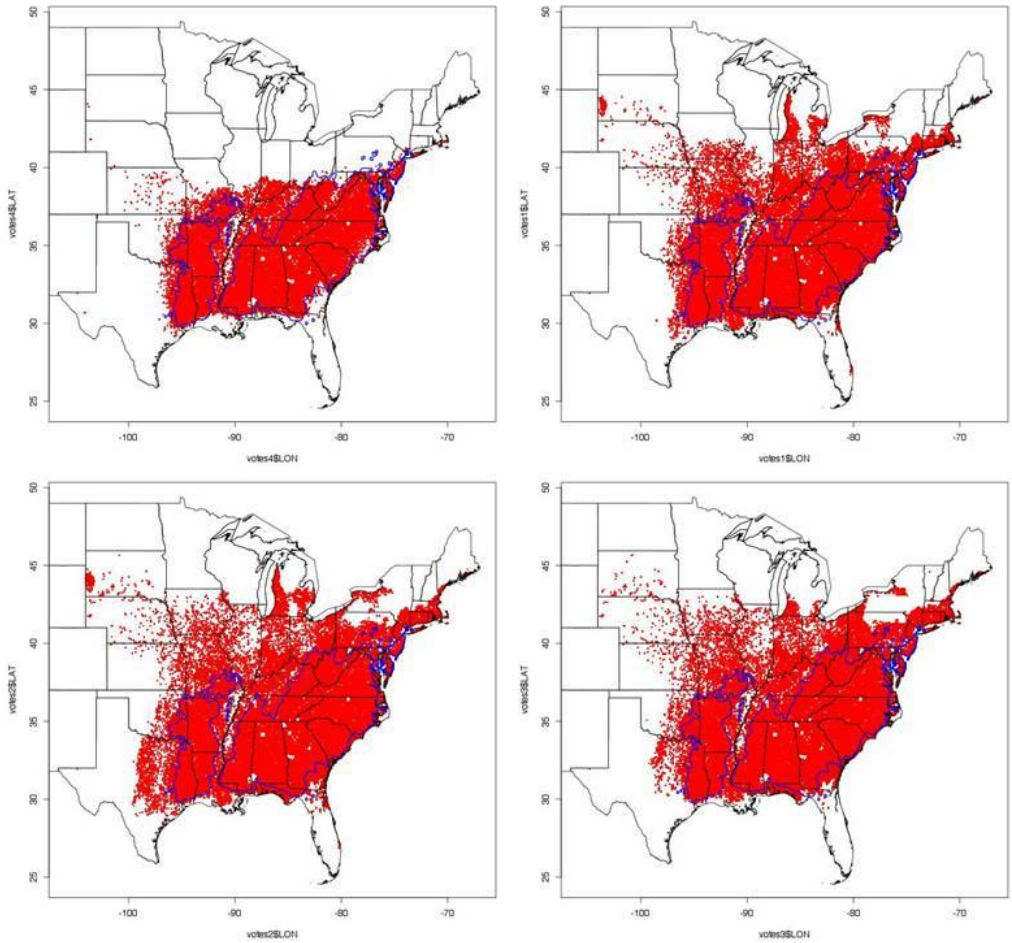


Figure A - 17. PA Shortleaf pine GFDLCM21 B1 models (top-left), 2030 (top-right), 2060 (bottom-left), 2090 (bottom-right)

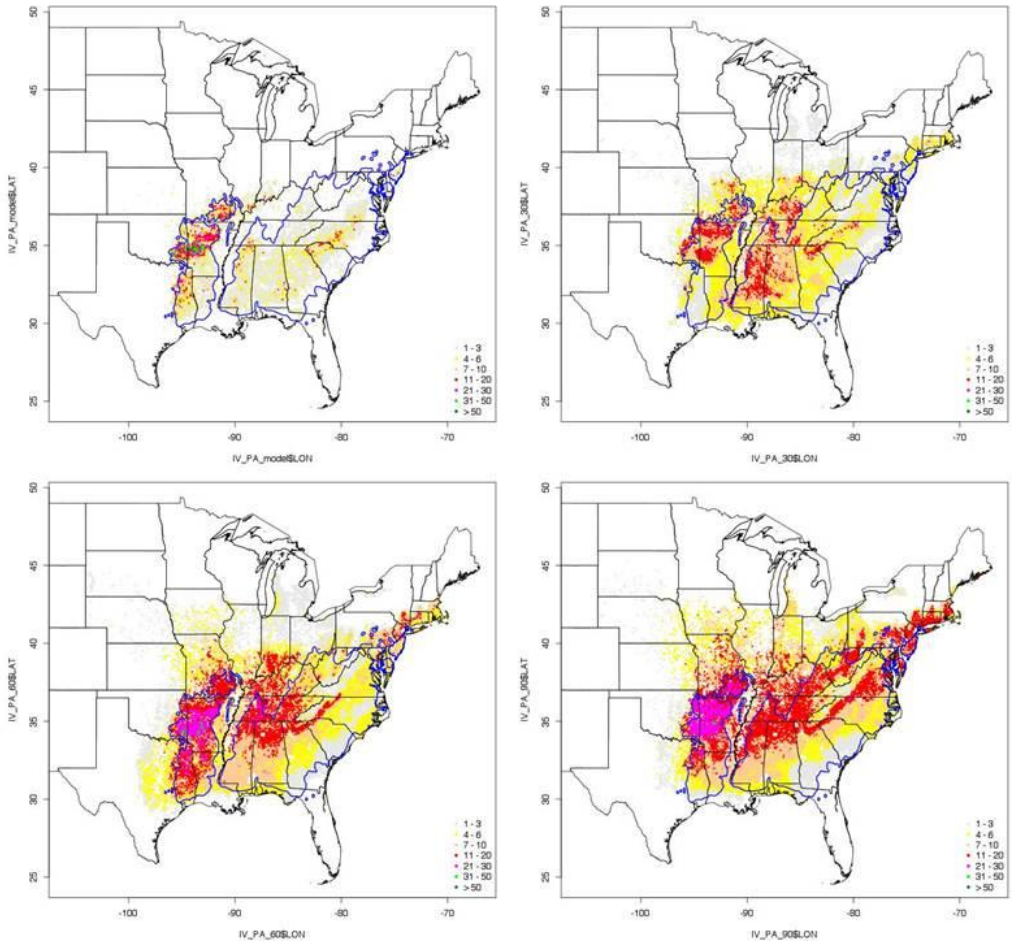


Figure A - 18. Combined IV and PA Shortleaf pine GFDLCM21 B1 models (top-left), 2030 (top-right), 2060 (bottom-left), 2090 (bottom-right)

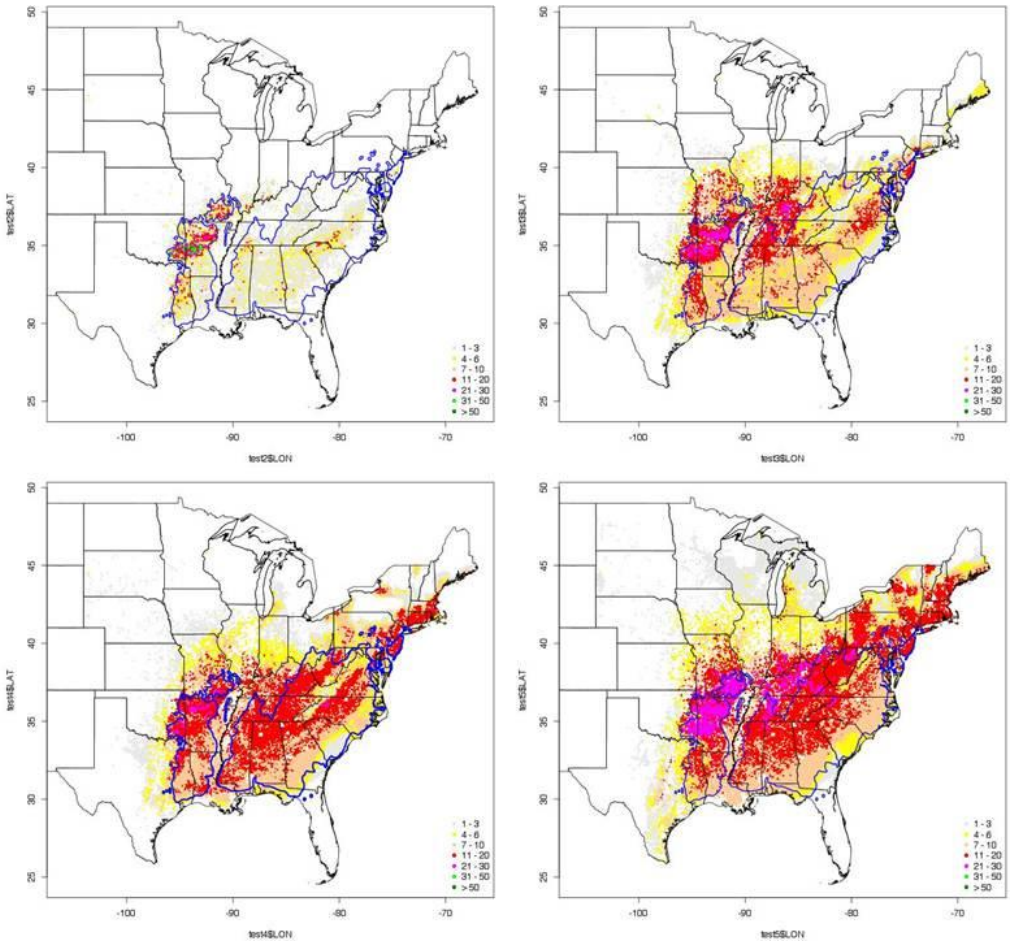


Figure A - 19. IV Shortleaf pine HADCM3 A2 models (top-left), 2030 (top-right), 2060 (bottom-left), 2090 (bottom-right)

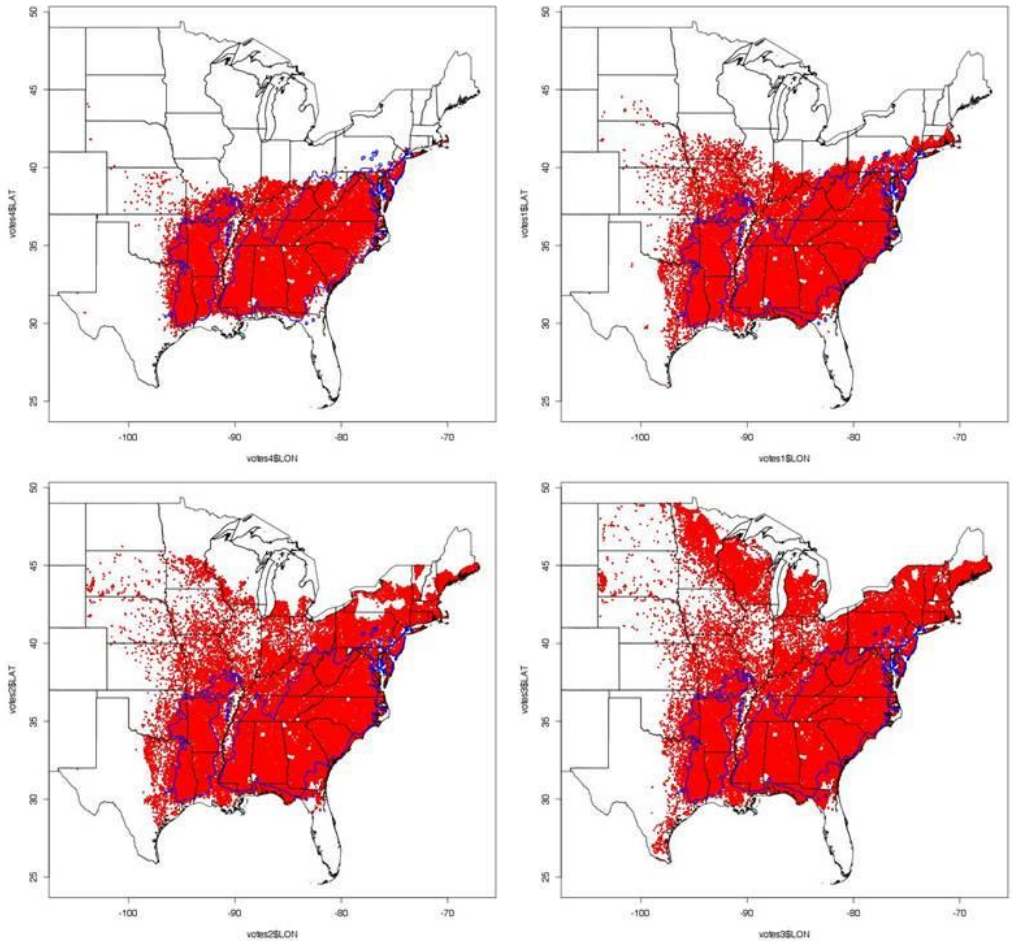


Figure A - 20. PA Shortleaf pine HADCM3 A2 models (top-left), 2030 (top-right), 2060 (bottom-left), 2090 (bottom-right)

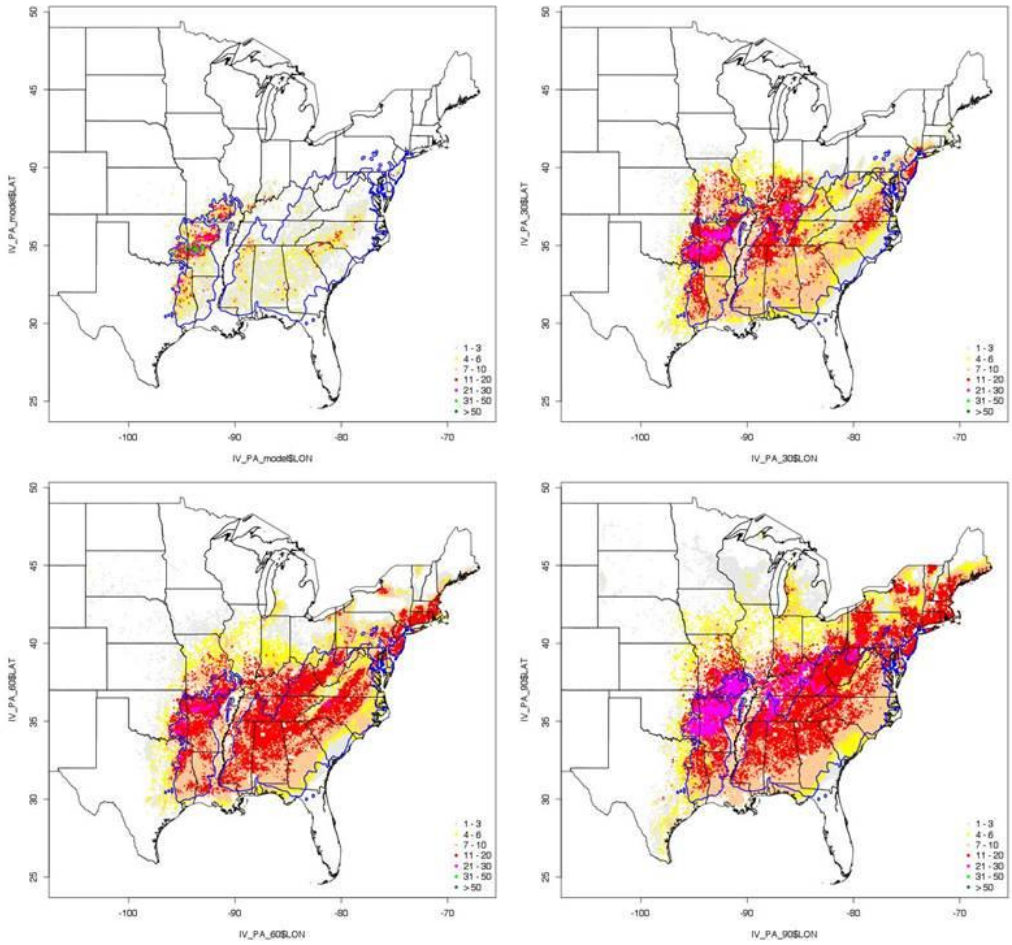


Figure A - 21. Combined IV and PA Shortleaf pine HADCM3 A2 models (top-left), 2030 (top-right), 2060 (bottom-left), 2090 (bottom-right)

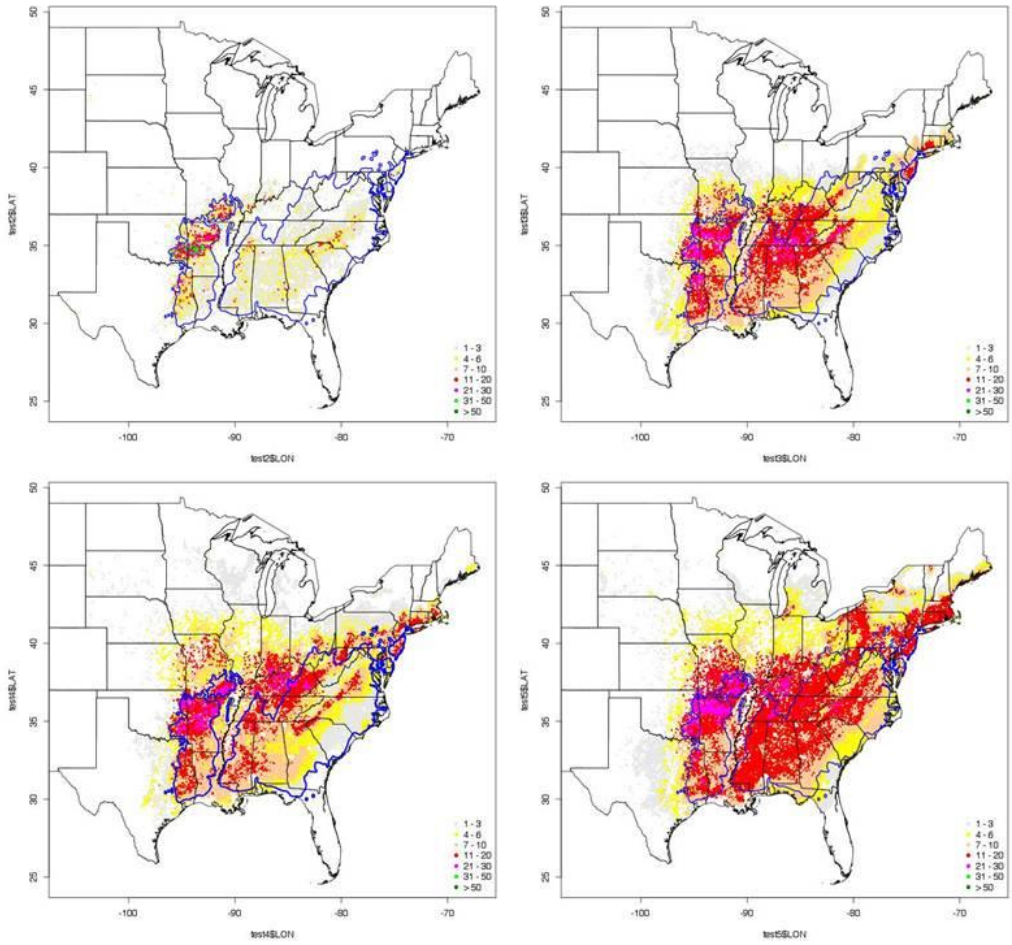


Figure A - 22. IV Shortleaf pine HADCM3 B2 models (top-left), 2030 (top-right), 2060 (bottom-left), 2090 (bottom-right)

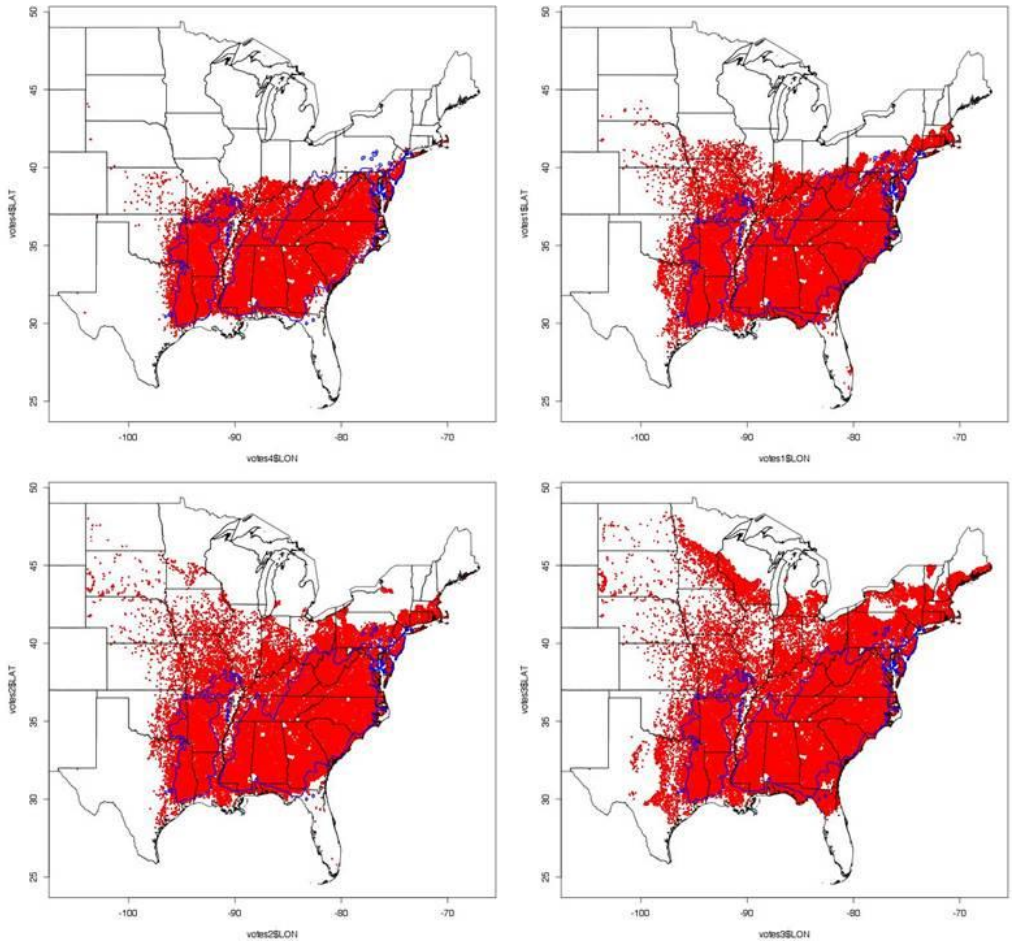


Figure A - 23. PA Shortleaf pine HADCM3 B2 models (top-left), 2030 (top-right), 2060 (bottom-left), 2090 (bottom-right)

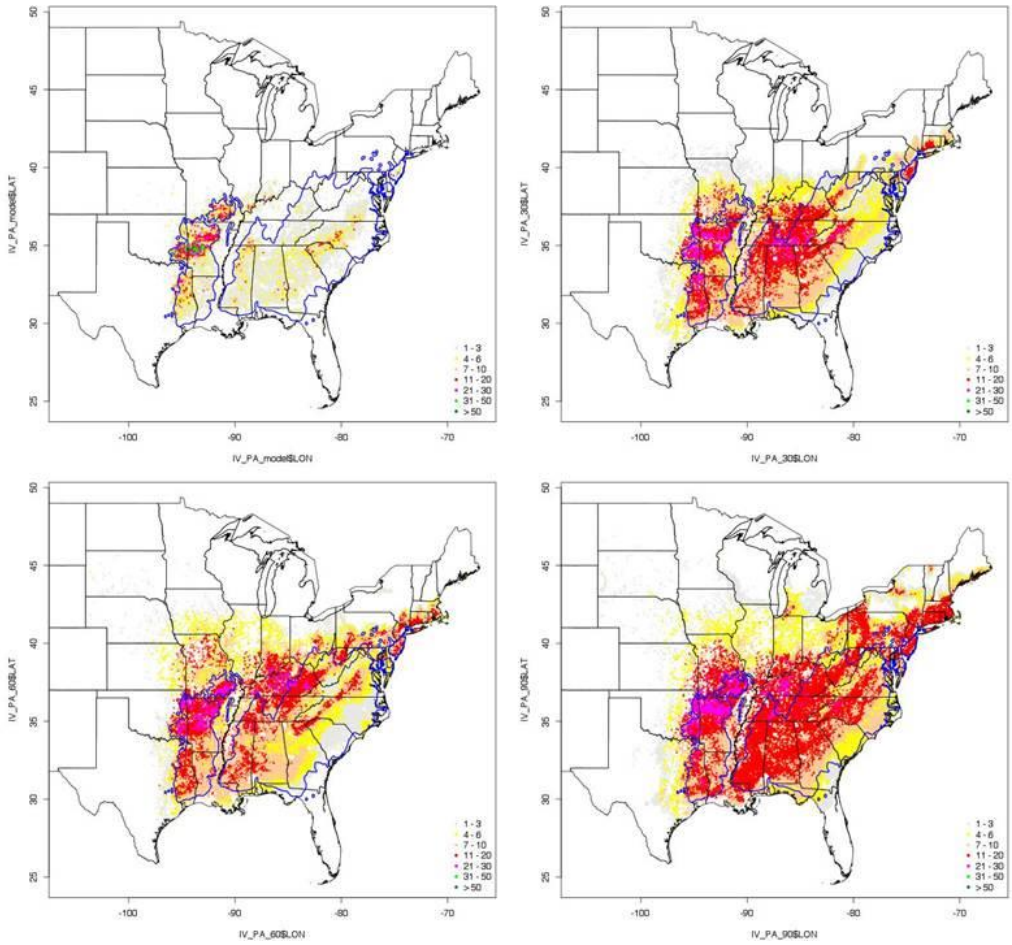


Figure A - 24. Combined IV and PA Shortleaf pine HADCM3 B2 models (top-left), 2030 (top-right), 2060 (bottom-left), 2090 (bottom-right)

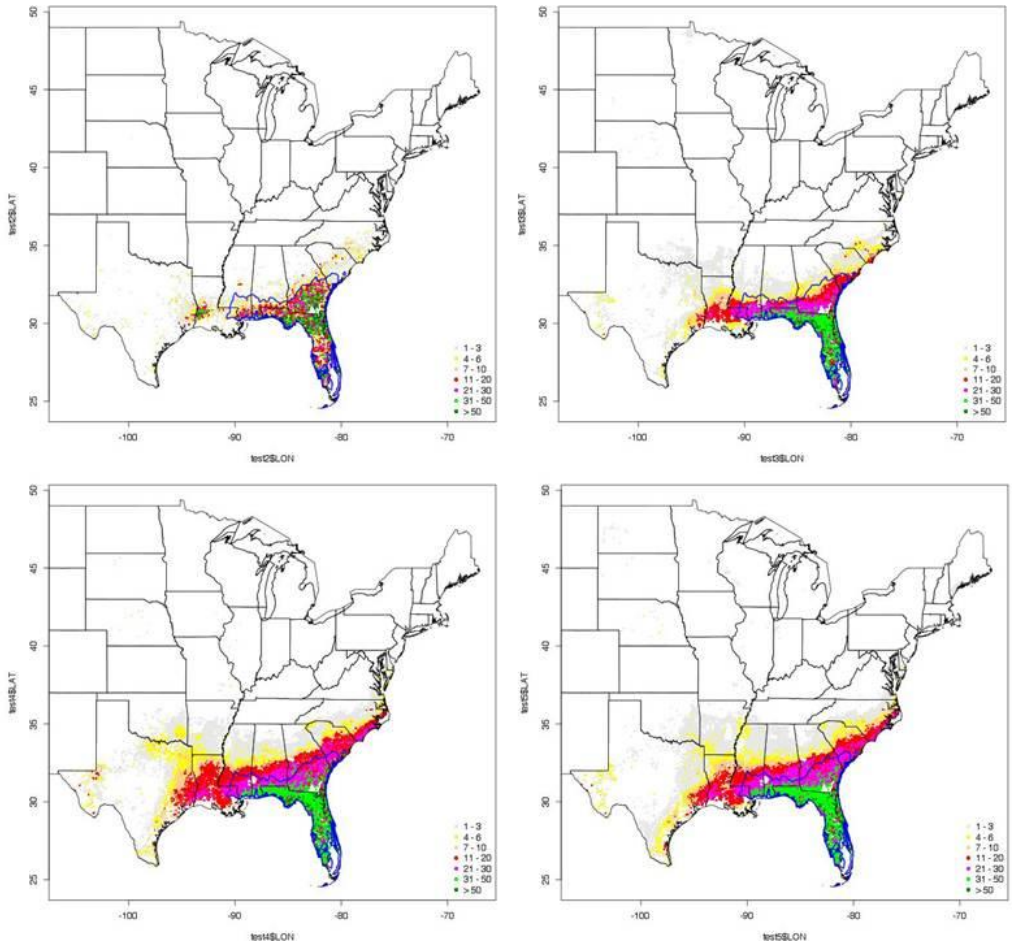


Figure A - 25. IV Slash pine CGCM3 A1B models (top-left), 2030 (top-right), 2060 (bottom-left), 2090 (bottom-right)

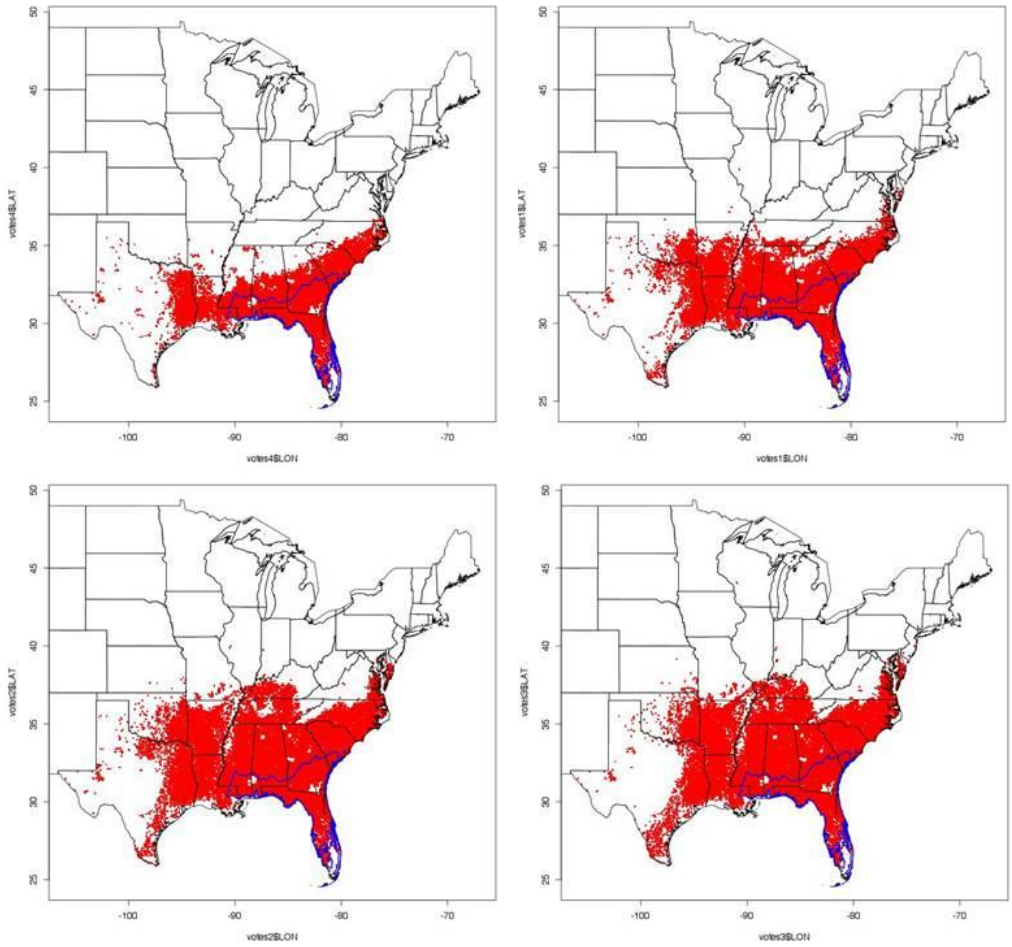


Figure A - 26. PA Slash pine CGCM3 A1B models (top-left), 2030 (top-right), 2060 (bottom-left), 2090 (bottom-right)

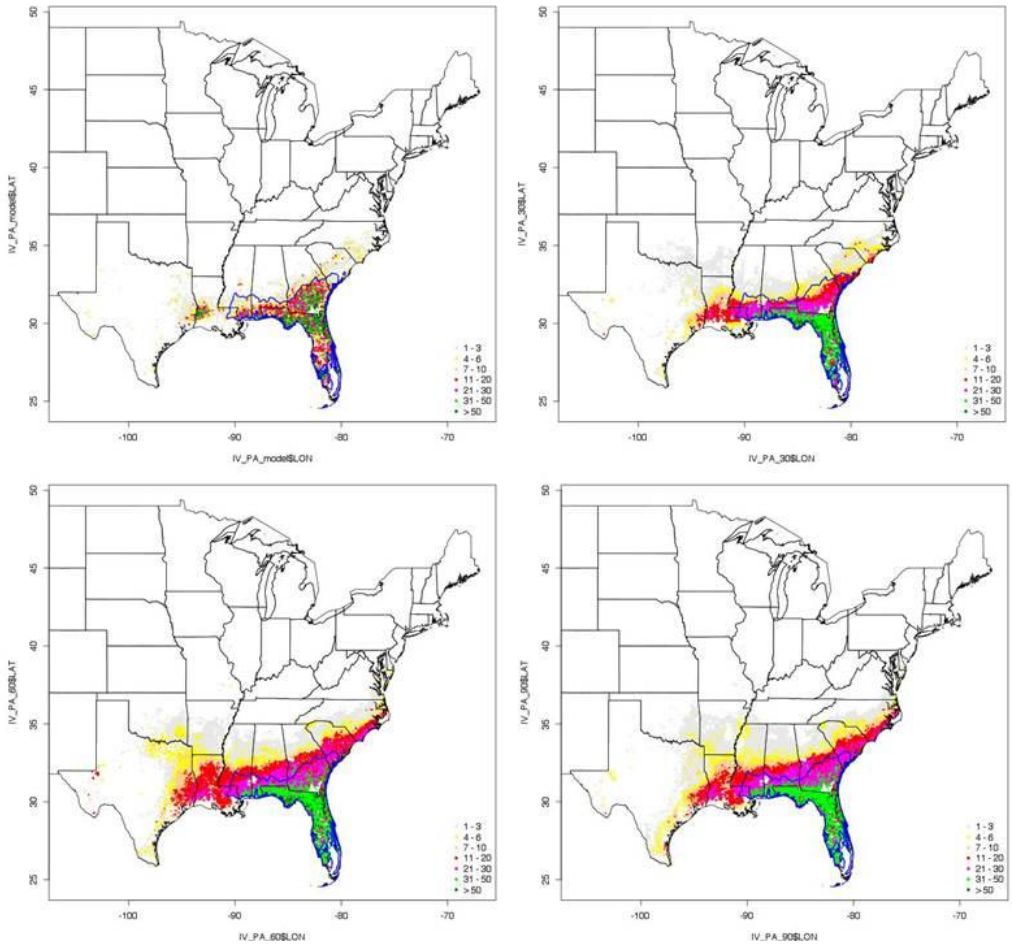


Figure A - 27. Combined IV and PA Slash pine CGCM3 A1B models (top-left), 2030 (top-right), 2060 (bottom-left), 2090 (bottom-right)

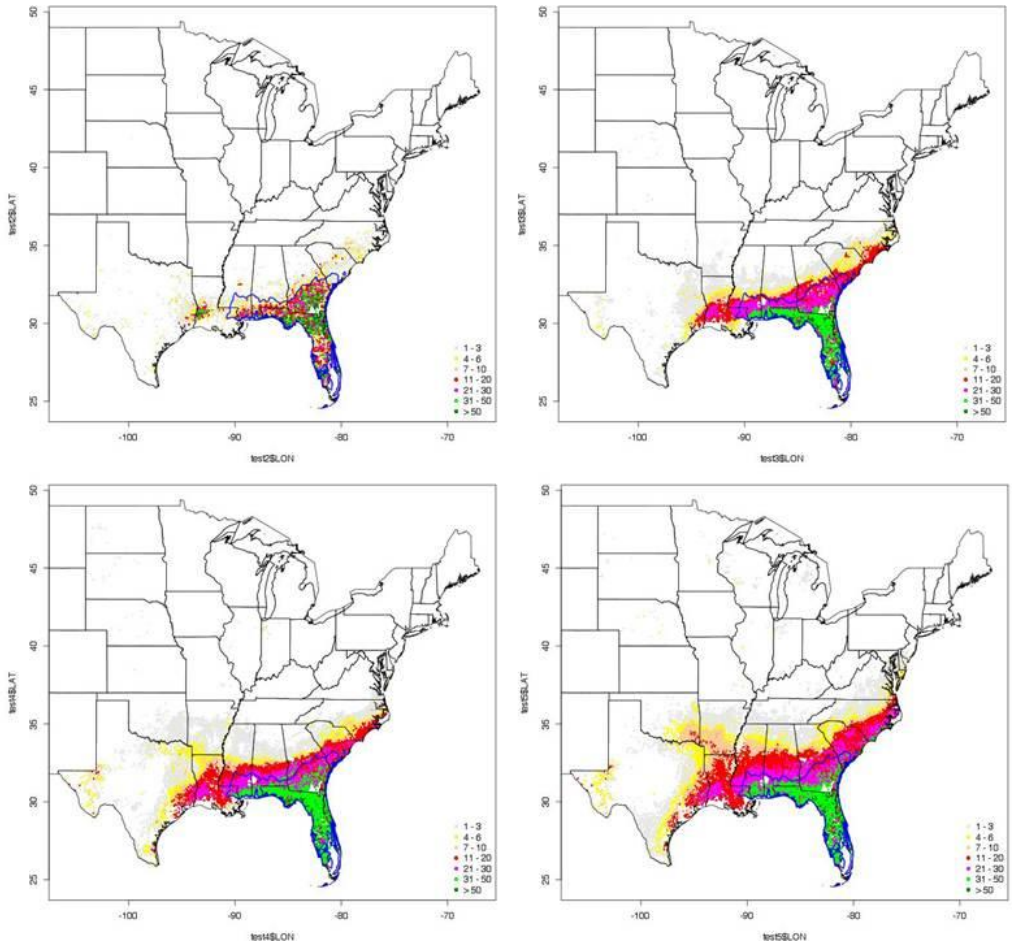


Figure A - 28. IV Slash pine CGCM3 A2 models (top-left), 2030 (top-right), 2060 (bottom-left), 2090 (bottom-right)

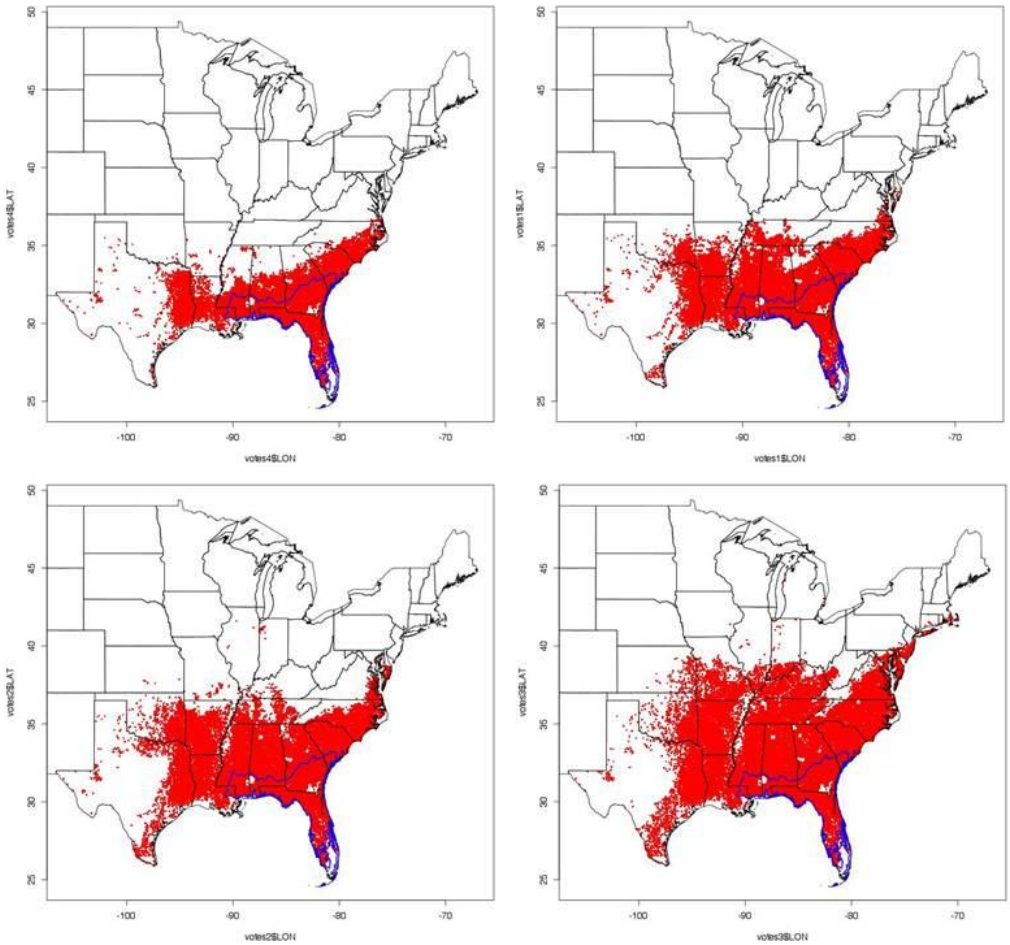


Figure A - 29. PA Slash pine CGCM3 A2 models (top-left), 2030 (top-right), 2060 (bottom-left), 2090 (bottom-right)

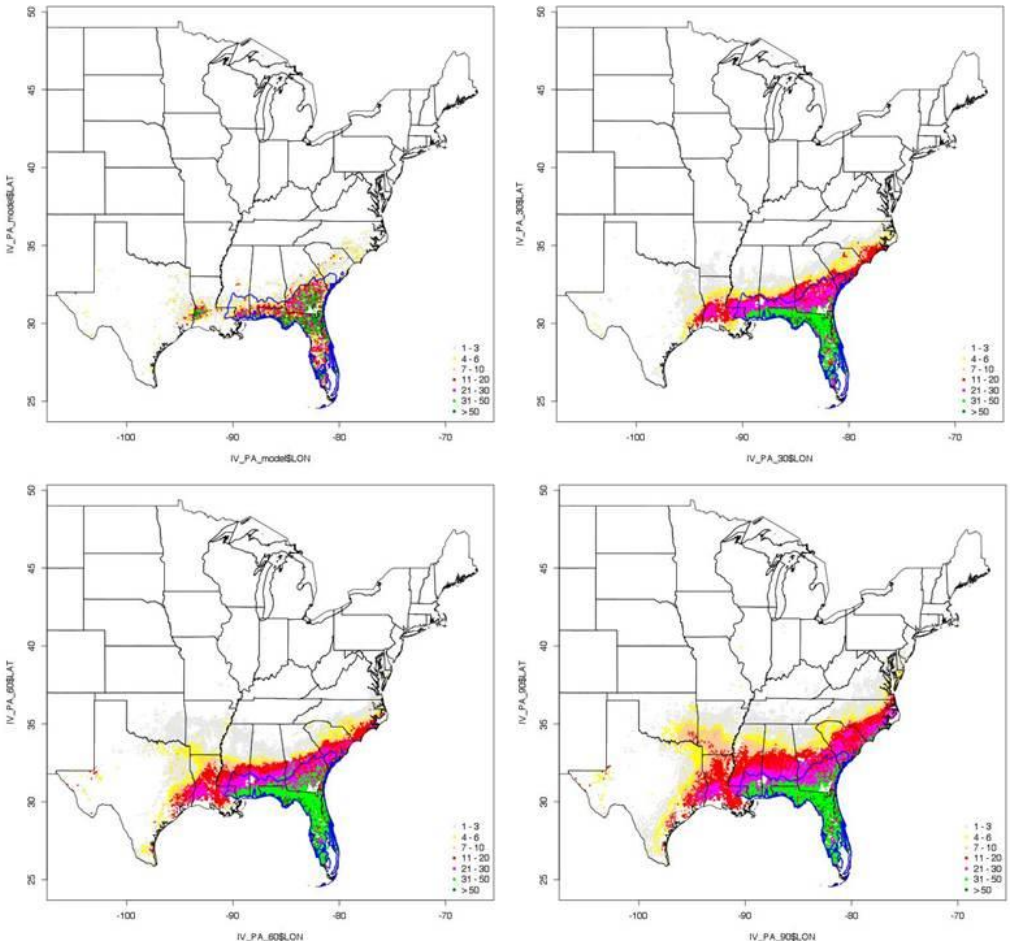


Figure A - 30. Combined IV and PA Slash pine CGCM3 A2 models (top-left), 2030 (top-right), 2060 (bottom-left), 2090 (bottom-right)

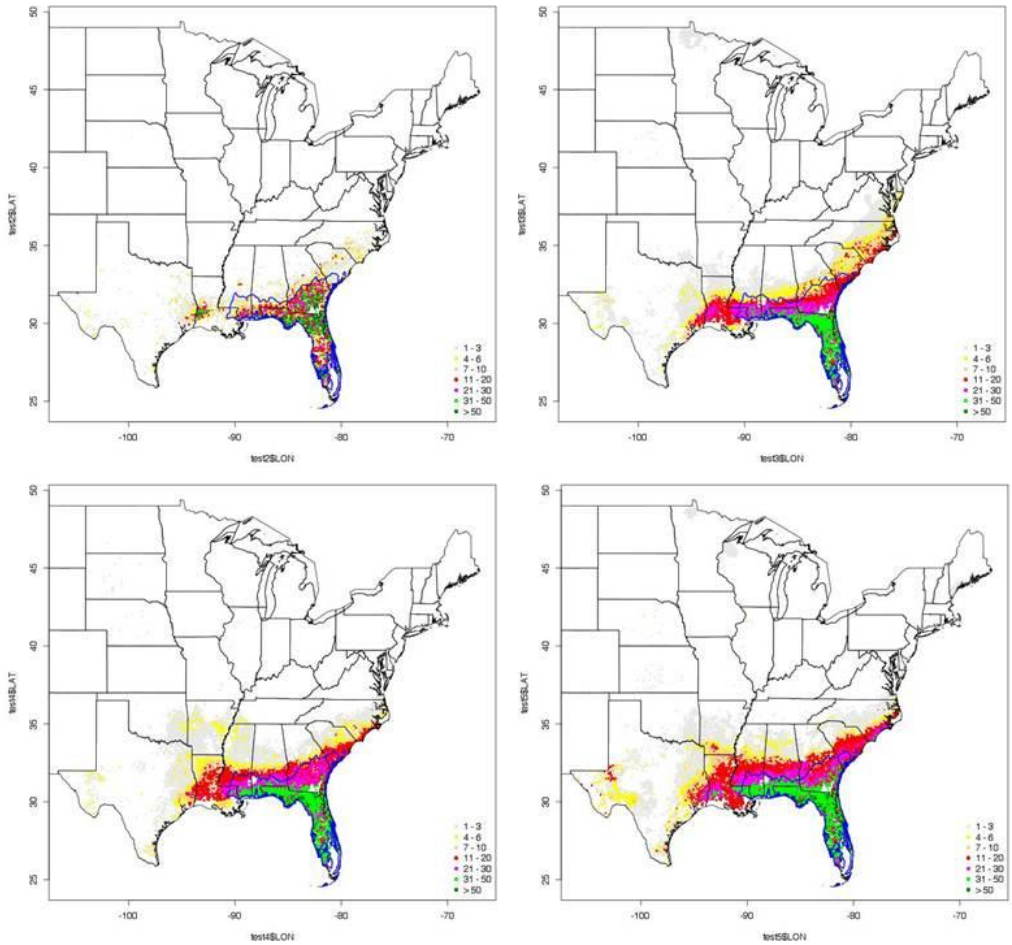


Figure A - 31. IV Slash pine CGCM3 B1 models (top-left), 2030 (top-right), 2060 (bottom-left), 2090 (bottom-right)

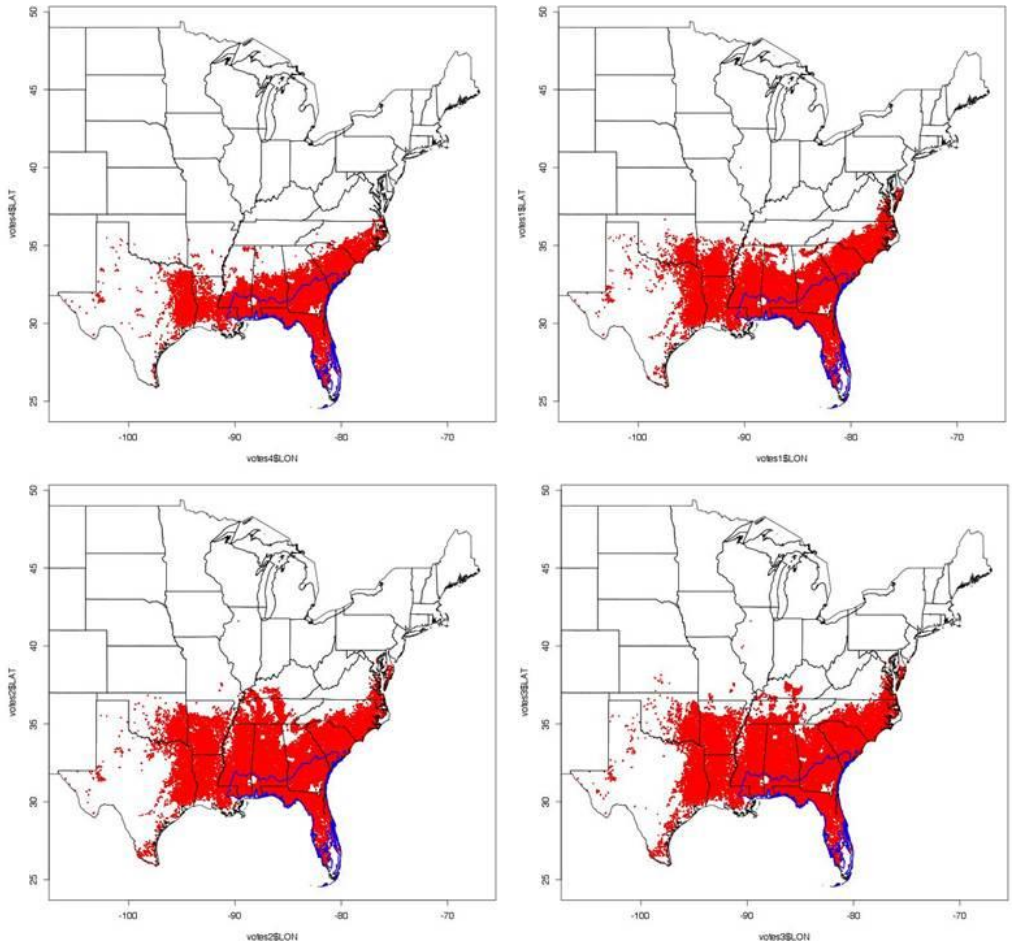


Figure A - 32. PA Slash pine CGCM3 B1 models (top-left), 2030 (top-right), 2060 (bottom-left), 2090 (bottom-right)

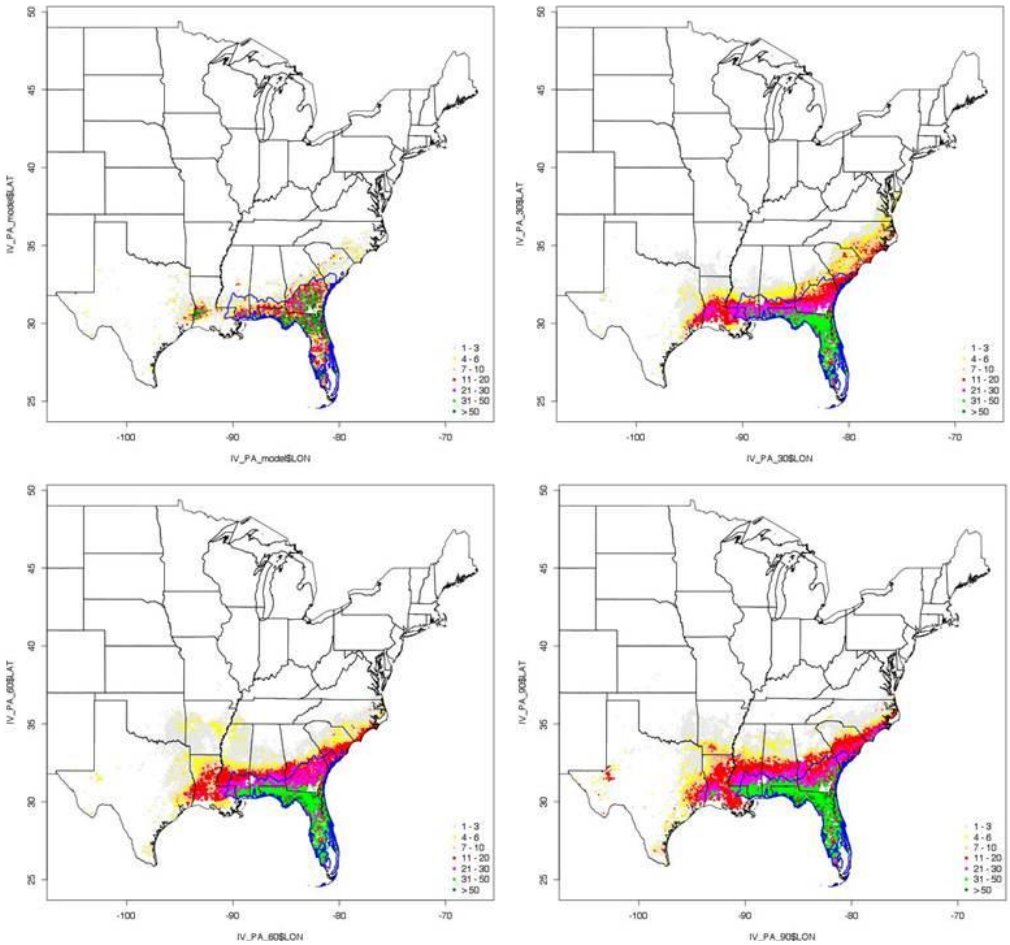


Figure A - 33. Combined IV and PA Slash pine CGCM3 B1 models (top-left), 2030 (top-right), 2060 (bottom-left), 2090 (bottom-right)

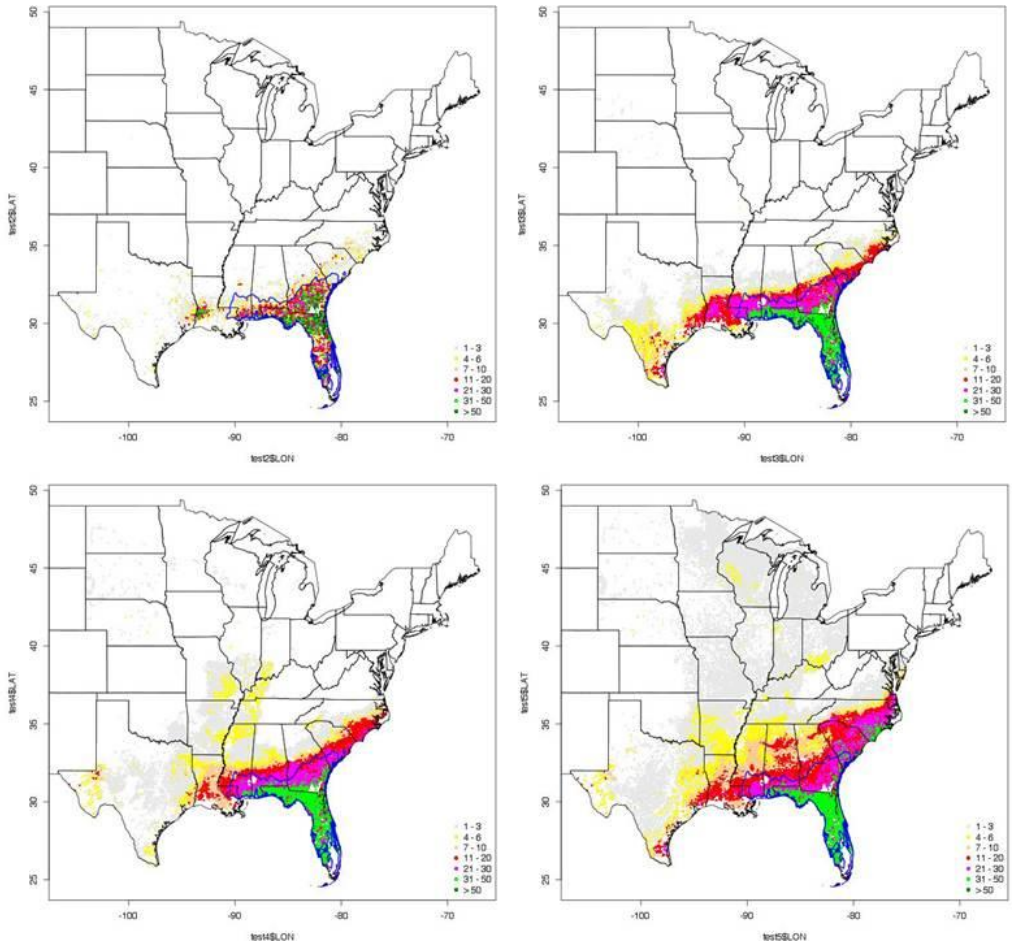


Figure A - 34. IV Slash pine GFDLCM21 A2 models (top-left), 2030 (top-right), 2060 (bottom-left), 2090 (bottom-right)

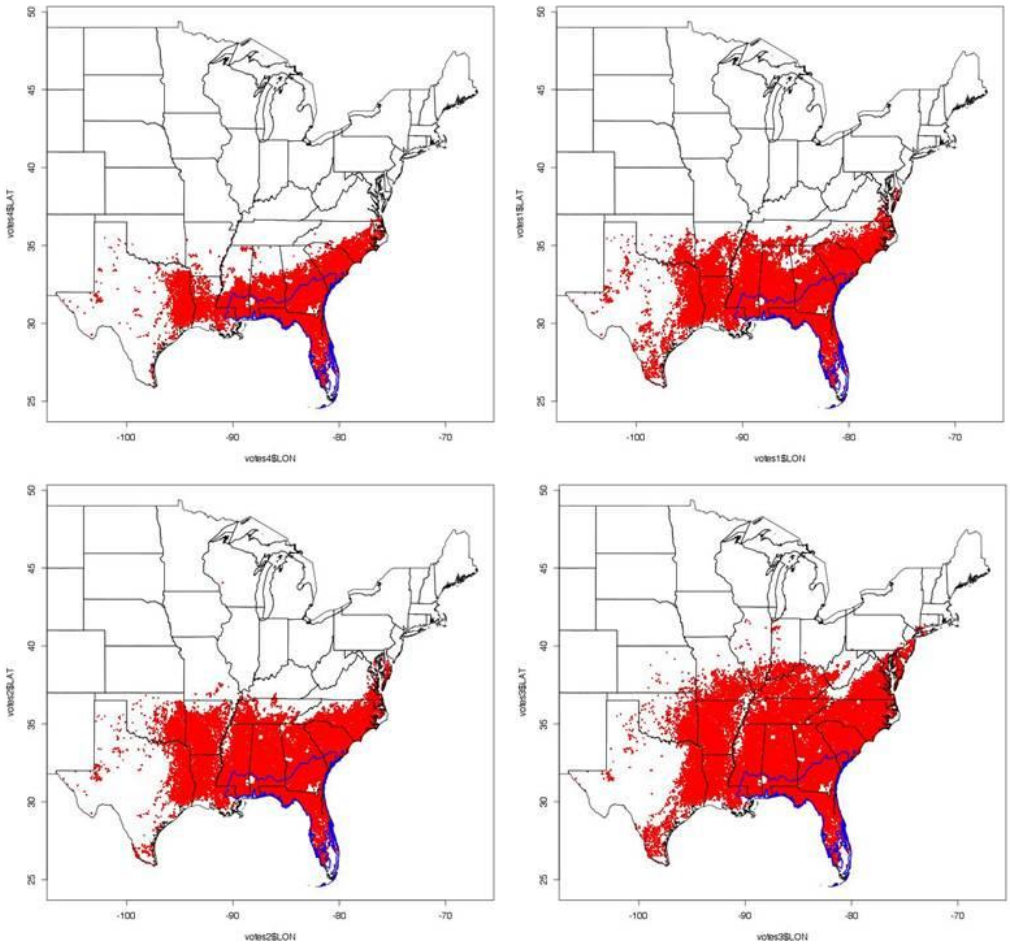


Figure A - 35. PA Slash pine GFDLCM21 A2 models (top-left), 2030 (top-right), 2060 (bottom-left), 2090 (bottom-right)

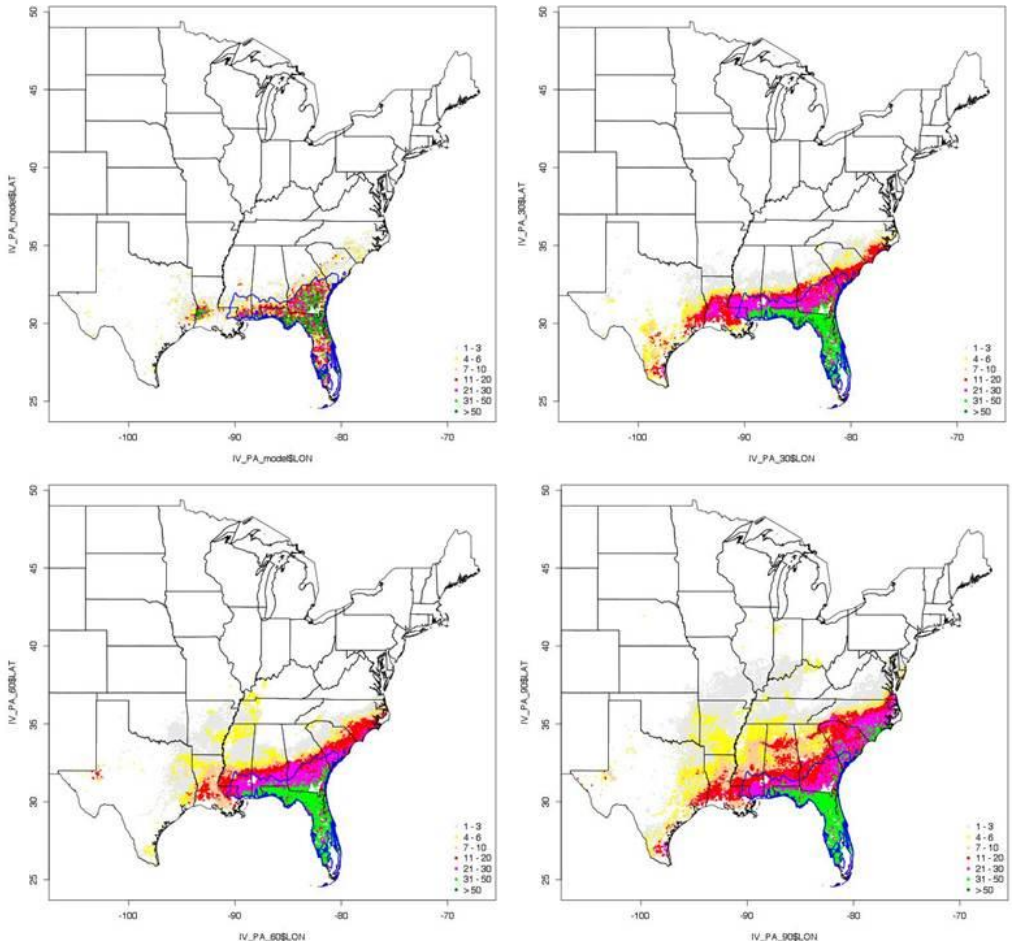


Figure A - 36. Combined IV and PA Slash pine GFDLCM21 A2 models (top-left), 2030 (top-right), 2060 (bottom-left), 2090 (bottom-right)

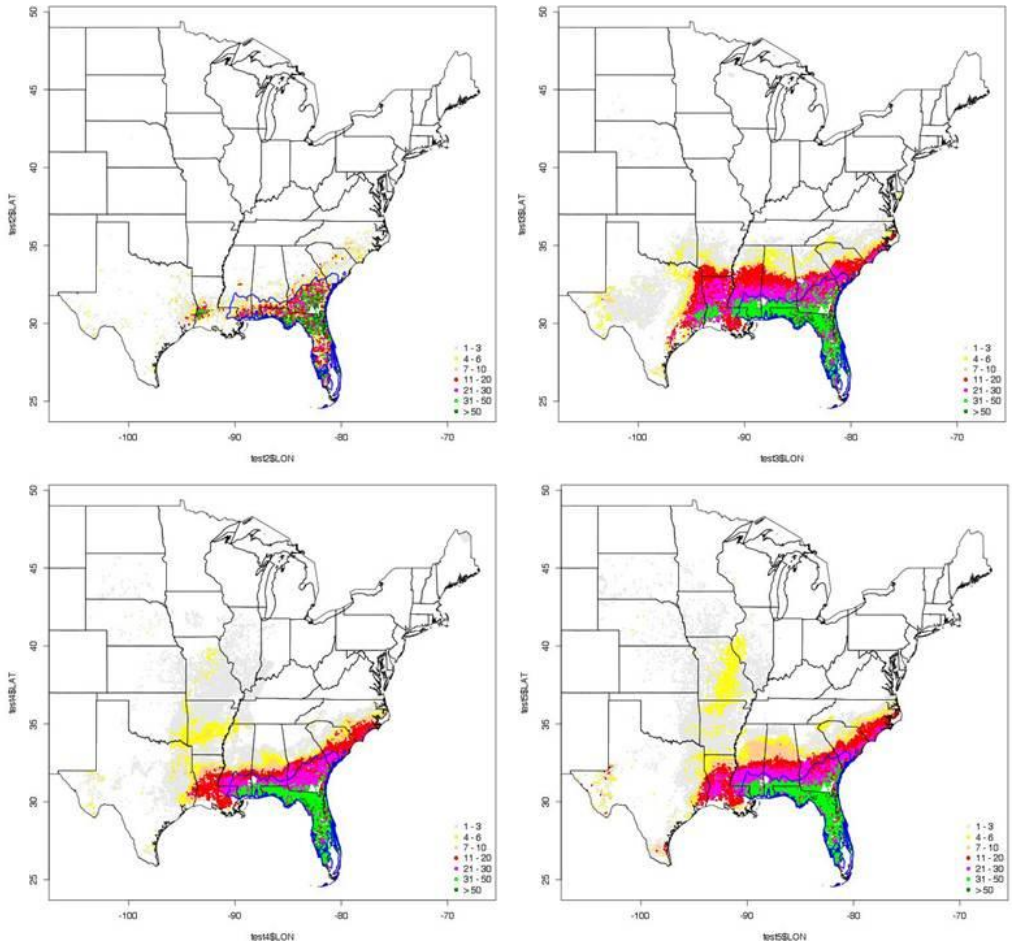


Figure A - 37. IV Slash pine GFDLCM21 B1 models (top-left), 2030 (top-right), 2060 (bottom-left), 2090 (bottom-right)

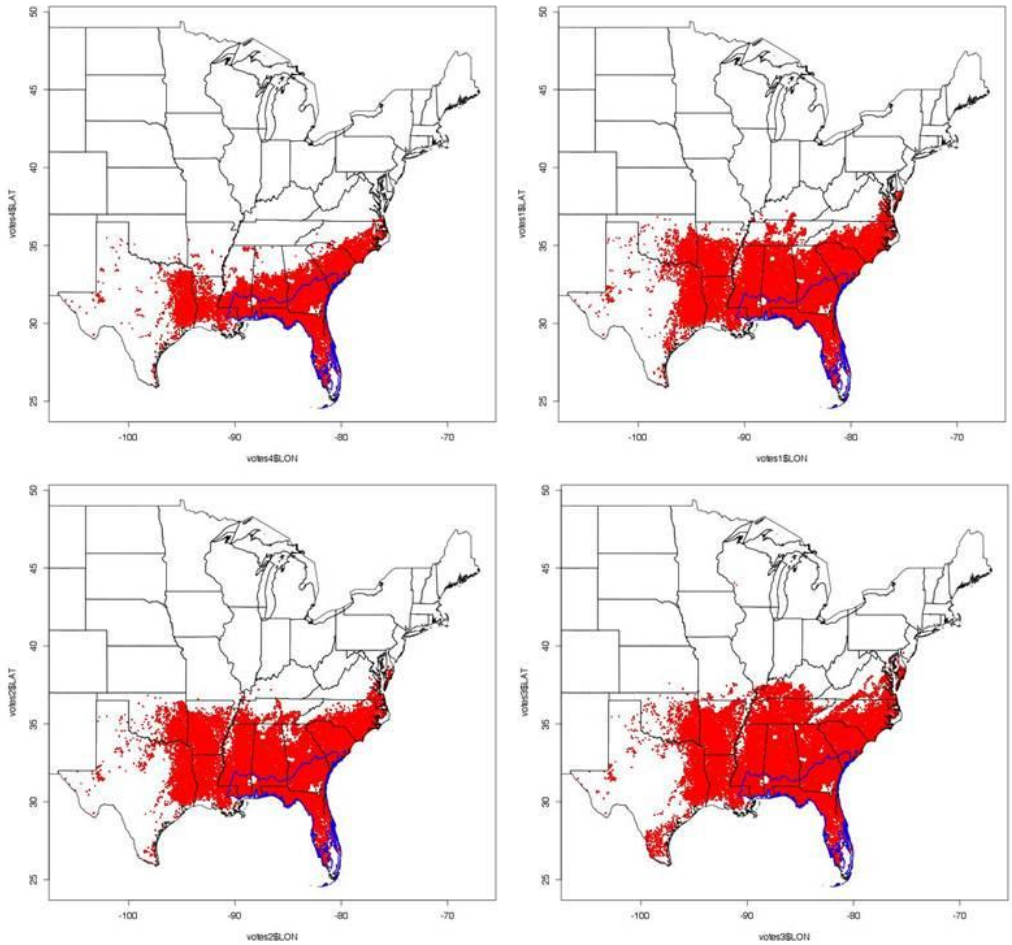


Figure A - 38. PA Slash pine GFDLCM21 B1 models (top-left), 2030 (top-right), 2060 (bottom-left), 2090 (bottom-right)

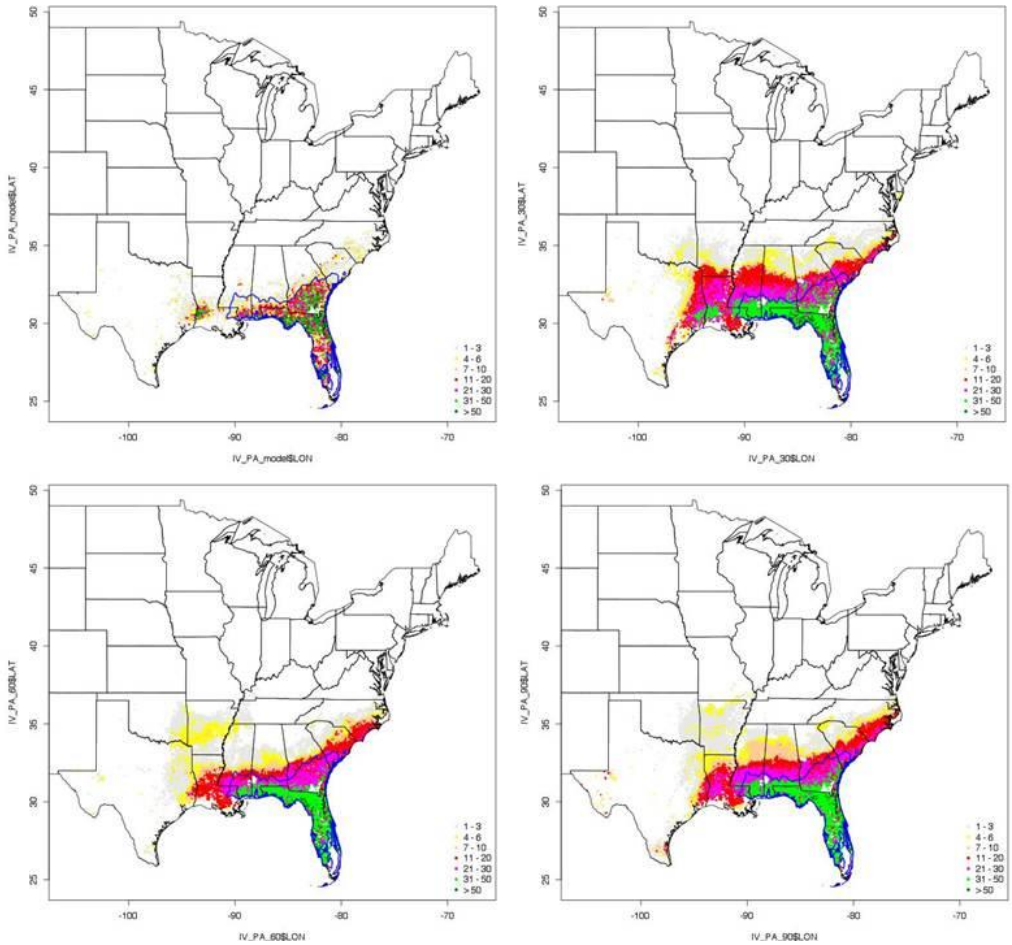


Figure A - 39. Combined IV and PA Slash pine GFDLCM21 B1 models (top-left), 2030 (top-right), 2060 (bottom-left), 2090 (bottom-right)

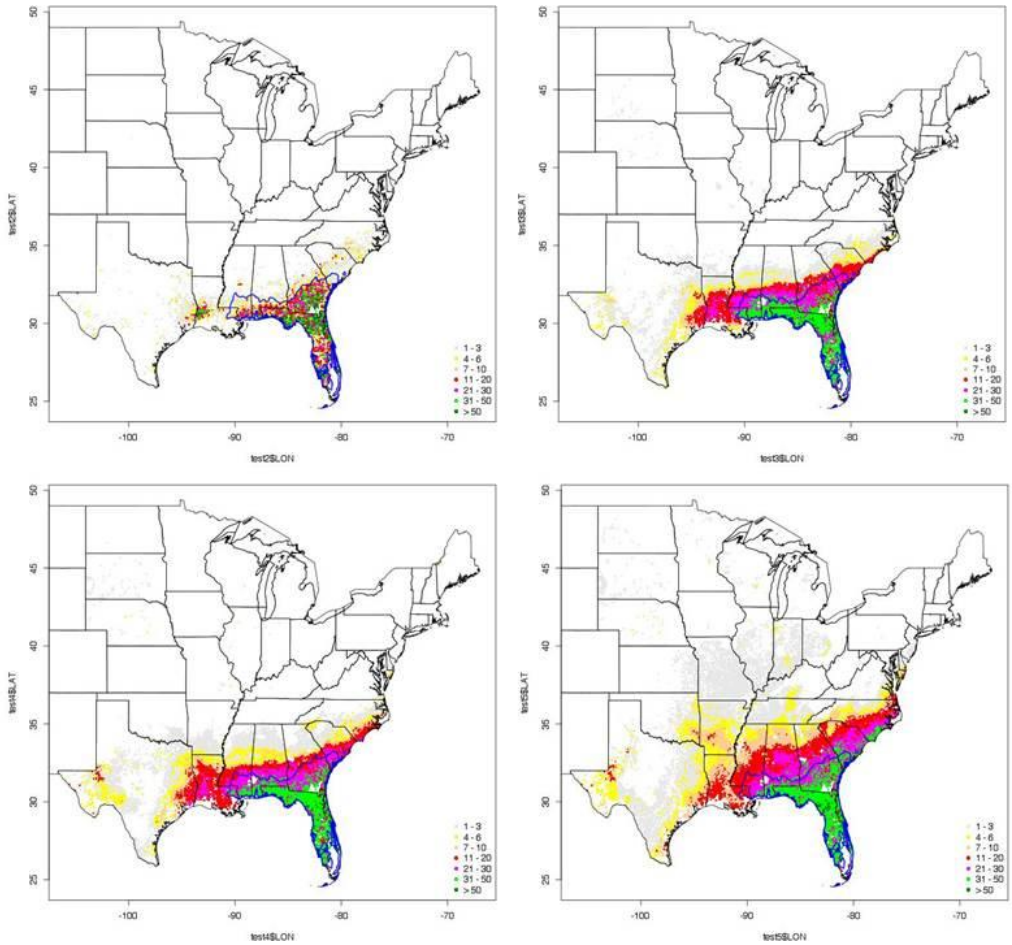


Figure A - 40. IV Slash pine HADCM3 A2 models (top-left), 2030 (top-right), 2060 (bottom-left), 2090 (bottom-right)

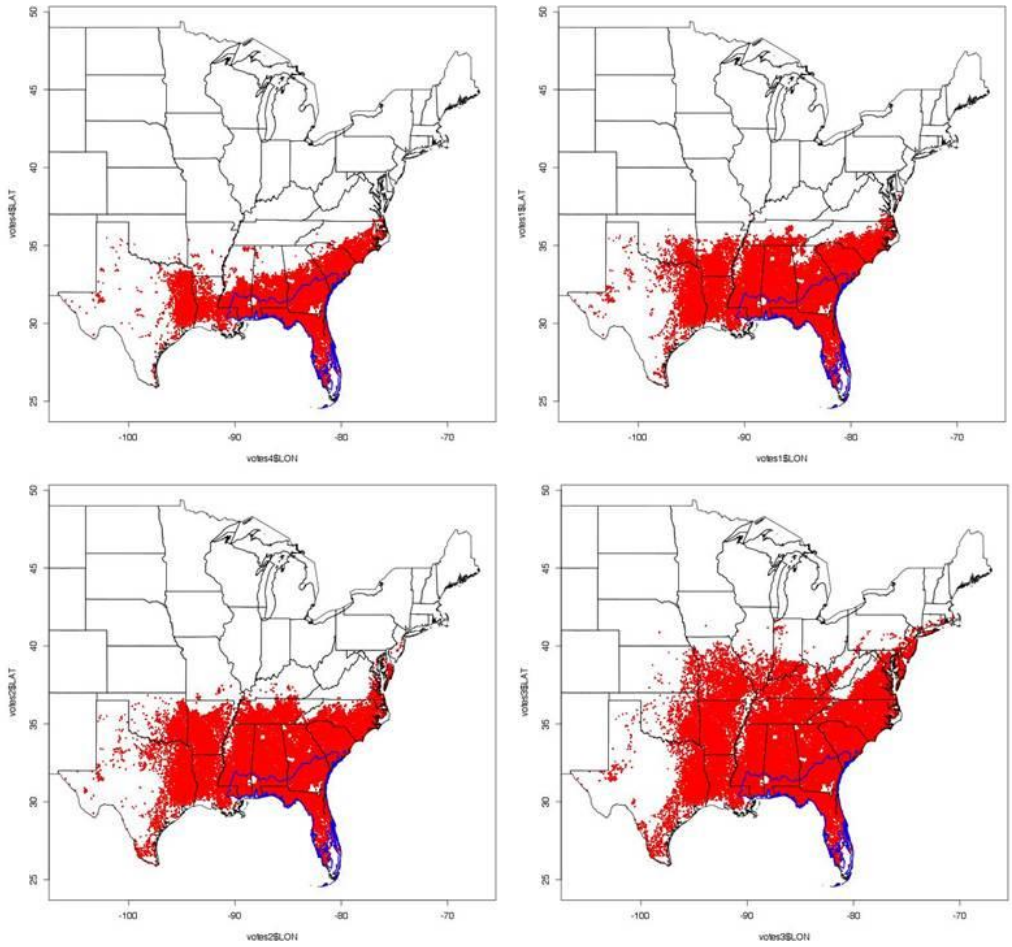


Figure A - 41. PA Slash pine HADCM3 A2 models (top-left), 2030 (top-right), 2060 (bottom-left), 2090 (bottom-right)

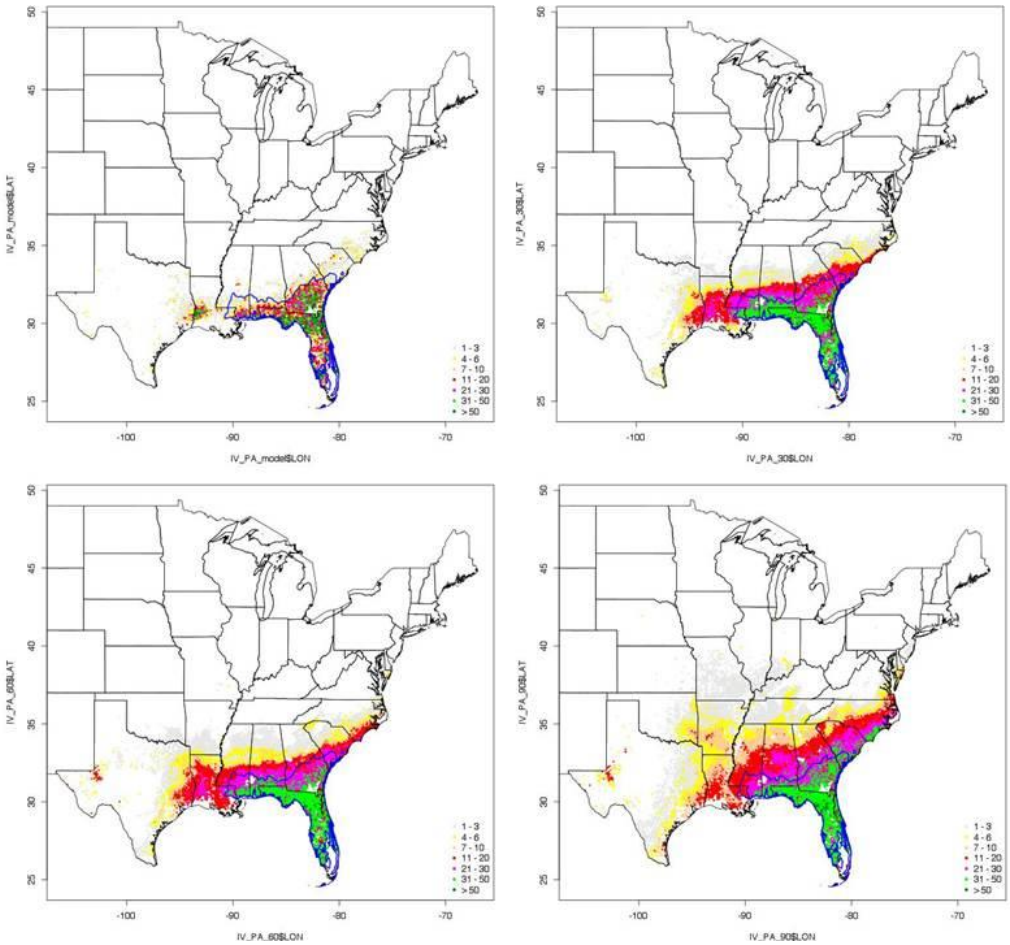


Figure A - 42. Combined IV and PA Slash pine HADCM3 A2 models (top-left), 2030 (top-right), 2060 (bottom-left), 2090 (bottom-right)

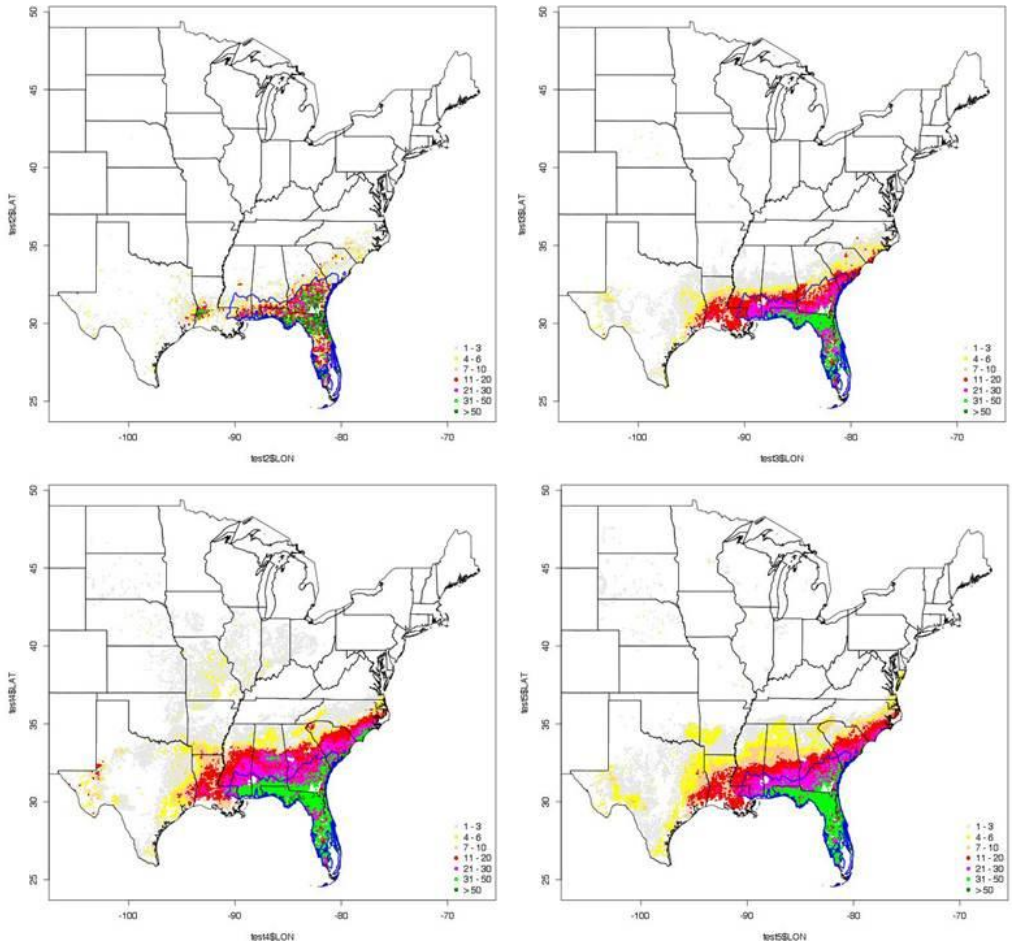


Figure A - 43. IV Slash pine HADCM3 B2 models (top-left), 2030 (top-right), 2060 (bottom-left), 2090 (bottom-right)

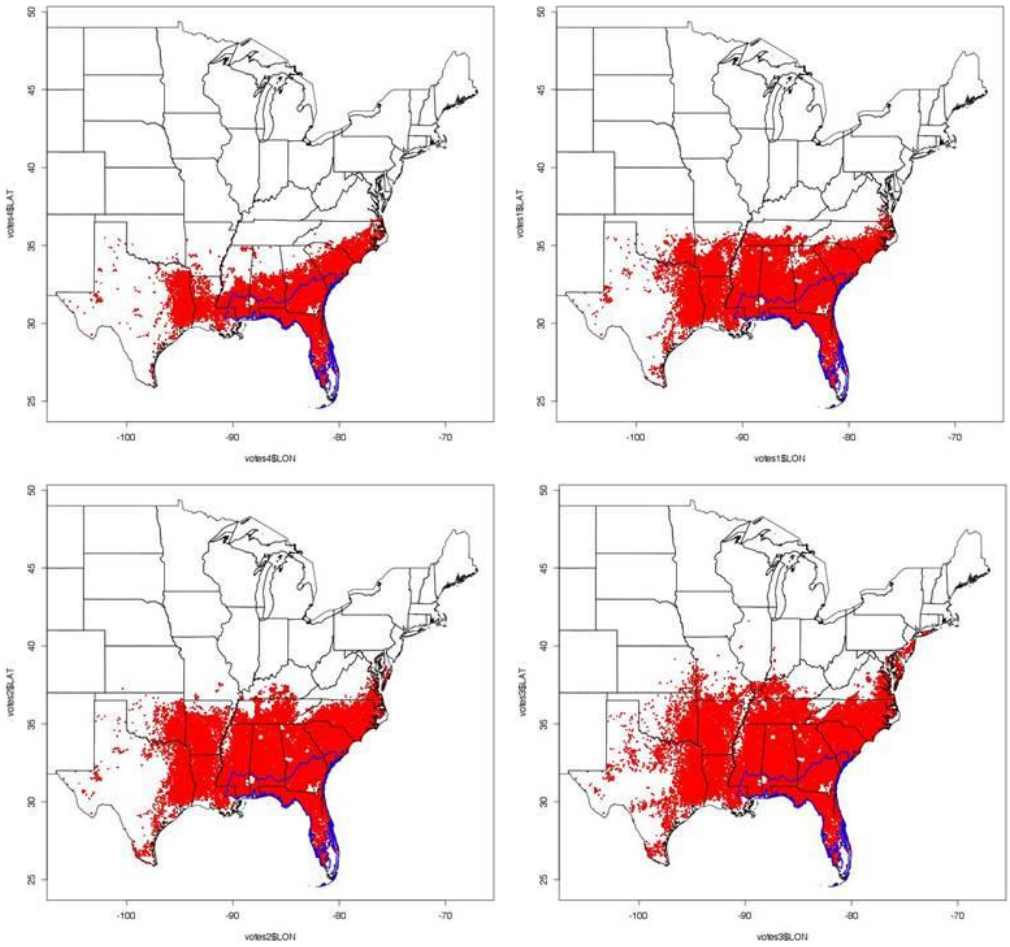


Figure A - 44. PA Slash pine HADCM3 B2 models (top-left), 2030 (top-right), 2060 (bottom-left), 2090 (bottom-right)

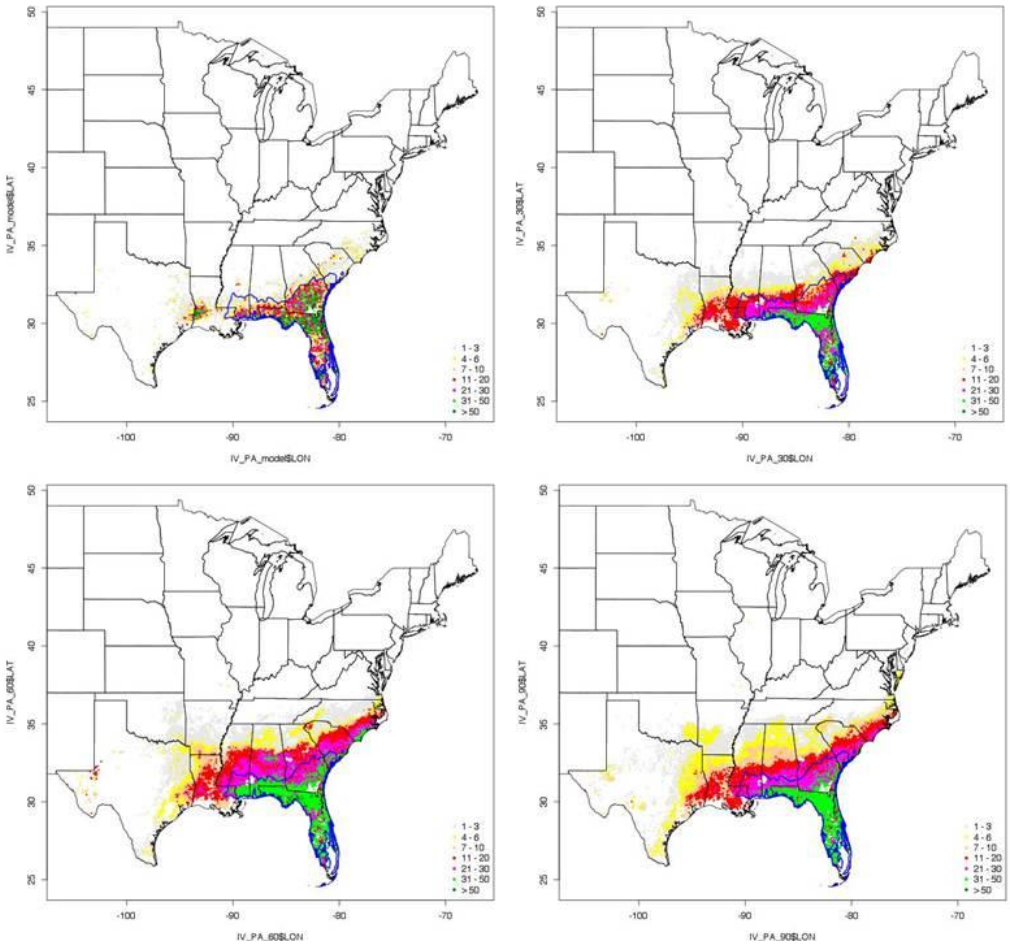


Figure A - 45. Combined IV and PA Slash pine HADCM3 B2 models (top-left), 2030 (top-right), 2060 (bottom-left), 2090 (bottom-right)

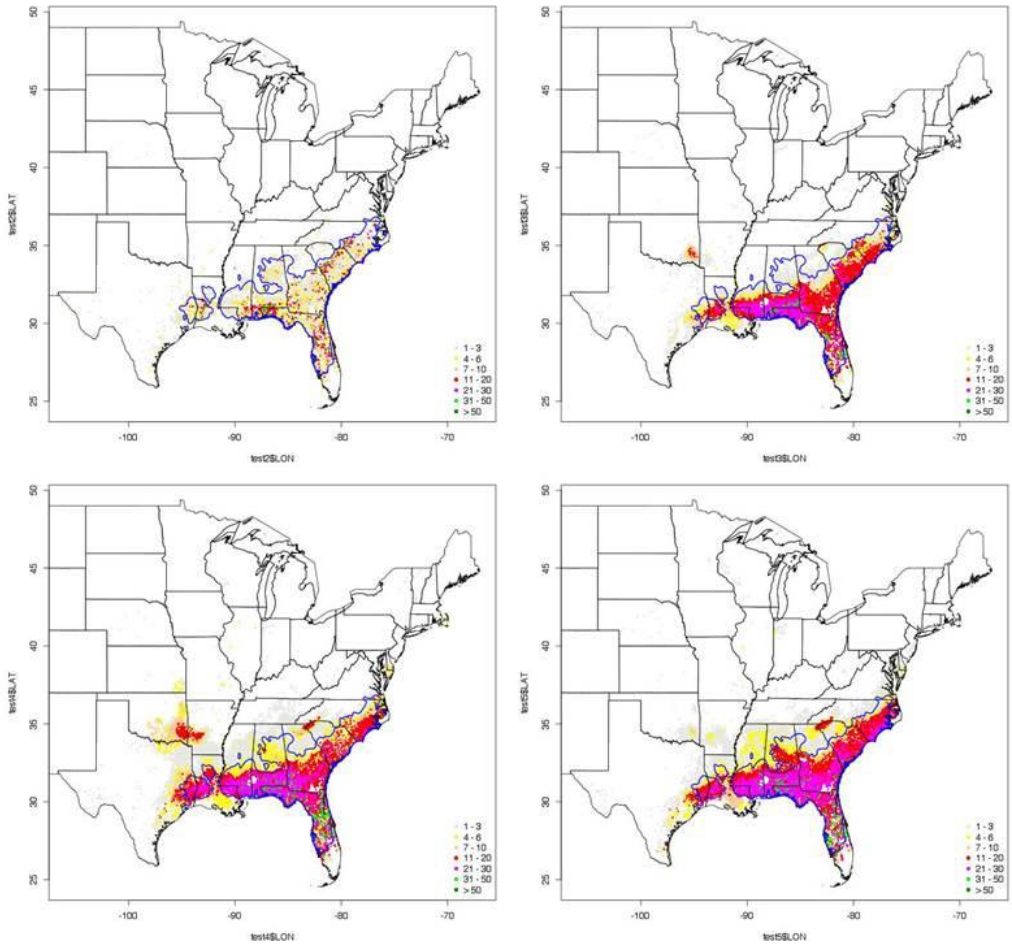


Figure A - 46. IV Longleaf pine CGCM3 A1B models (top-left), 2030 (top-right), 2060 (bottom-left), 2090 (bottom-right)

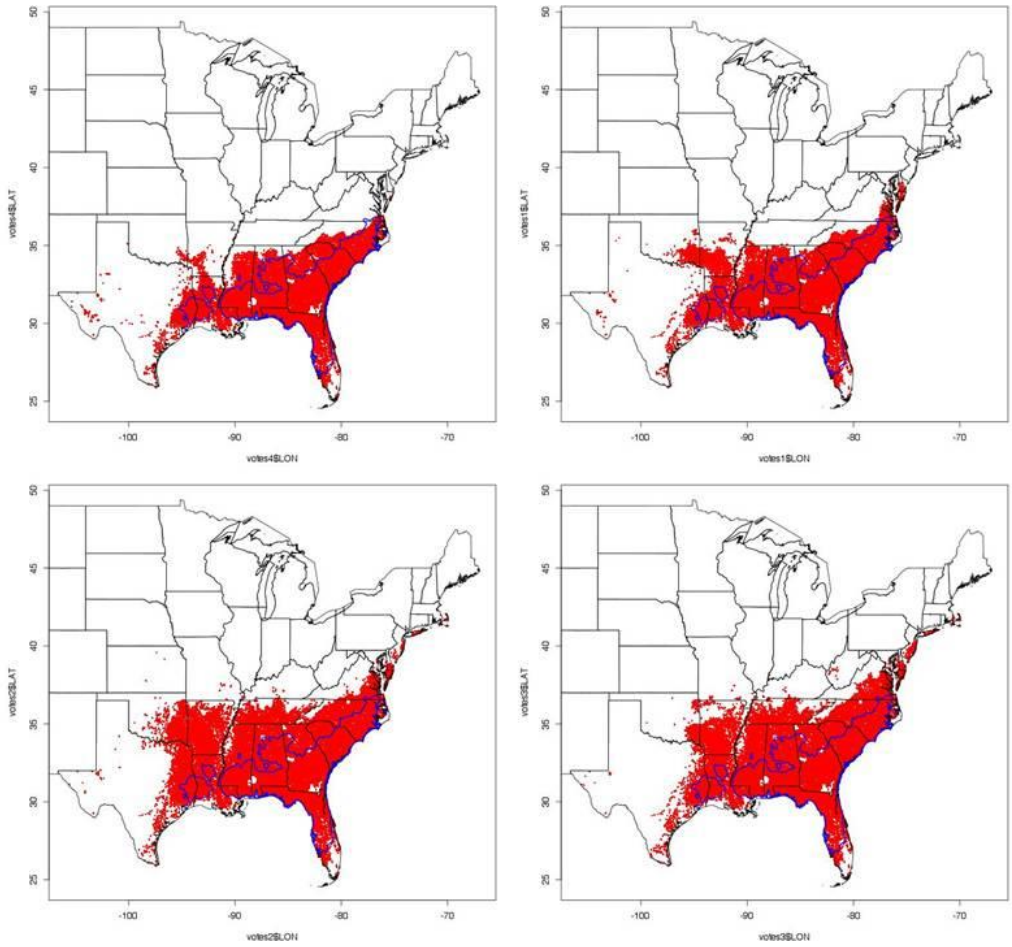


Figure A - 47. PA Longleaf pine CGCM3 A1B models (top-left), 2030 (top-right), 2060 (bottom-left), 2090 (bottom-right)

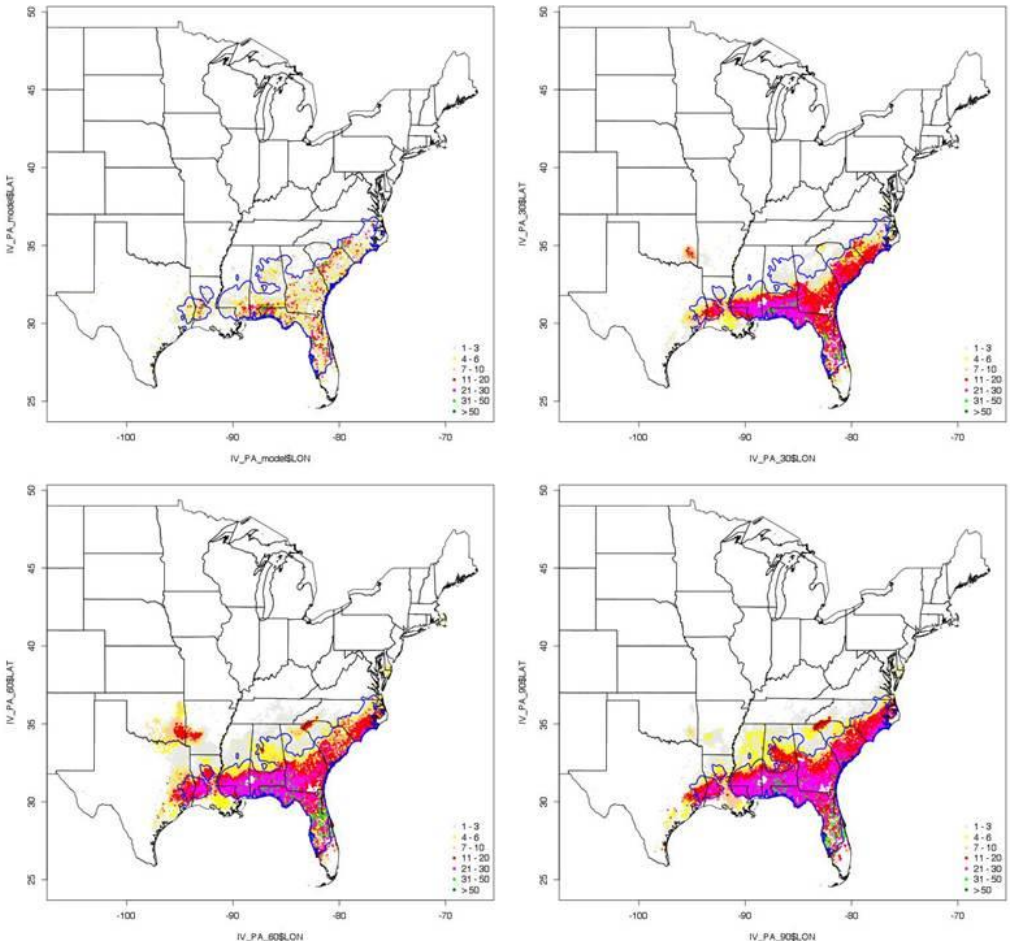


Figure A - 48. Combined IV and PA Longleaf pine CGCM3 A1B models (top-left), 2030 (top-right), 2060 (bottom-left), 2090 (bottom-right)

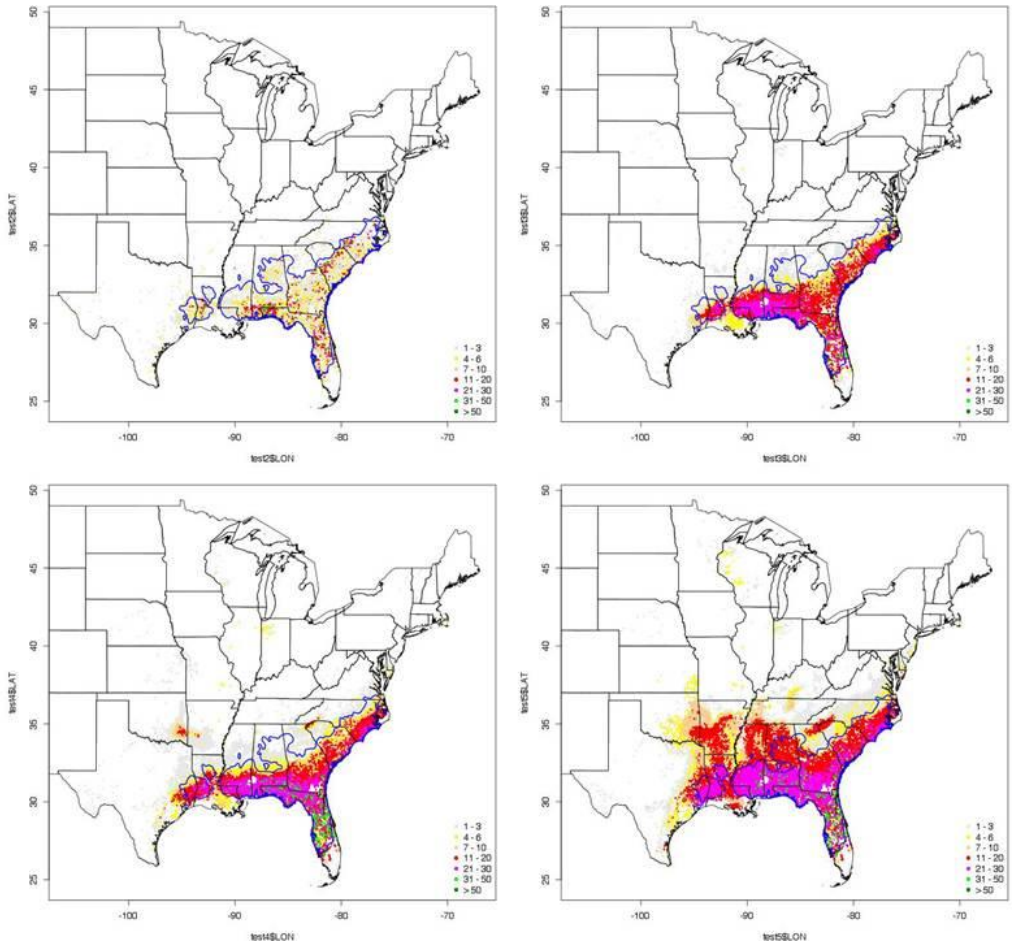


Figure A - 49. IV Longleaf pine CGCM3 A2 models (top-left), 2030 (top-right), 2060 (bottom-left), 2090 (bottom-right)

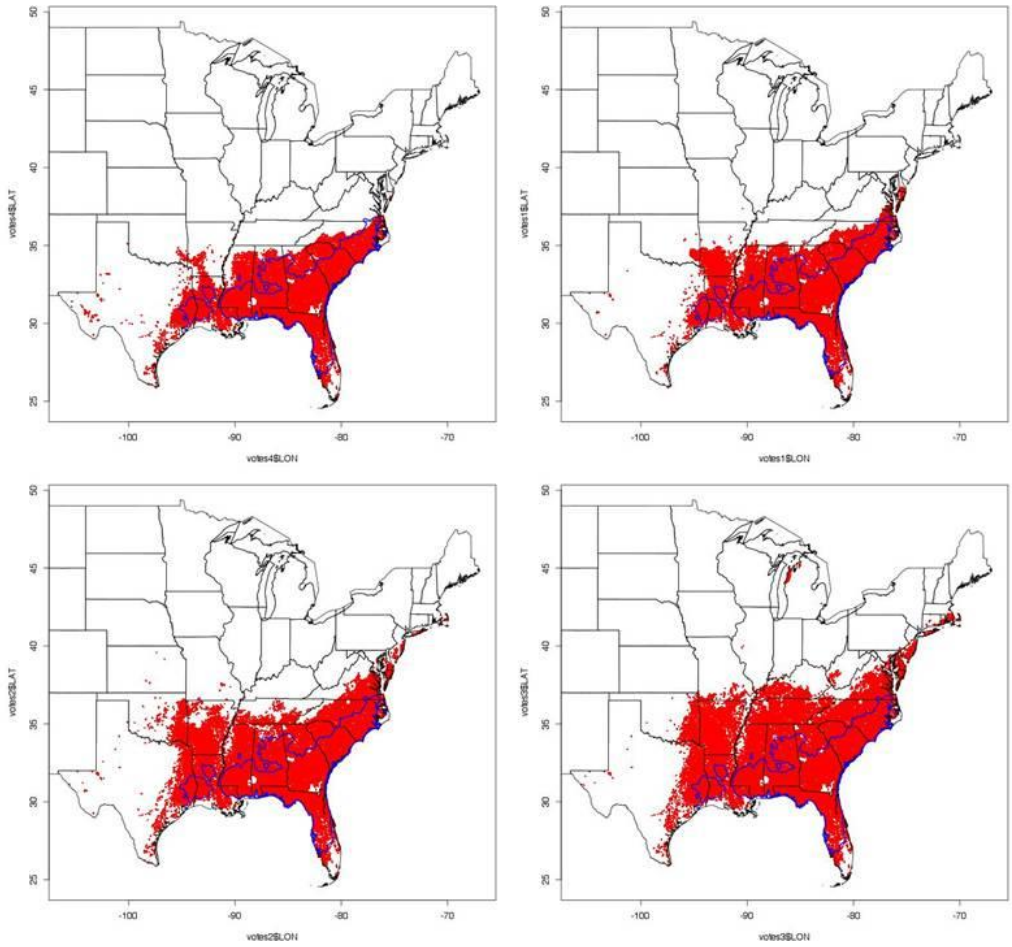


Figure A - 50. PA Longleaf pine CGCM3 A2 models (top-left), 2030 (top-right), 2060 (bottom-left), 2090 (bottom-right)

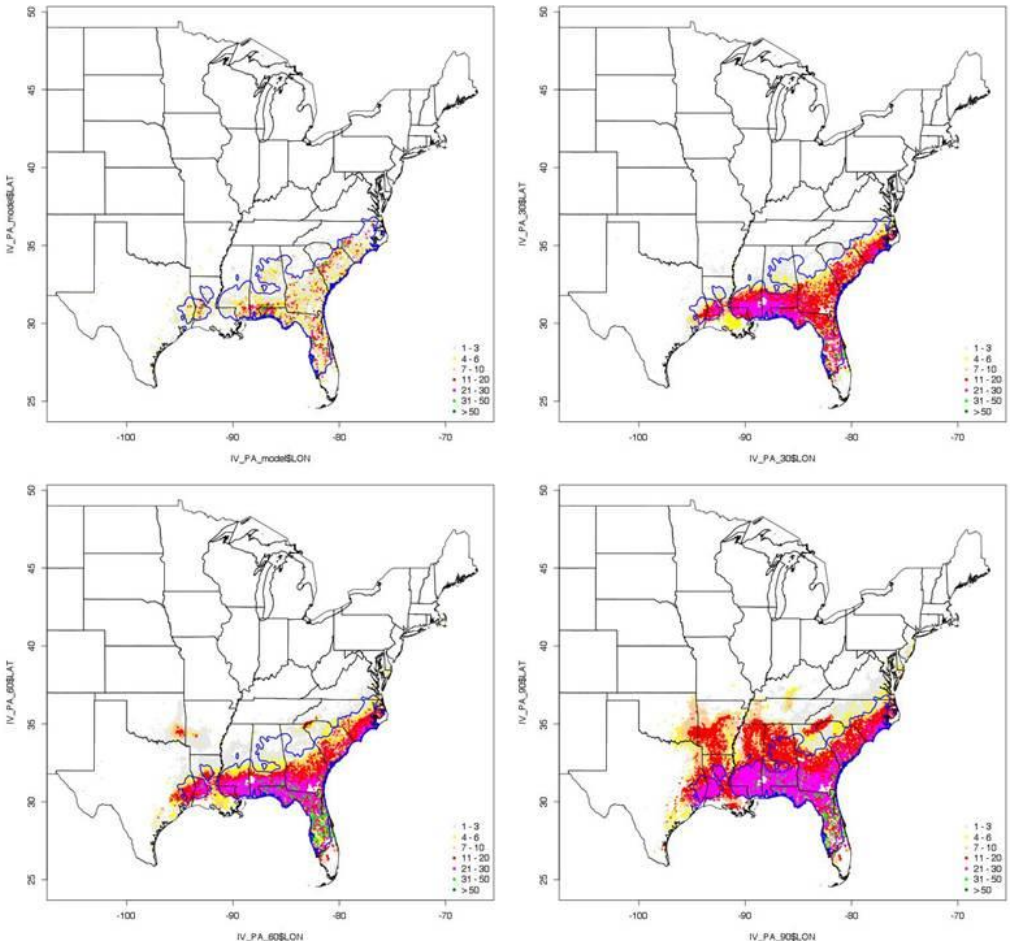


Figure A - 51. Combined IV and PA Longleaf pine CGCM3 A2 models (top-left), 2030 (top-right), 2060 (bottom-left), 2090 (bottom-right). Also referenced as Figure 17 in Results section.

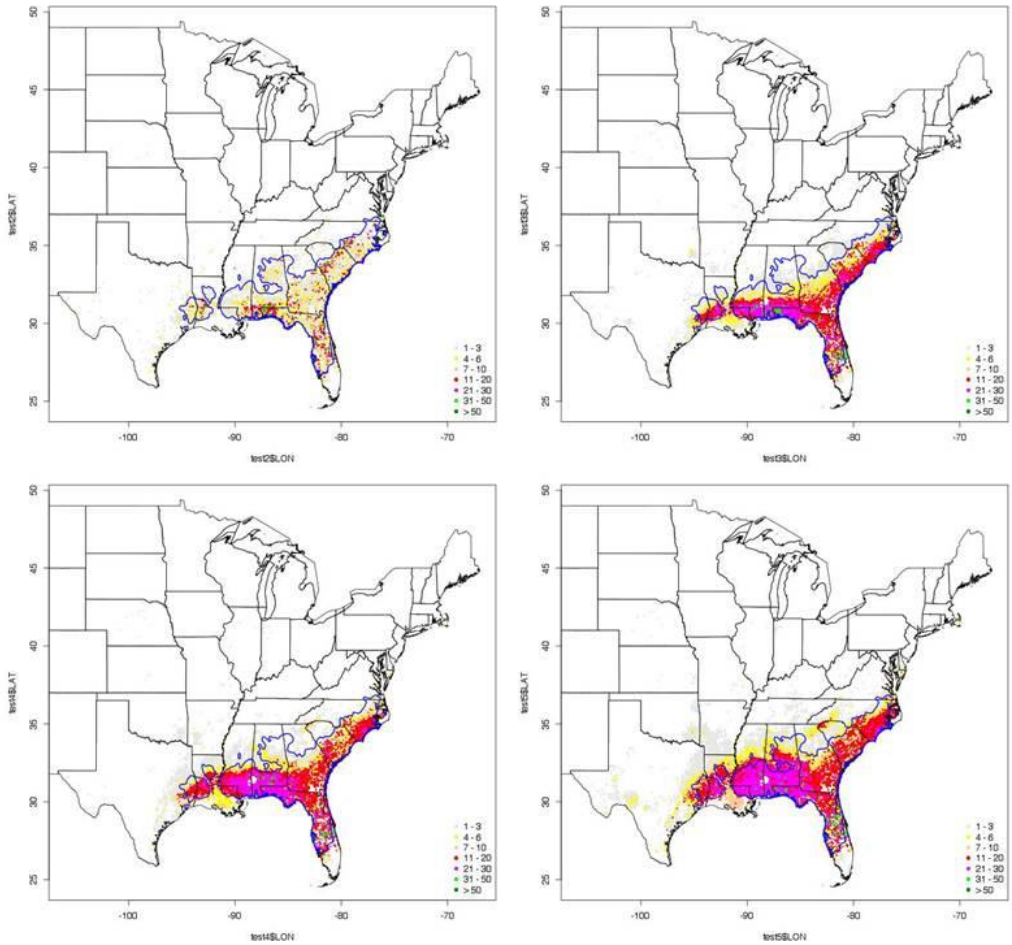


Figure A - 52. IV Longleaf pine CGCM3 B1 models (top-left), 2030 (top-right), 2060 (bottom-left), 2090 (bottom-right)

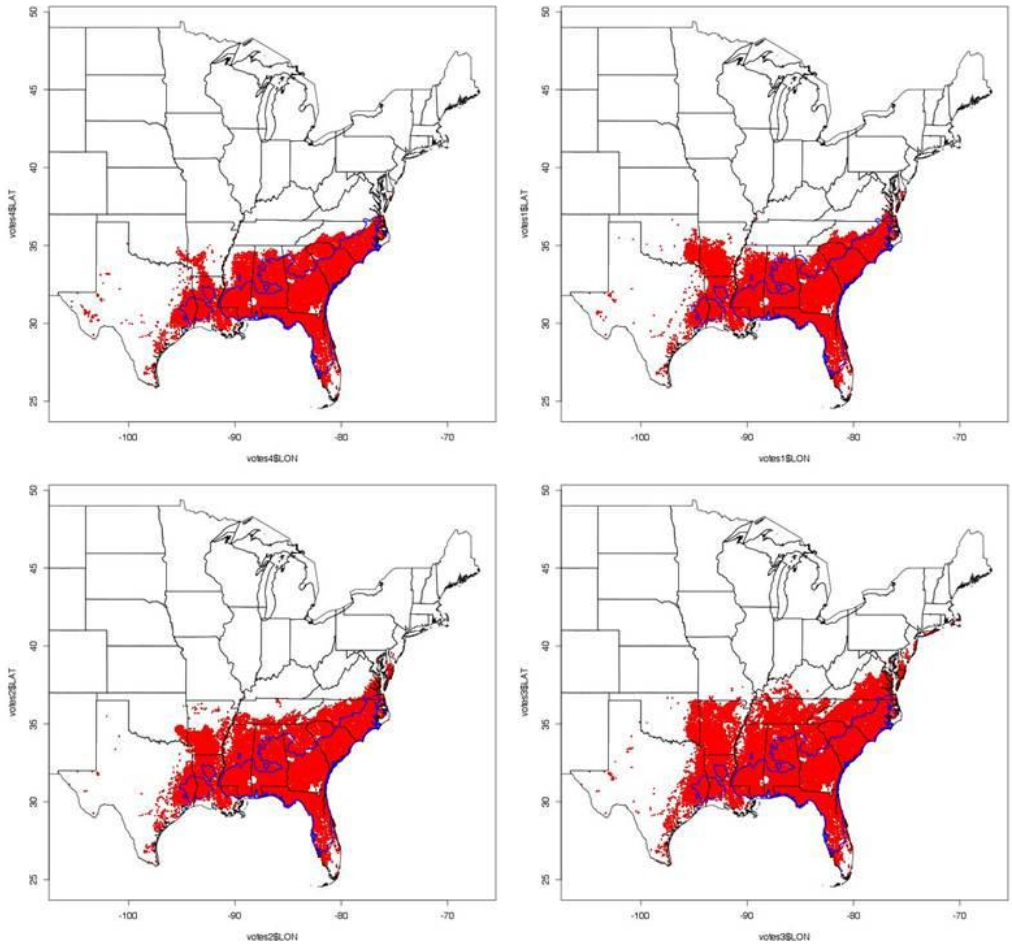


Figure A - 53. PA Longleaf pine CGCM3 B1 models (top-left), 2030 (top-right), 2060 (bottom-left), 2090 (bottom-right)

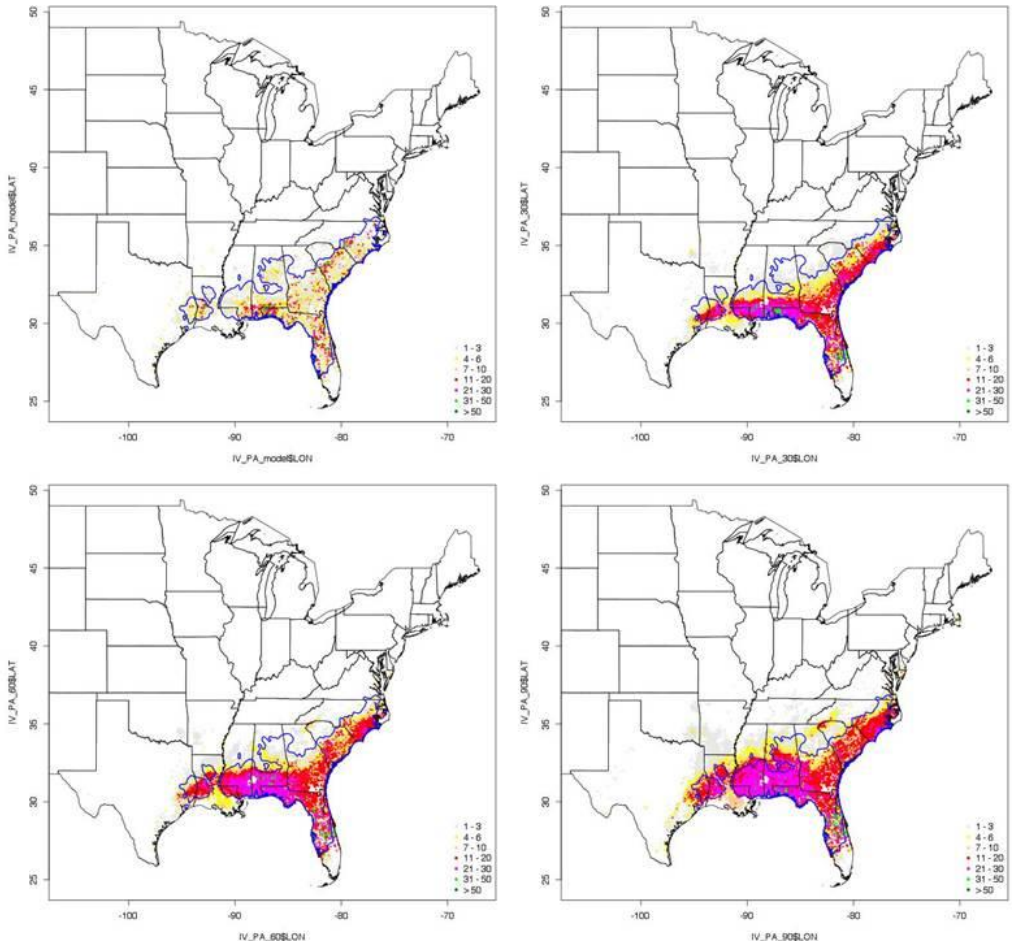


Figure A - 54. Combined IV and PA Longleaf pine CGCM3 B1 models (top-left), 2030 (top-right), 2060 (bottom-left), 2090 (bottom-right)

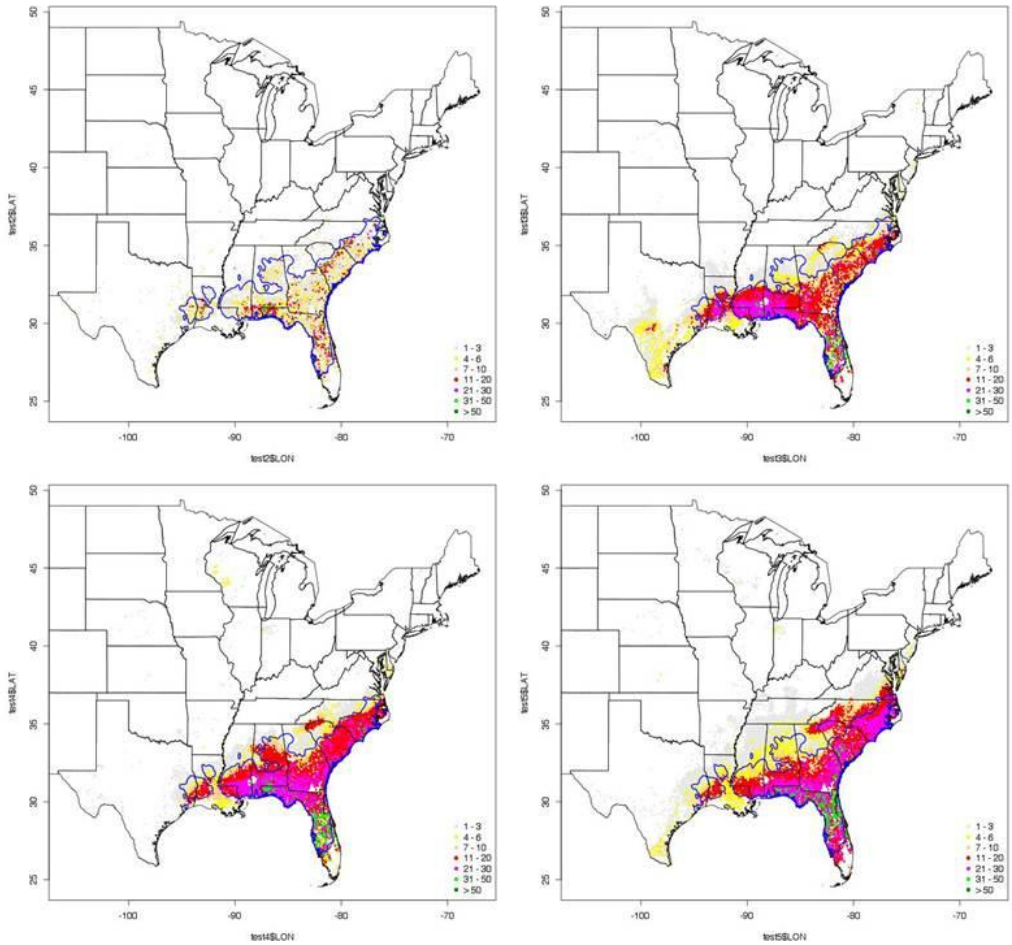


Figure A - 55. IV Longleaf pine GFDLCM21 A2 models (top-left), 2030 (top-right), 2060 (bottom-left), 2090 (bottom-right)

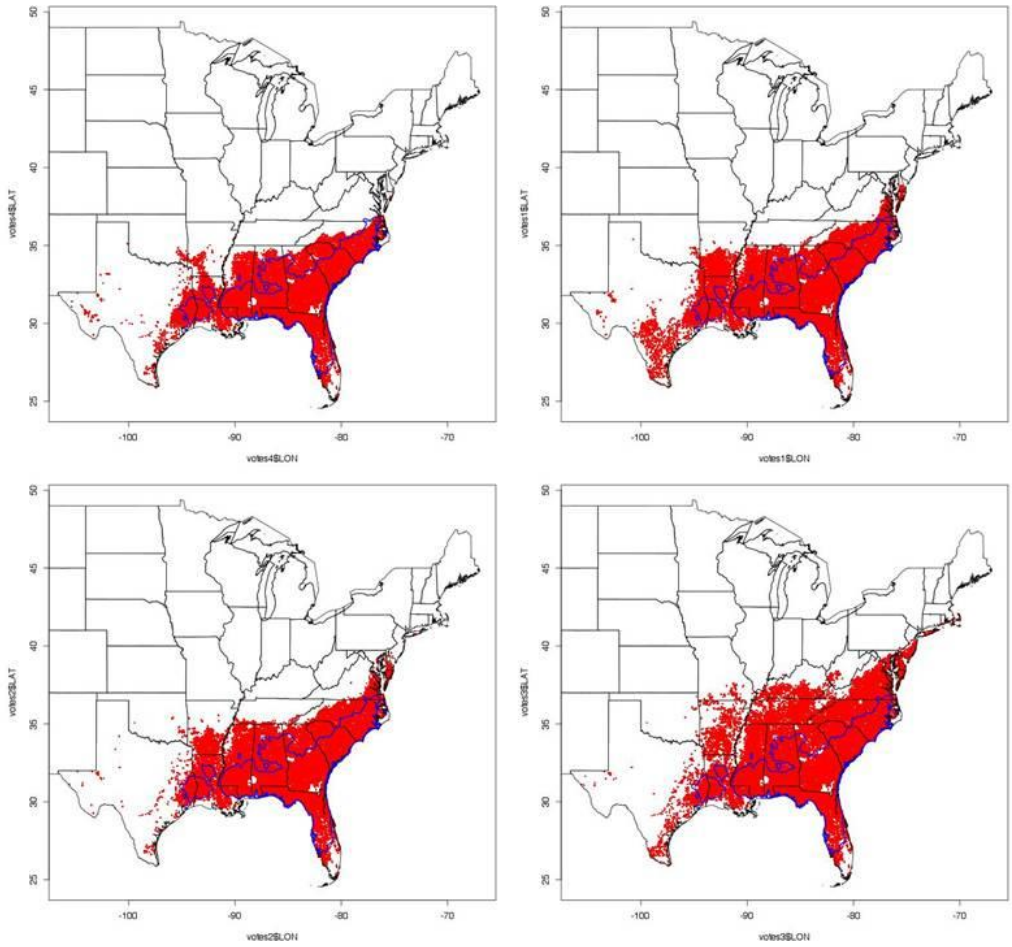


Figure A - 56. PA Longleaf pine GFDLCM321 A2 models (top-left), 2030 (top-right), 2060 (bottom-left), 2090 (bottom-right)

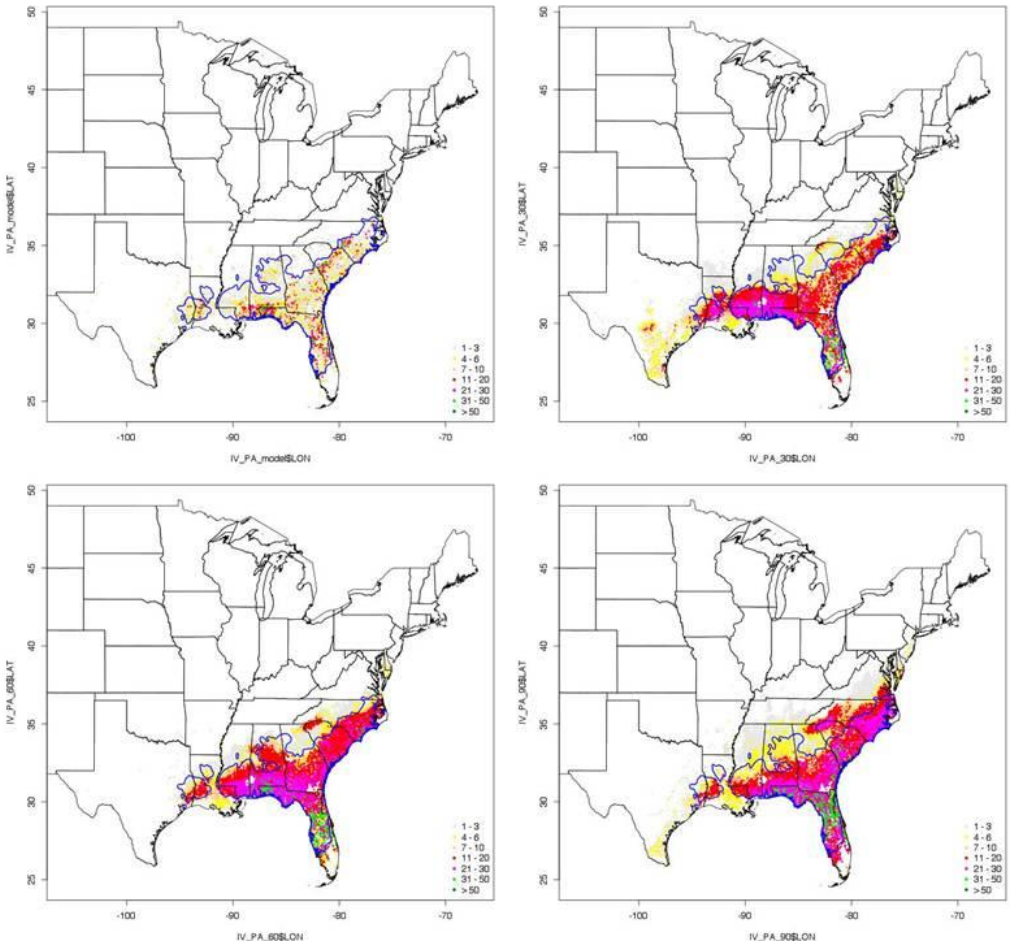


Figure A - 57. Combined IV and PA Longleaf pine GFDLCM21 A2 models (top-left), 2030 (top-right), 2060 (bottom-left), 2090 (bottom-right)

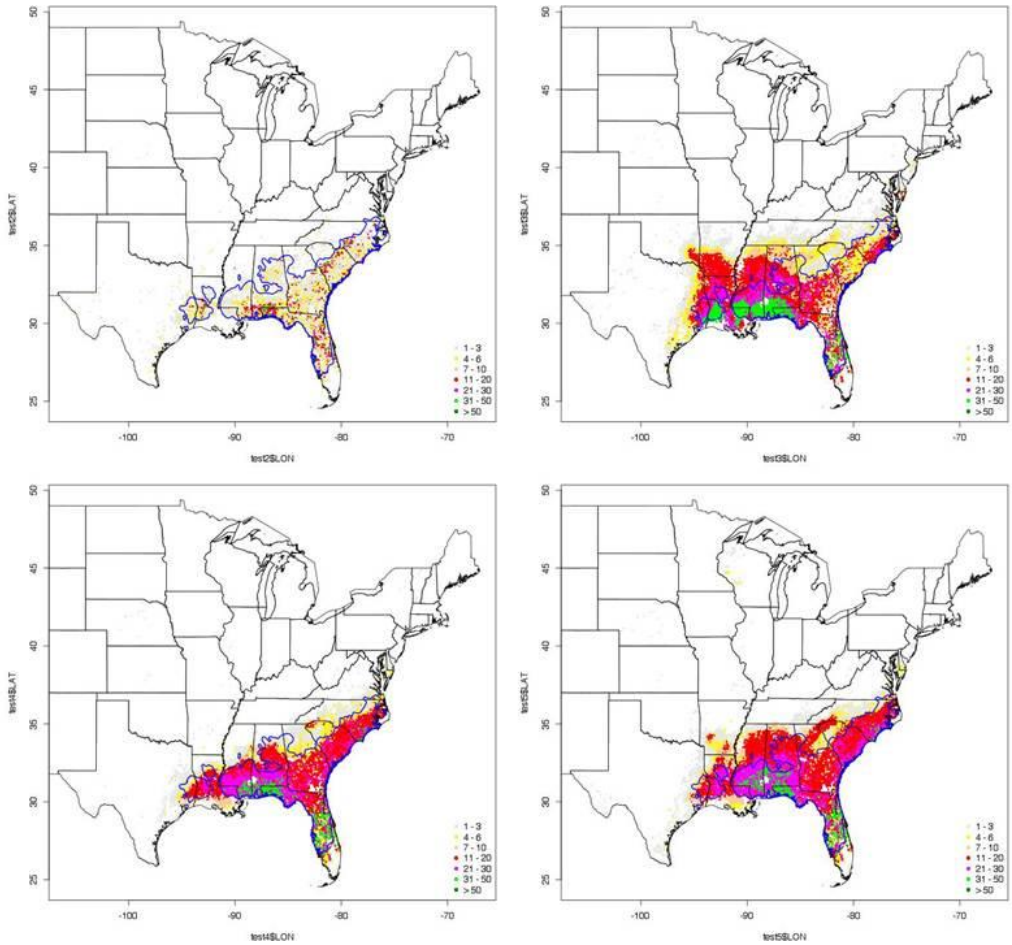


Figure A - 58. IV Longleaf pine GFDLCM21 B1 models (top-left), 2030 (top-right), 2060 (bottom-left), 2090 (bottom-right)

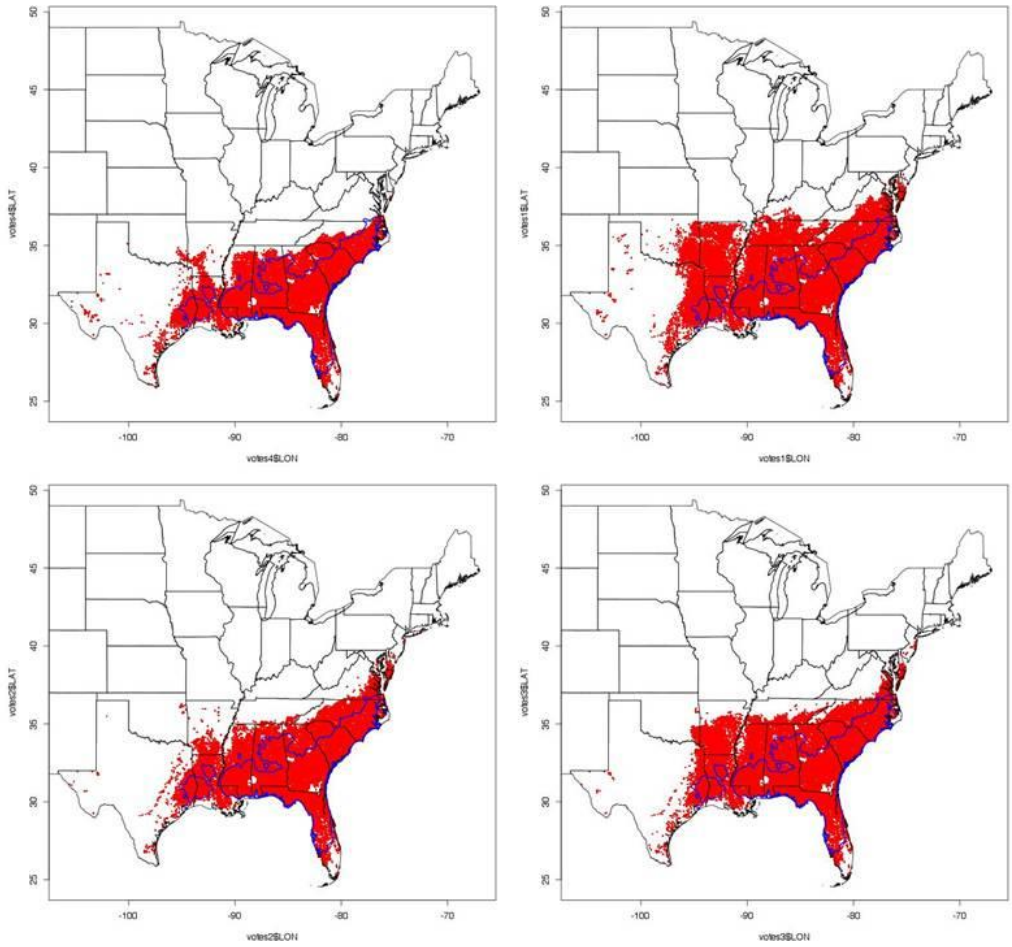


Figure A - 59. PA Longleaf pine GFDLCM21 B1 models (top-left), 2030 (top-right), 2060 (bottom-left), 2090 (bottom-right)

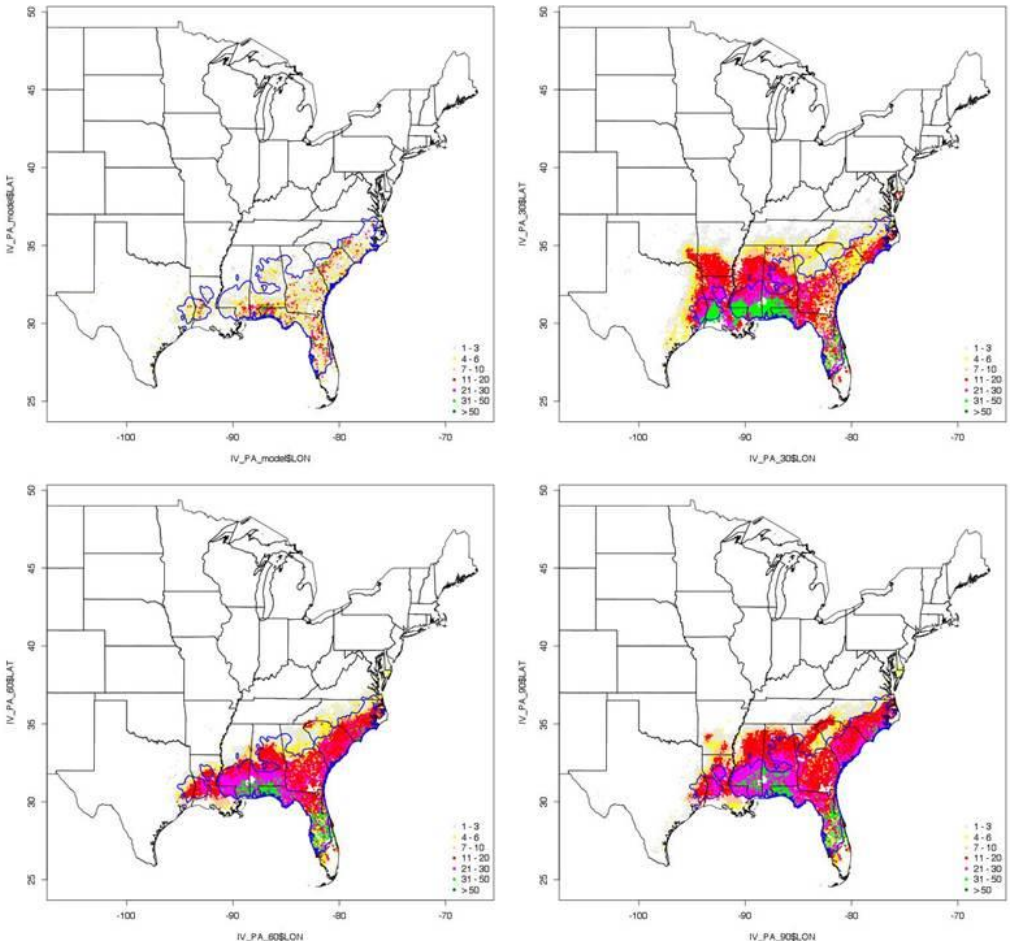


Figure A - 60. Combined IV and PA Longleaf pine GFDLCM21 B1 models (top-left), 2030 (top-right), 2060 (bottom-left), 2090 (bottom-right)

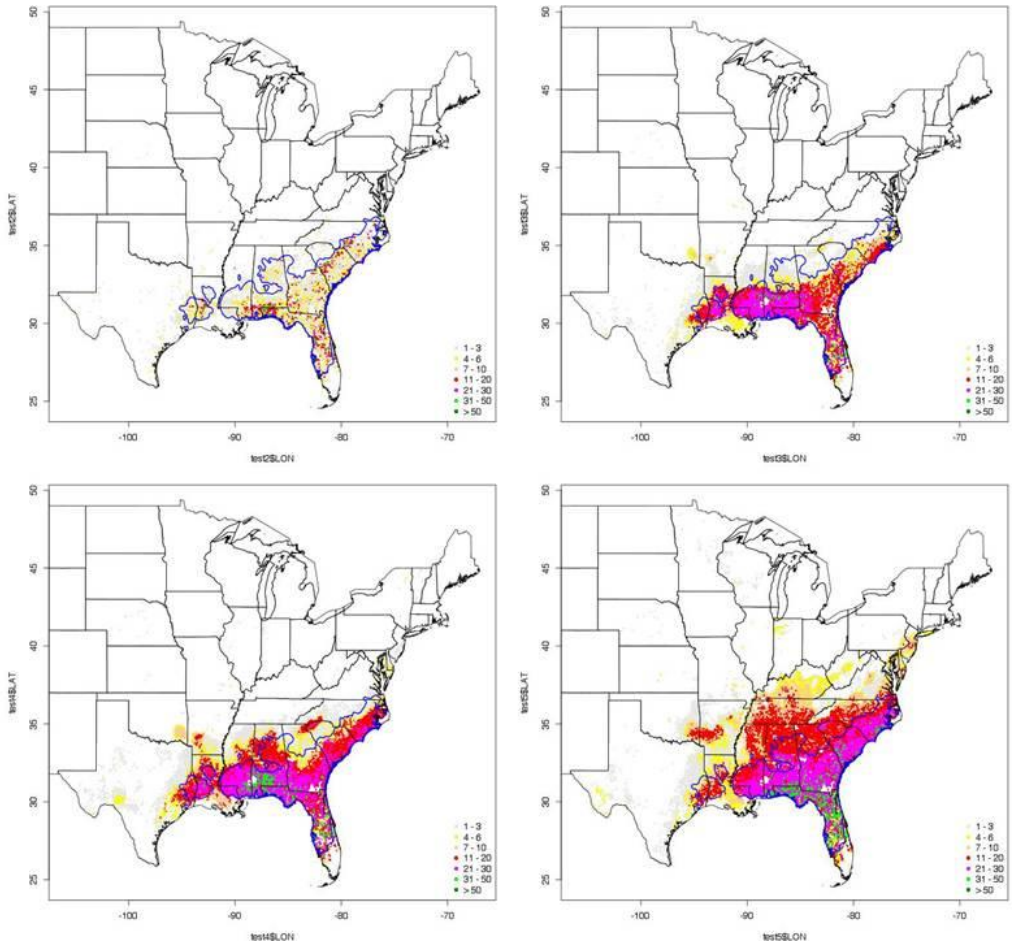


Figure A - 61. IV Longleaf pine HADCM3 A2 models (top-left), 2030 (top-right), 2060 (bottom-left), 2090 (bottom-right). Also referenced at Figure 20 in Results.

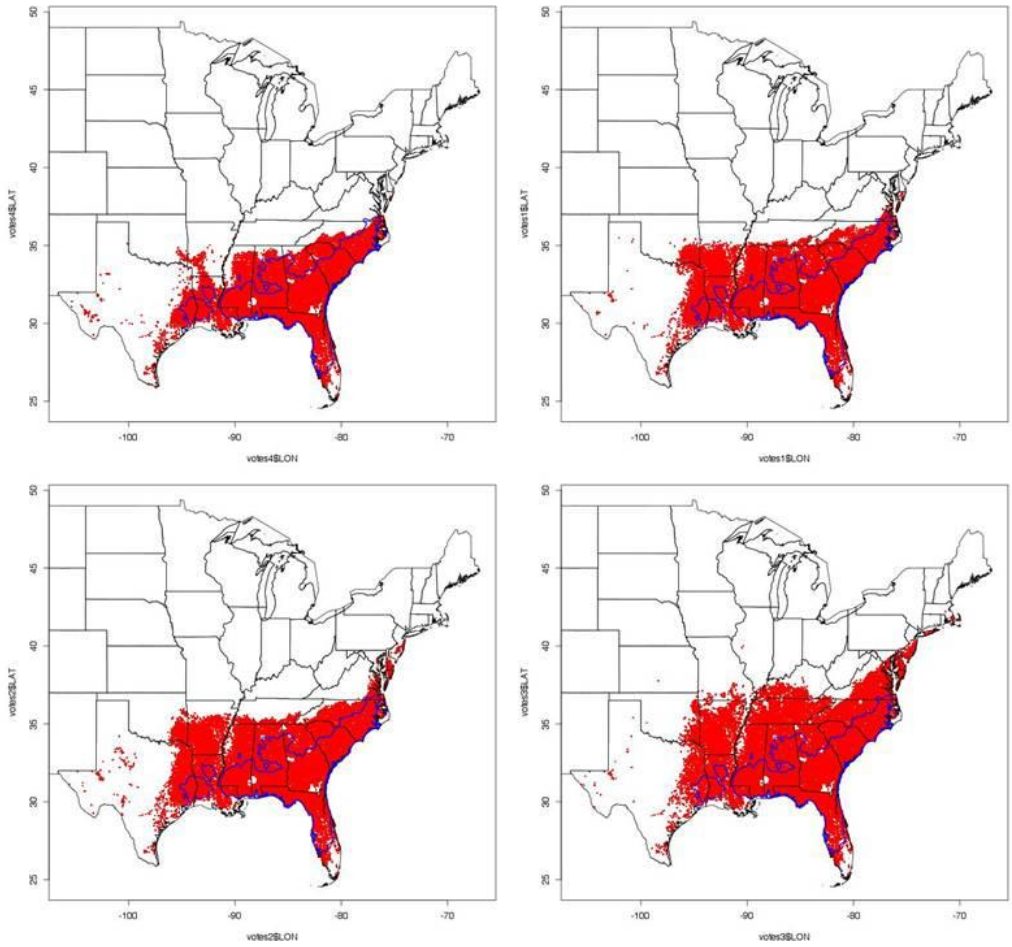


Figure A - 62. PA Longleaf pine HADCM3 A2 models (top-left), 2030 (top-right), 2060 (bottom-left), 2090 (bottom-right)

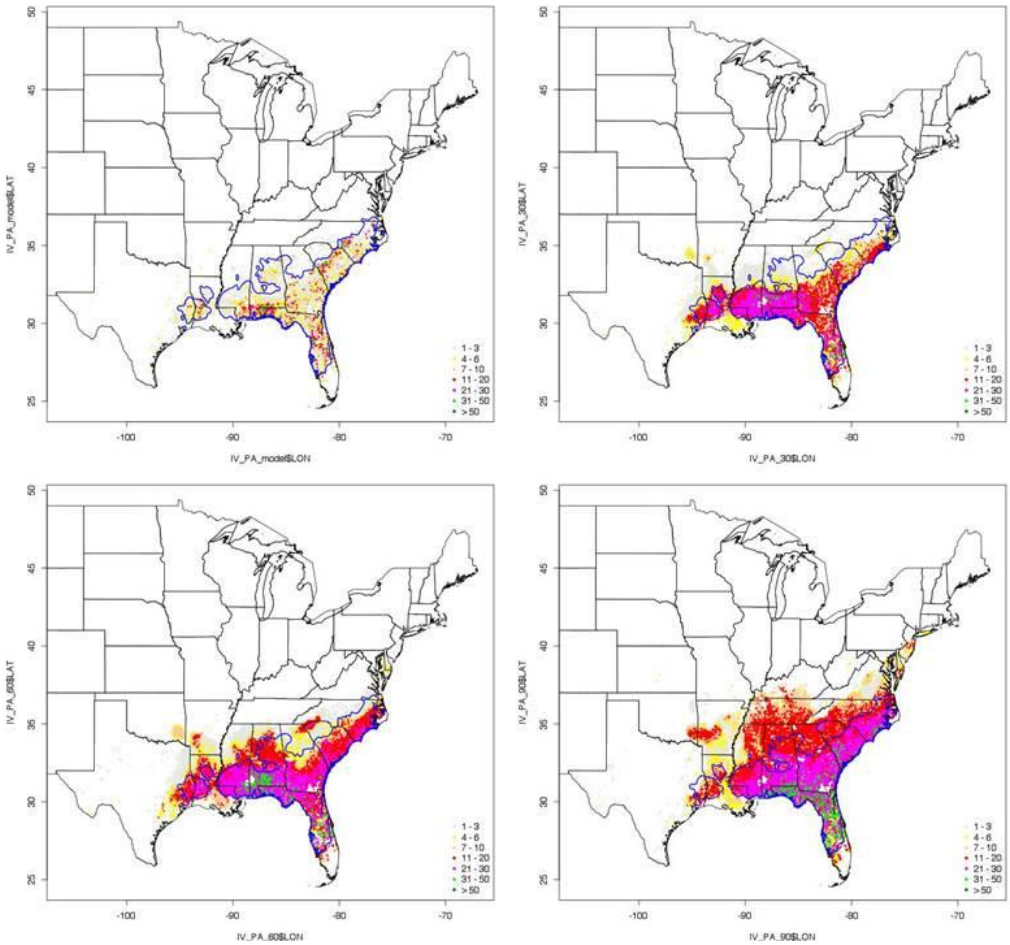


Figure A - 63. Combined IV and PA Longleaf pine HADCM3 A2 models (top-left), 2030 (top-right), 2060 (bottom-left), 2090 (bottom-right)

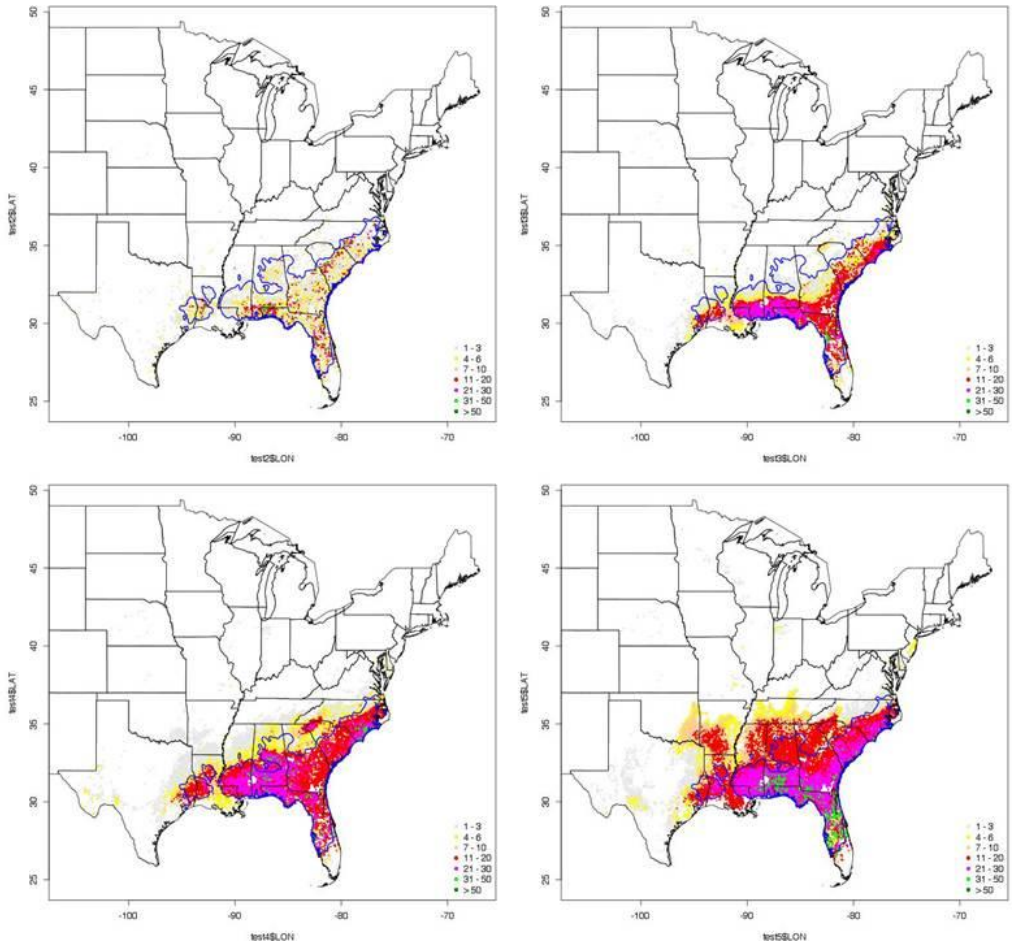


Figure A - 64. IV Longleaf pine HADCM3 B2 models (top-left), 2030 (top-right), 2060 (bottom-left), 2090 (bottom-right)

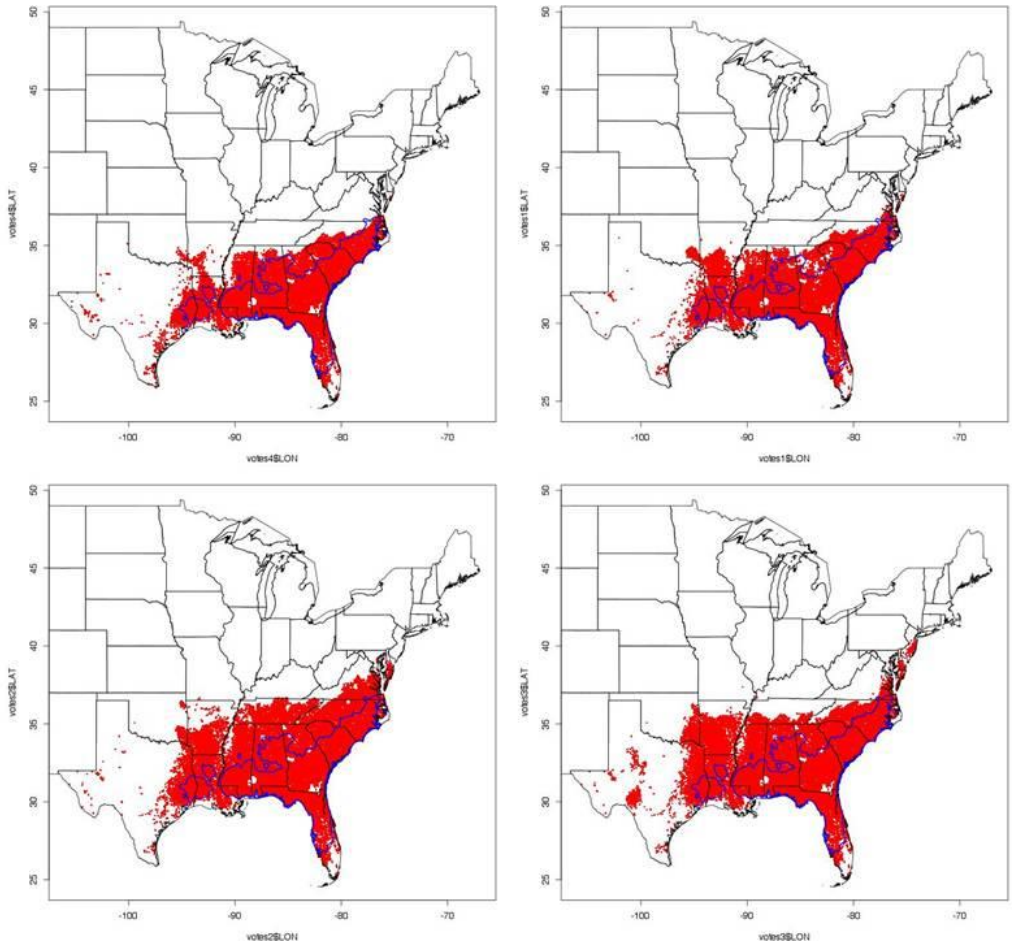


Figure A - 65. PA Longleaf pine HADCM3 B2 models (top-left), 2030 (top-right), 2060 (bottom-left), 2090 (bottom-right)

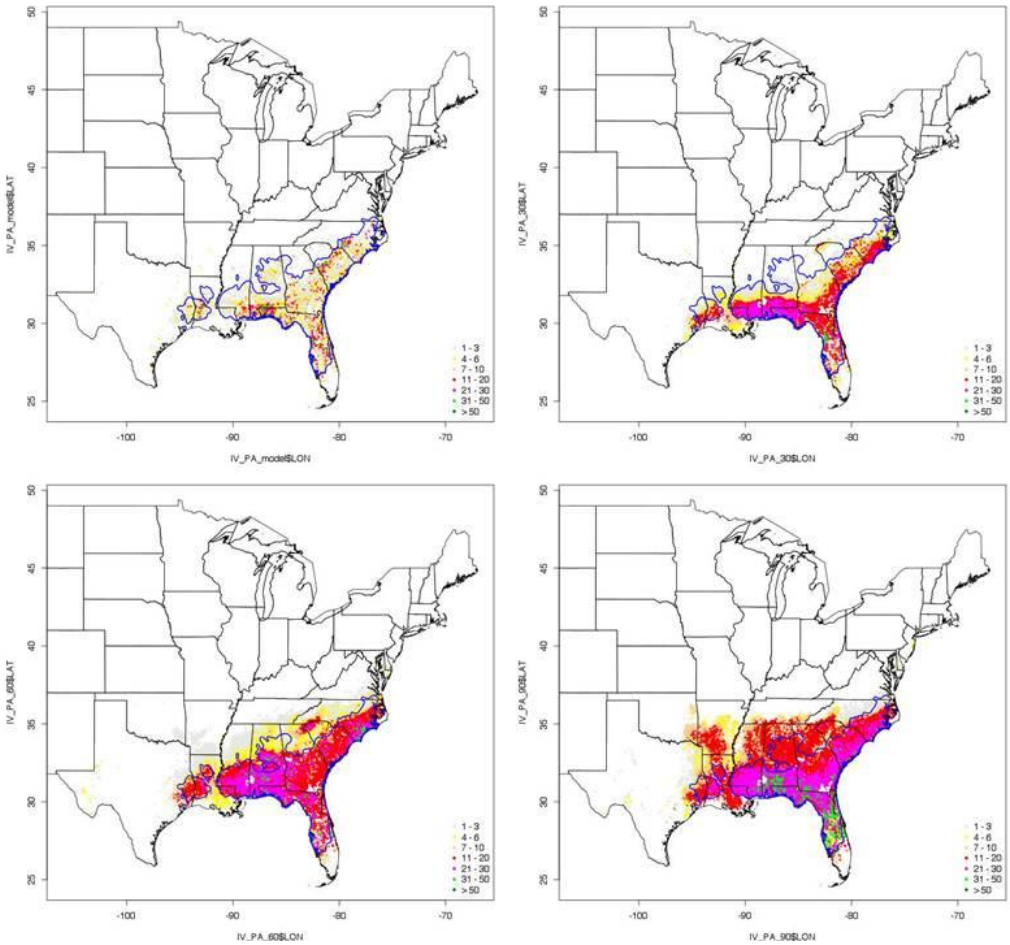


Figure A - 66. Combined IV and PA Longleaf pine HADCM3 B2 models (top-left), 2030 (top-right), 2060 (bottom-left), 2090 (bottom-right)

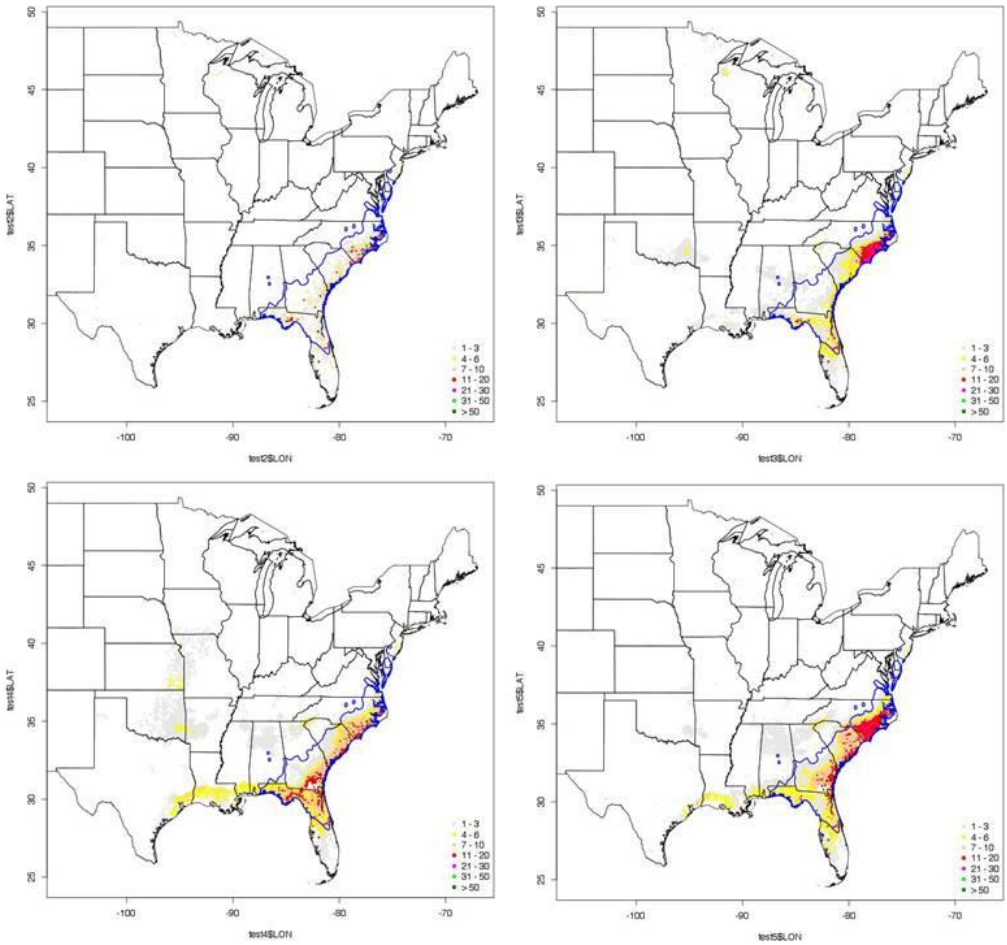


Figure A - 67. IV Pond pine CGCM3 A1B models (top-left), 2030 (top-right), 2060 (bottom-left), 2090 (bottom-right)

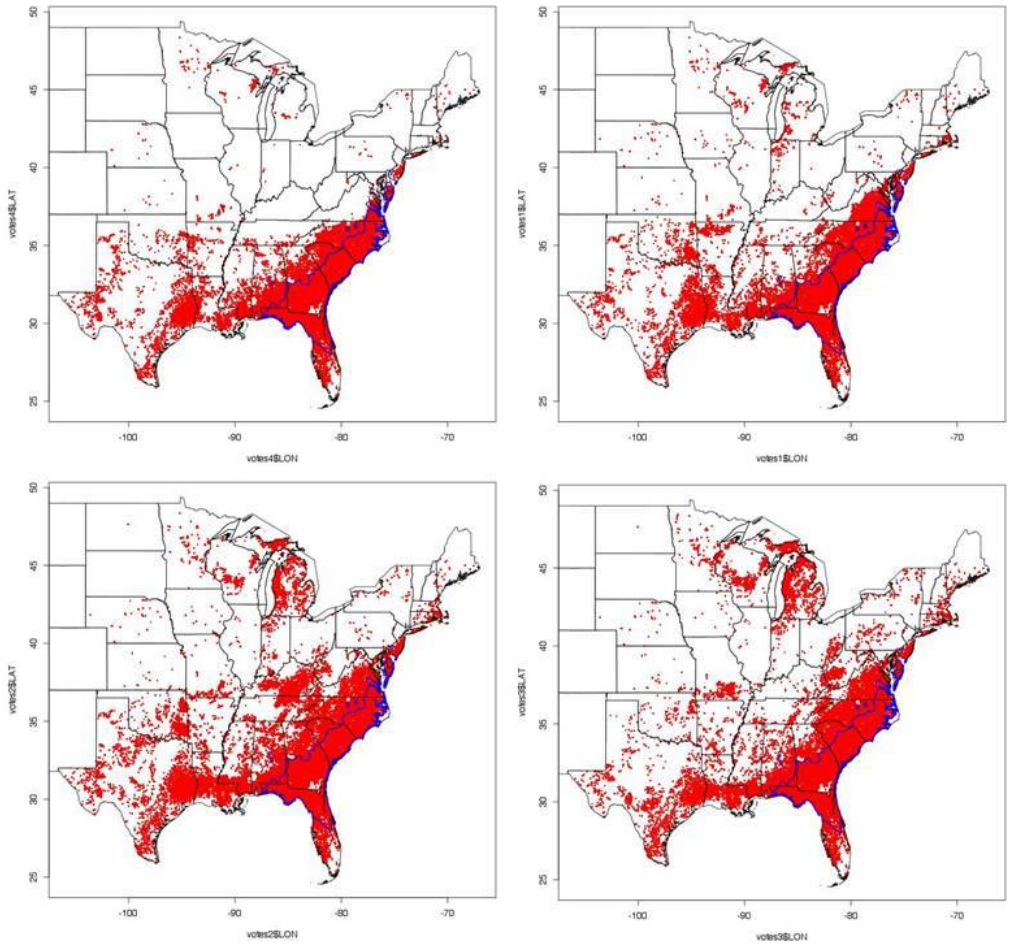


Figure A - 68. PA Pond pine CGCM3 A1B models (top-left), 2030 (top-right), 2060 (bottom-left), 2090 (bottom-right). Also referenced as Figure 37 in Results.

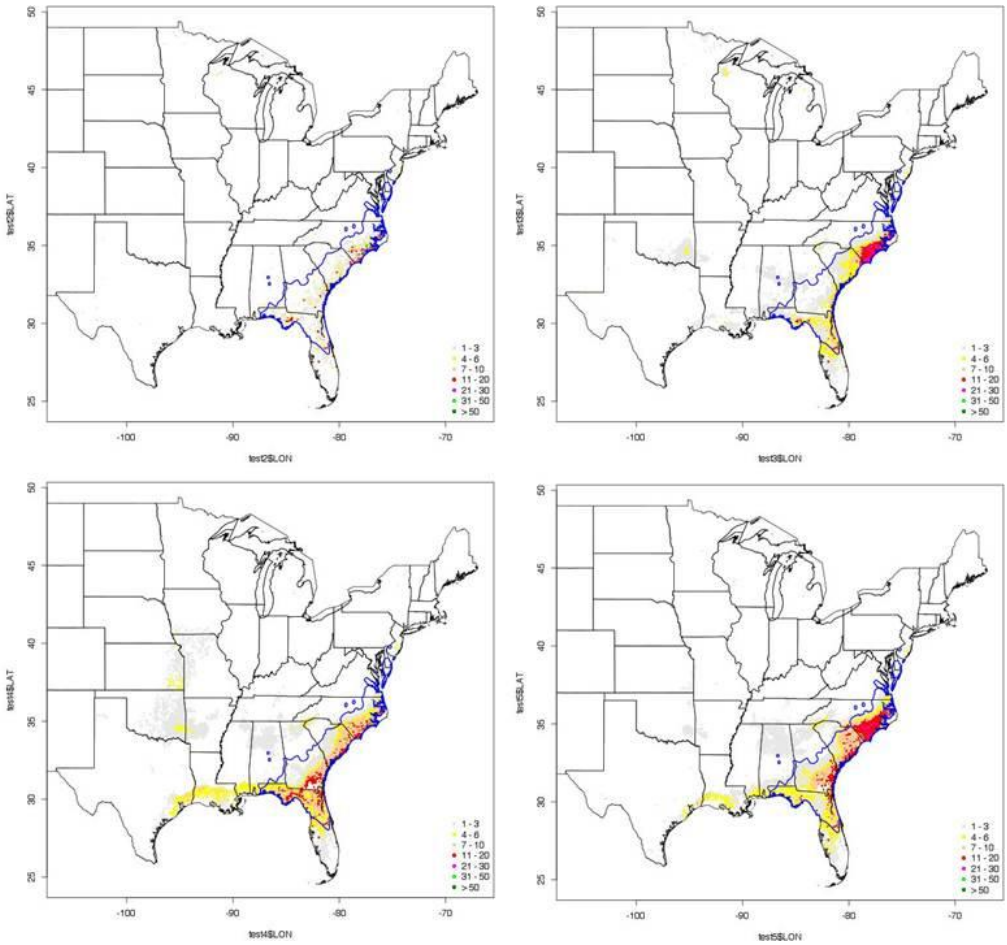


Figure A - 69. Combined IV and PA Pond pine CGCM3 A1B models (top-left), 2030 (top-right), 2060 (bottom-left), 2090 (bottom-right). Also referenced as Figure 39 in Results.

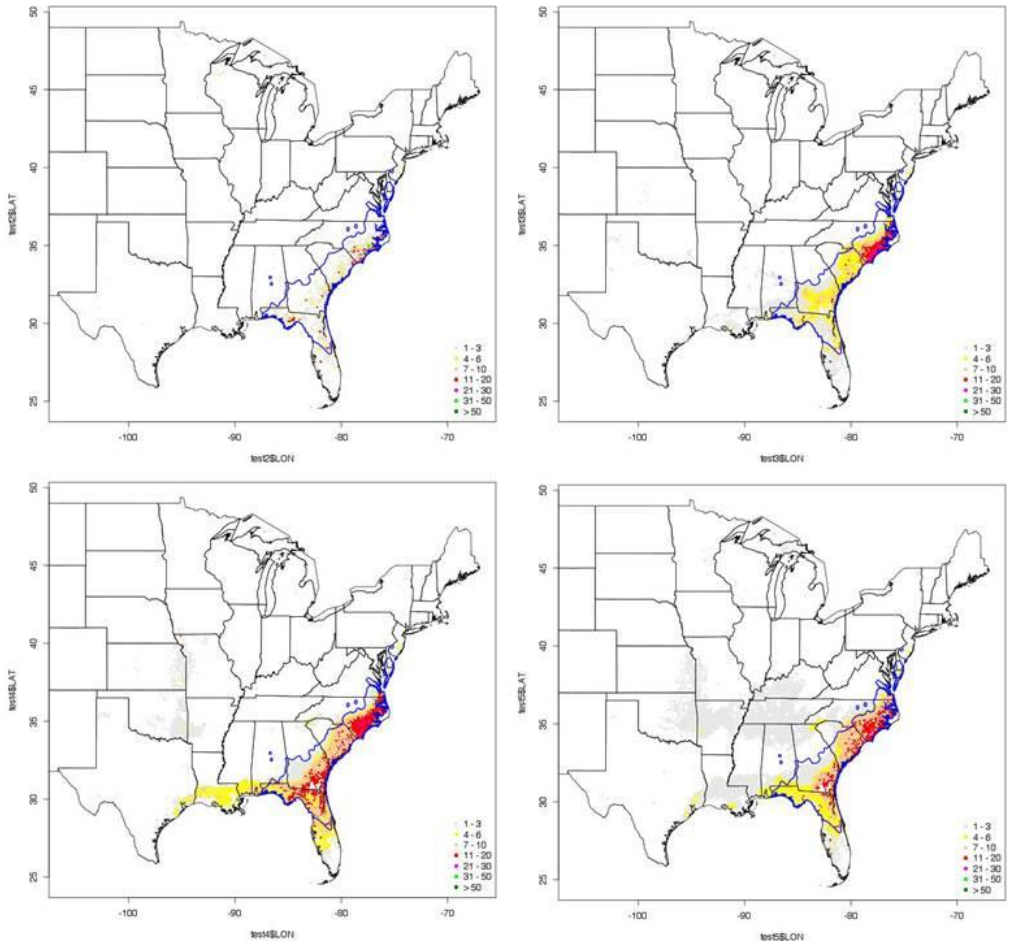


Figure A - 70. IV Pond pine CGCM3 A2 models (top-left), 2030 (top-right), 2060 (bottom-left), 2090 (bottom-right)

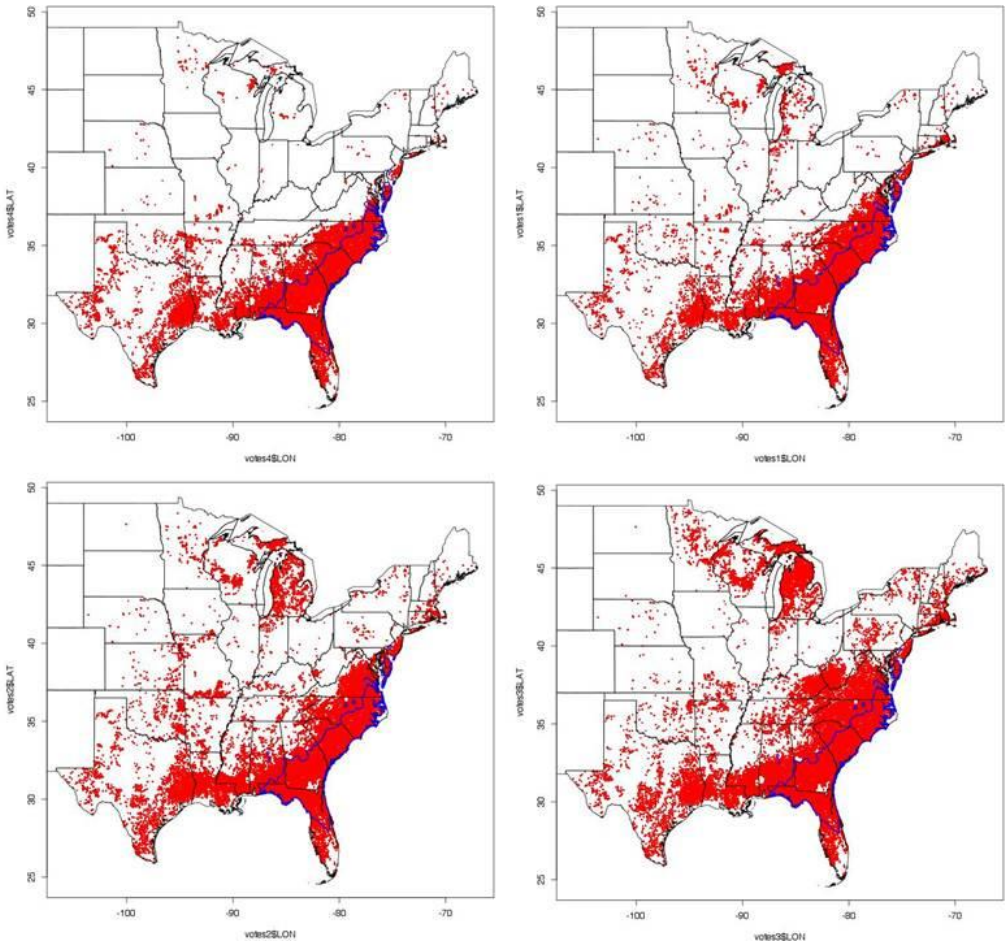


Figure A - 71. PA Pond pine CGCM3 A2 models (top-left), 2030 (top-right), 2060 (bottom-left), 2090 (bottom-right)

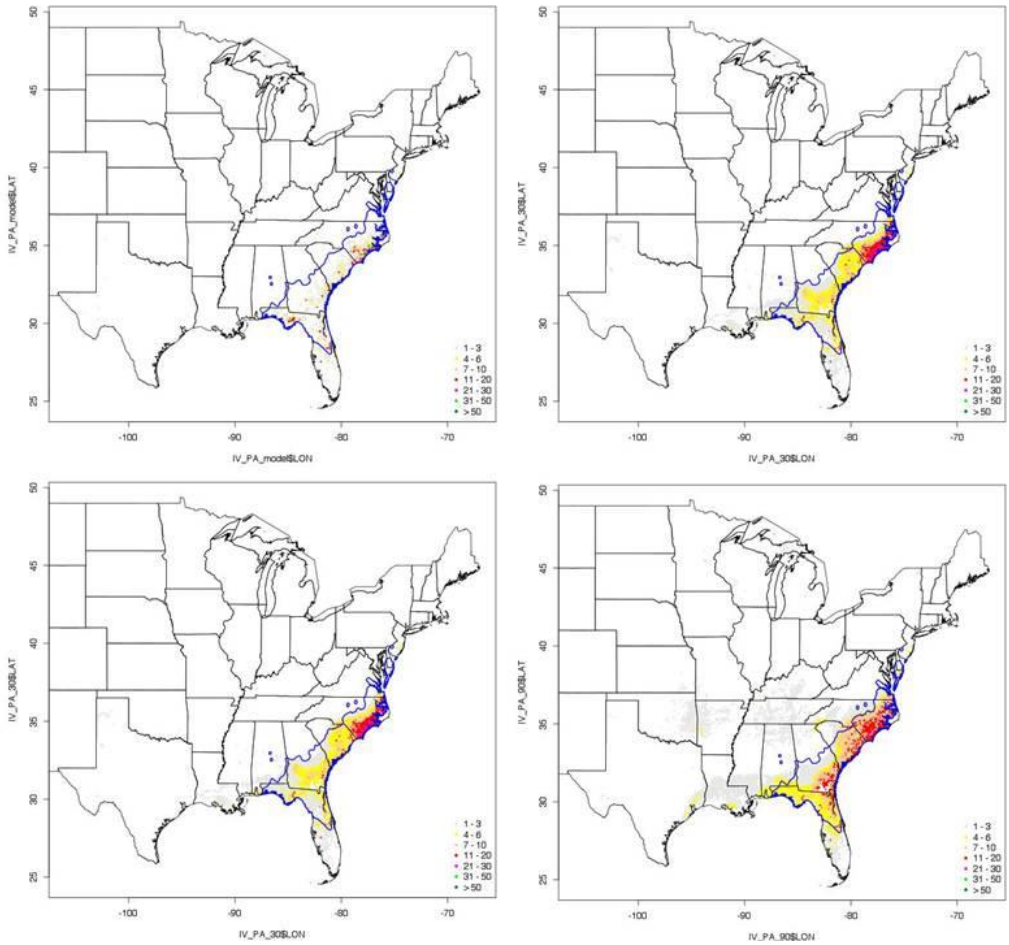


Figure A - 72. Combined IV and PA Pond pine CGCM3 A2 models (top-left), 2030 (top-right), 2060 (bottom-left), 2090 (bottom-right)

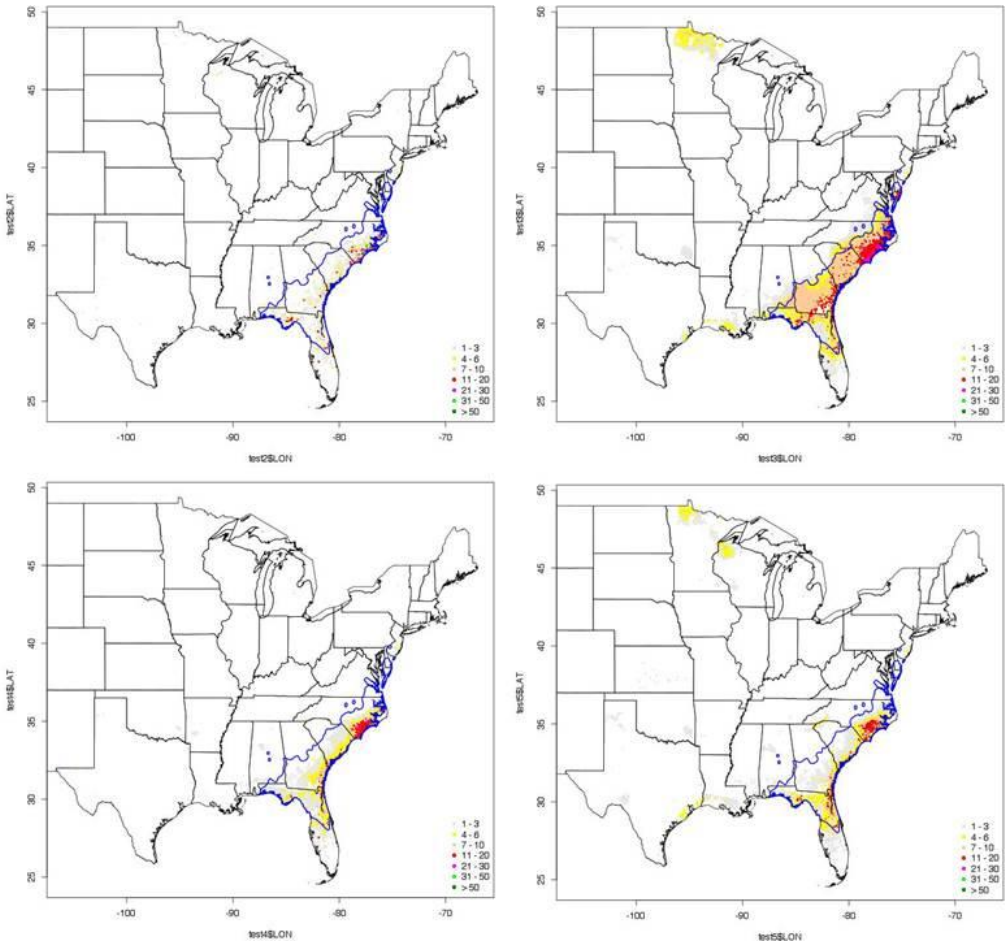


Figure A - 73. IV Pond pine CGCM3 B1 models (top-left), 2030 (top-right), 2060 (bottom-left), 2090 (bottom-right)

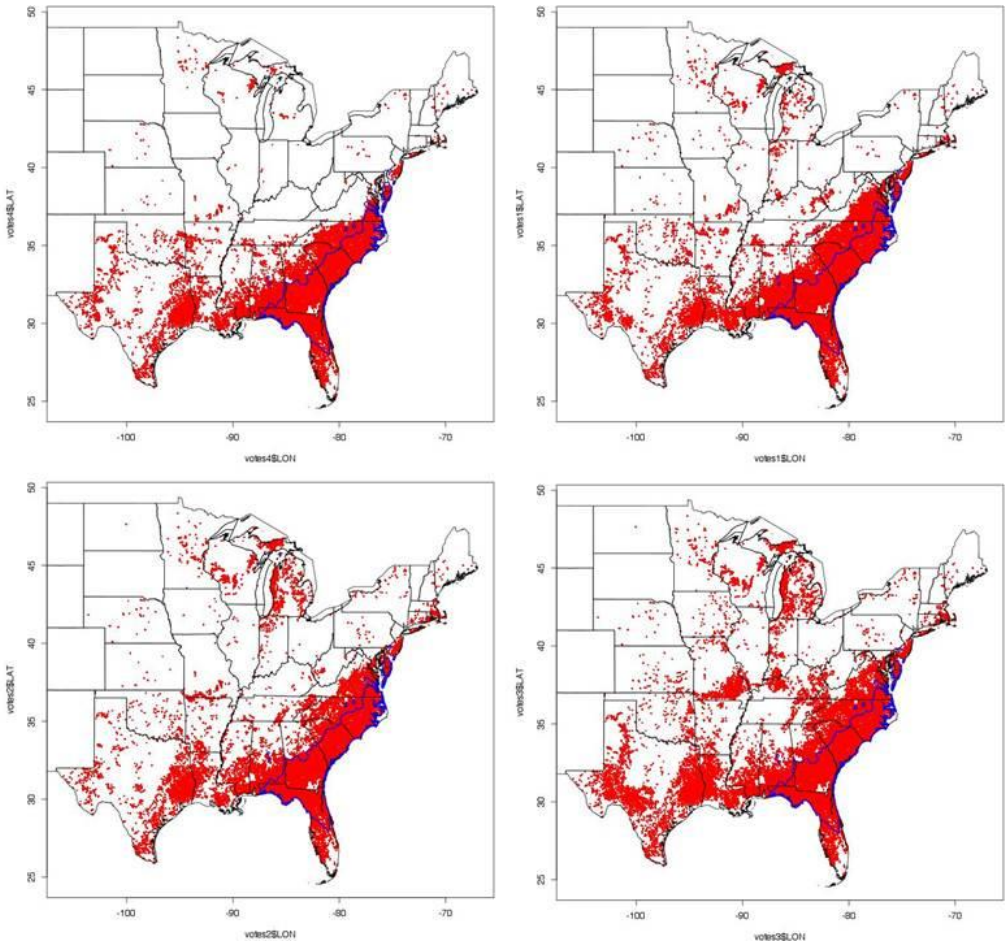


Figure A - 74. PA Pond pine CGCM3 B1 models (top-left), 2030 (top-right), 2060 (bottom-left), 2090 (bottom-right)

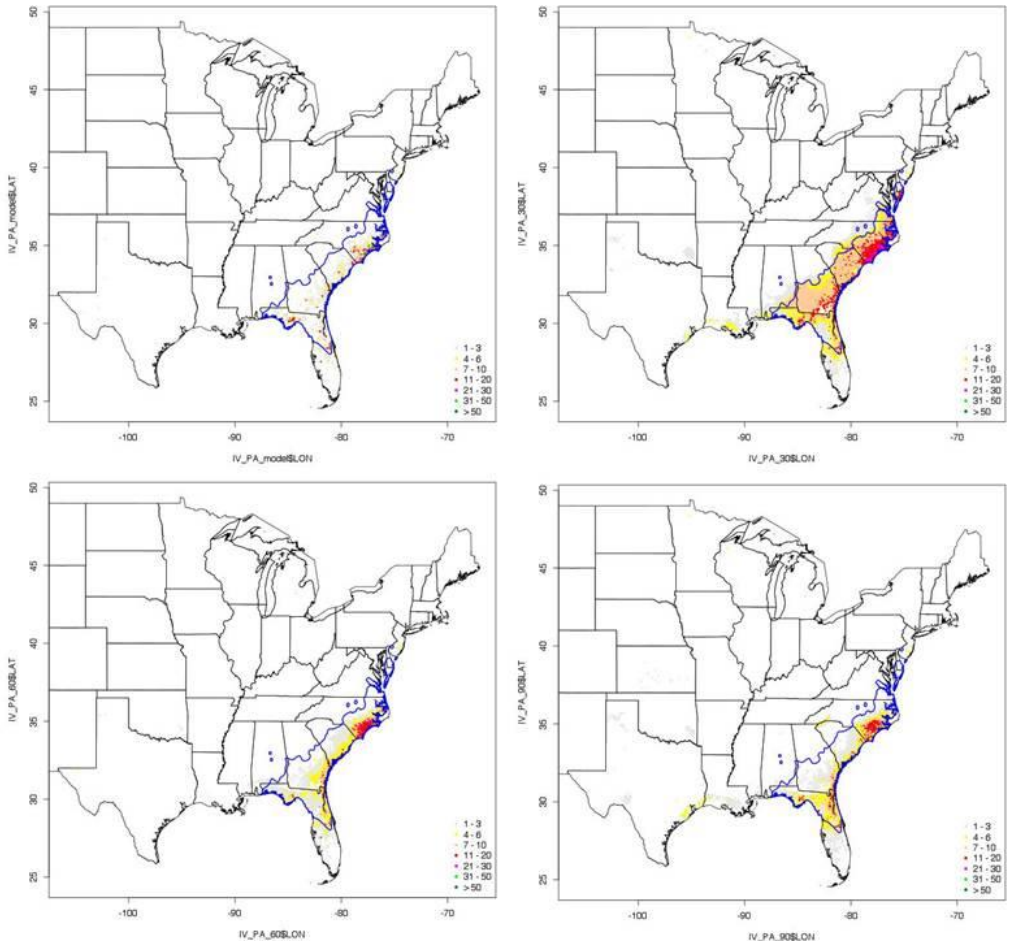


Figure A - 75. Combined IV and PA Pond pine CGCM3 B1 models (top-left), 2030 (top-right), 2060 (bottom-left), 2090 (bottom-right)

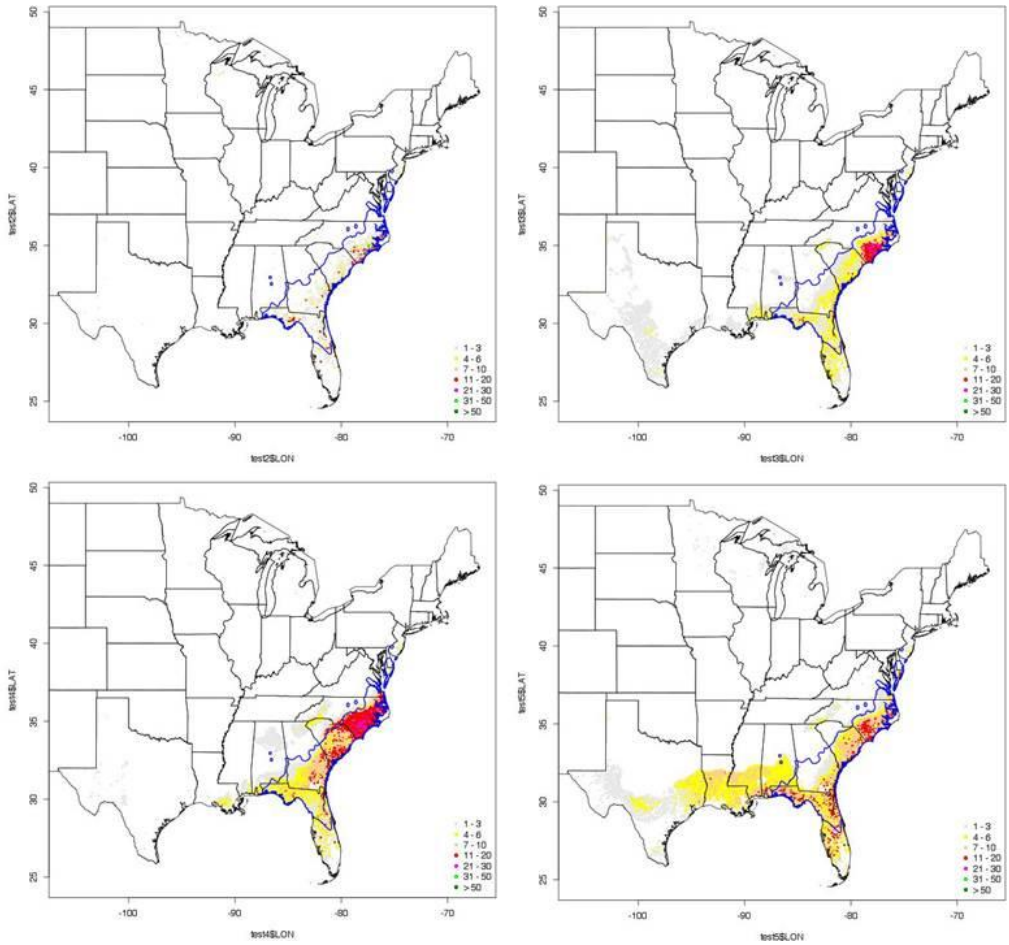


Figure A - 76. IV Pond pine GFDLCM21 A2 models (top-left), 2030 (top-right), 2060 (bottom-left), 2090 (bottom-right)

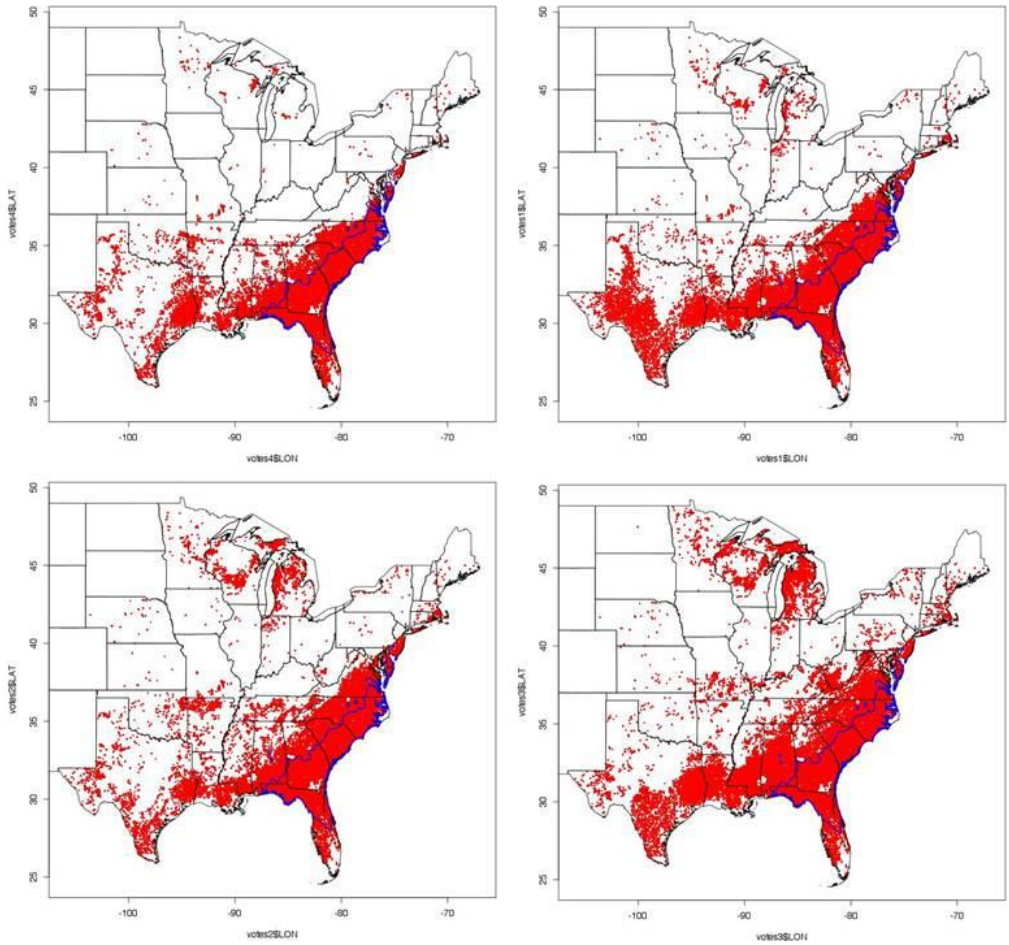


Figure A - 77. PA Pond pine GFDLCM21 A2 models (top-left), 2030 (top-right), 2060 (bottom-left), 2090 (bottom-right)

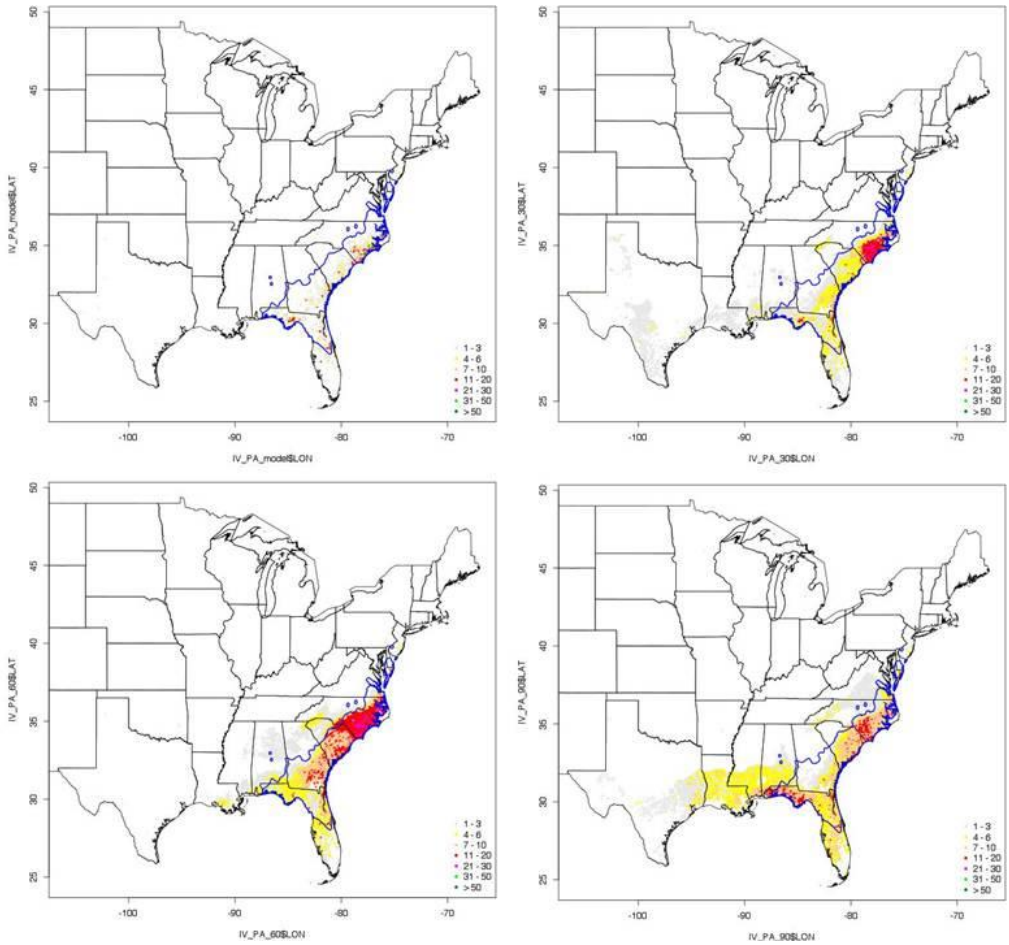


Figure A - 78. Combined IV and PA Pond pine GFDLCM21 A2 models (top-left), 2030 (top-right), 2060 (bottom-left), 2090 (bottom-right)

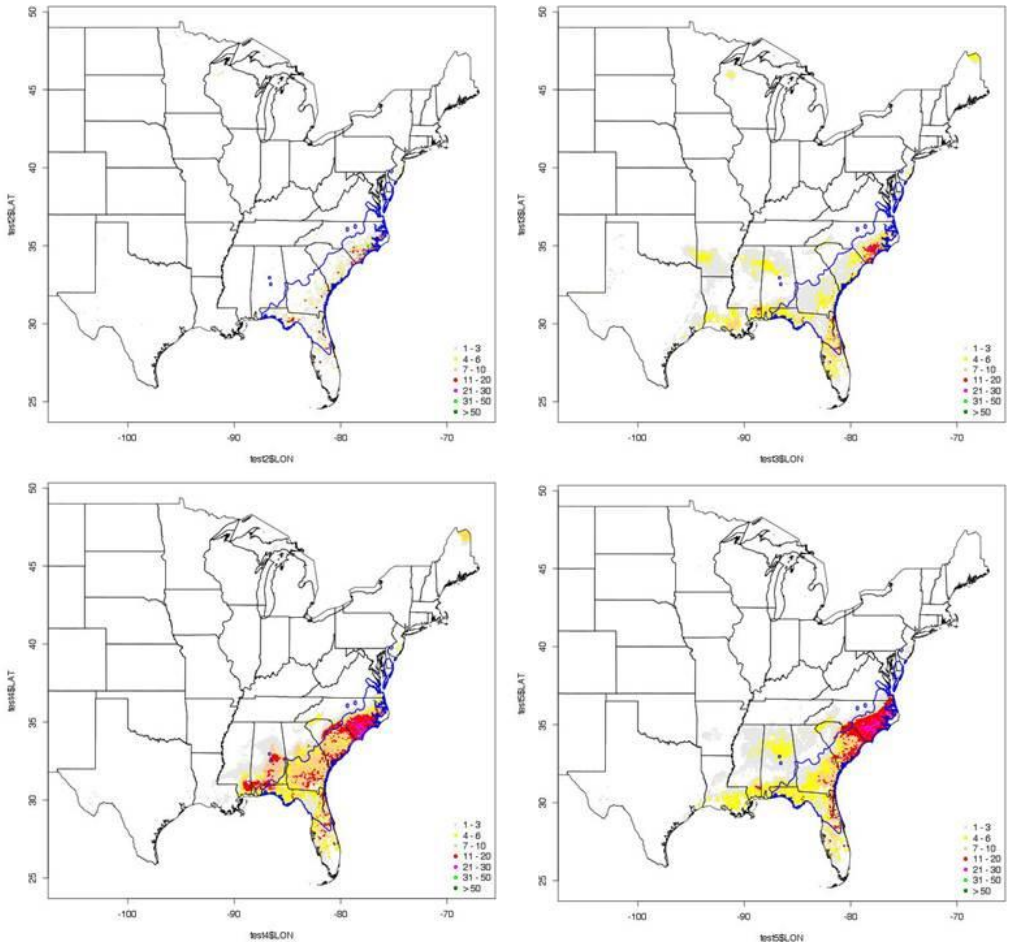


Figure A - 79. IV Pond pine GFDLCM21 B1 models (top-left), 2030 (top-right), 2060 (bottom-left), 2090 (bottom-right)

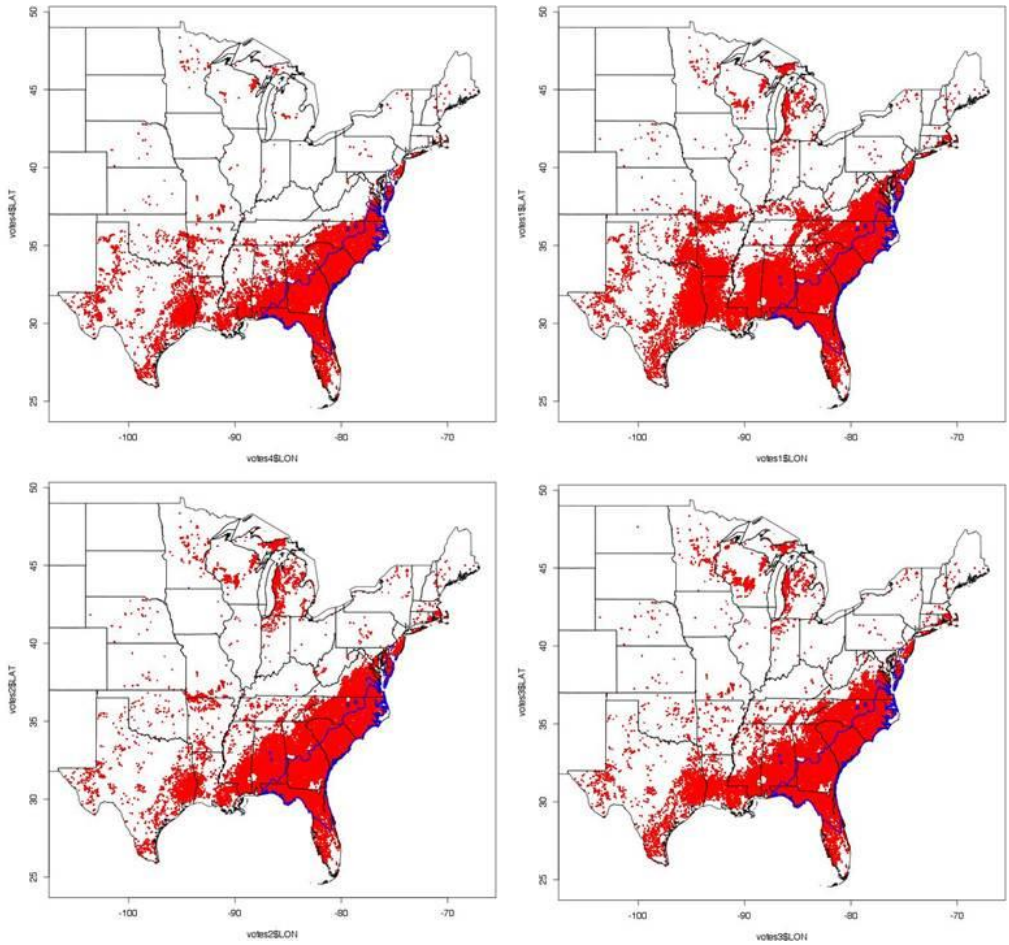


Figure A - 80. PA Pond pine GFDLCM3 B1 models (top-left), 2030 (top-right), 2060 (bottom-left), 2090 (bottom-right)

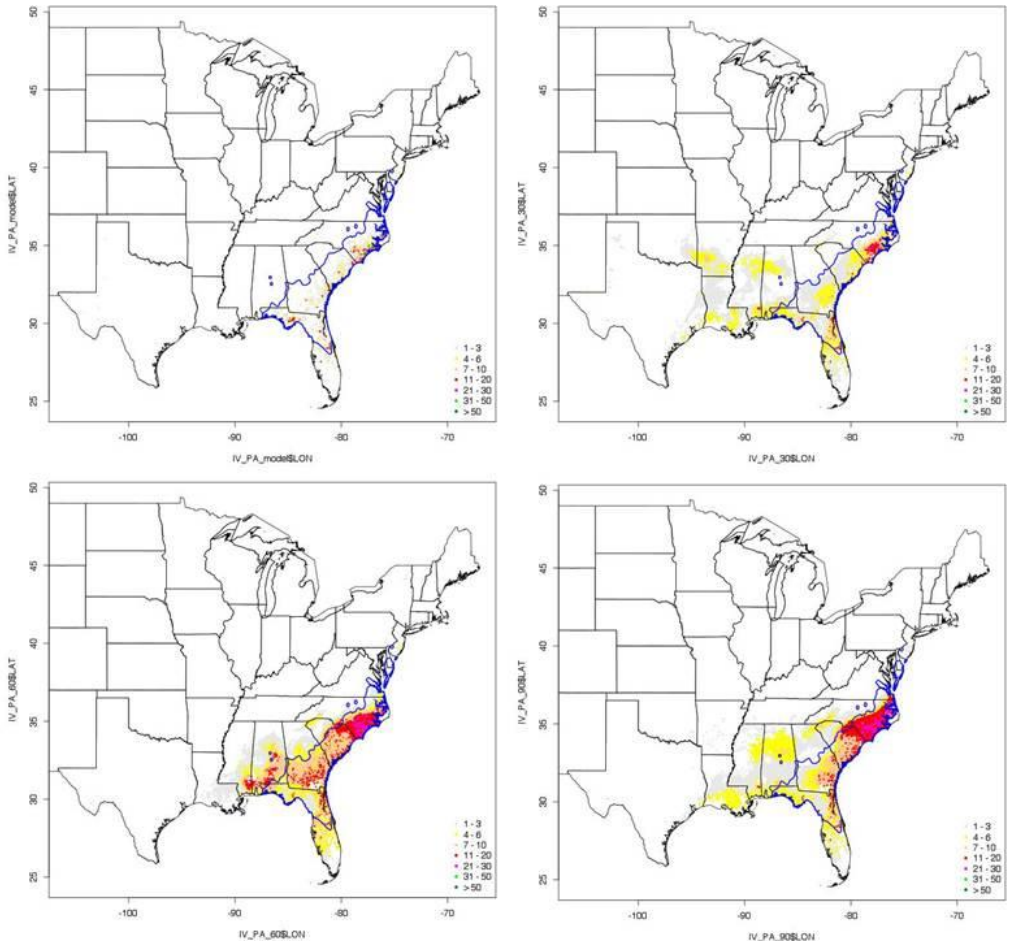


Figure A - 81. Combined IV and PA Pond pine GFDLCM21 B1 models (top-left), 2030 (top-right), 2060 (bottom-left), 2090 (bottom-right)

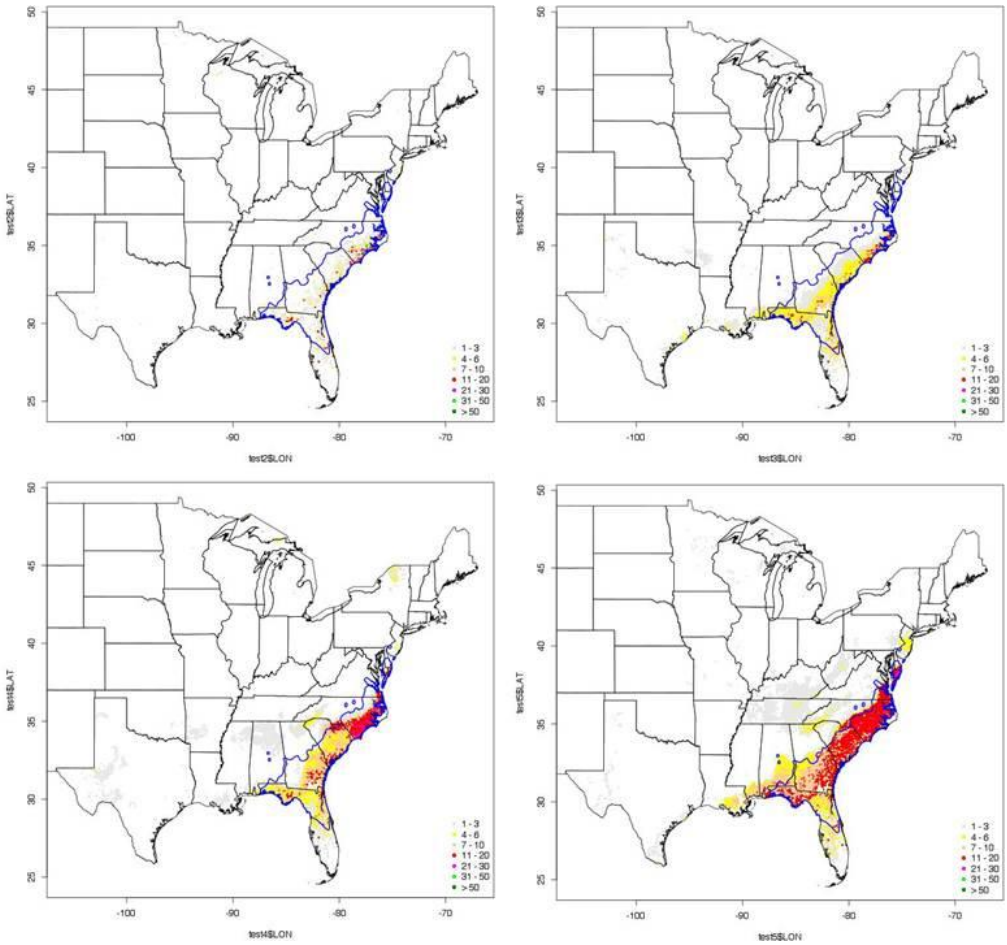


Figure A - 82. IV Pond pine HADCM3 A2 models (top-left), 2030 (top-right), 2060 (bottom-left), 2090 (bottom-right)

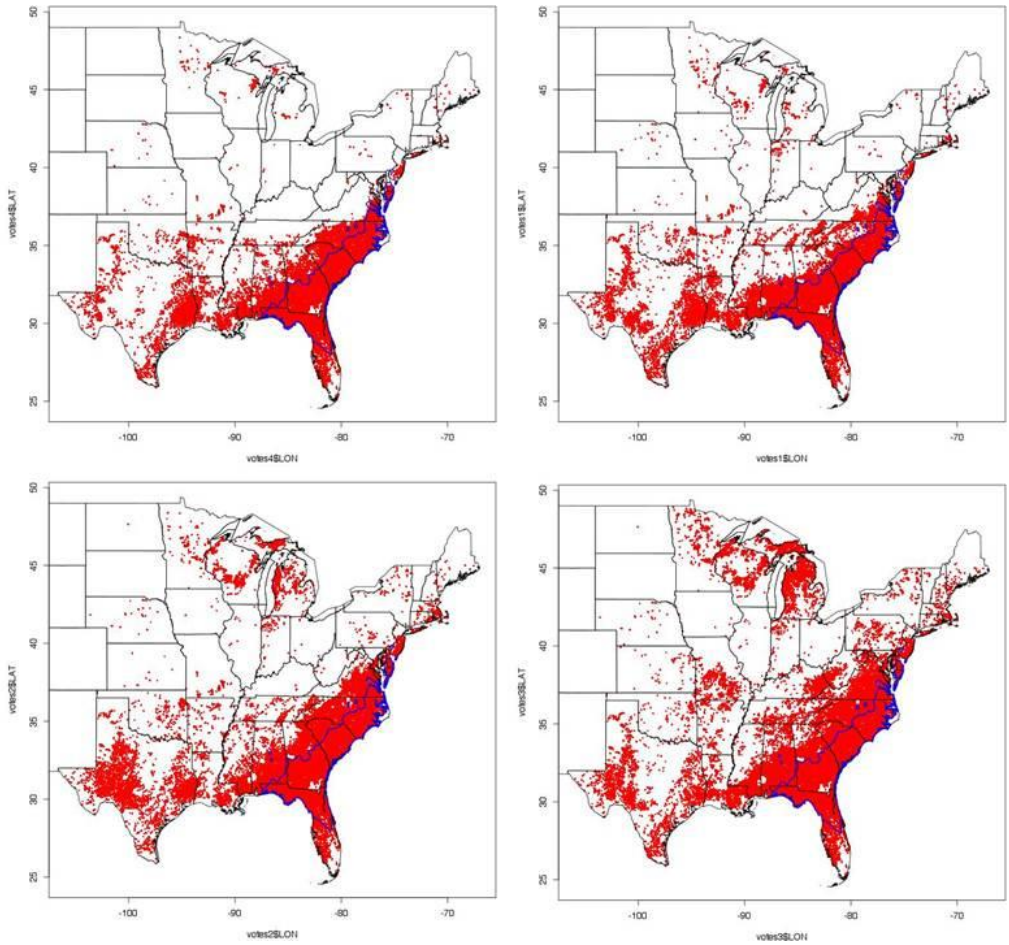


Figure A - 83. PA Pond pine HADCM3 A2 models (top-left), 2030 (top-right), 2060 (bottom-left), 2090 (bottom-right)

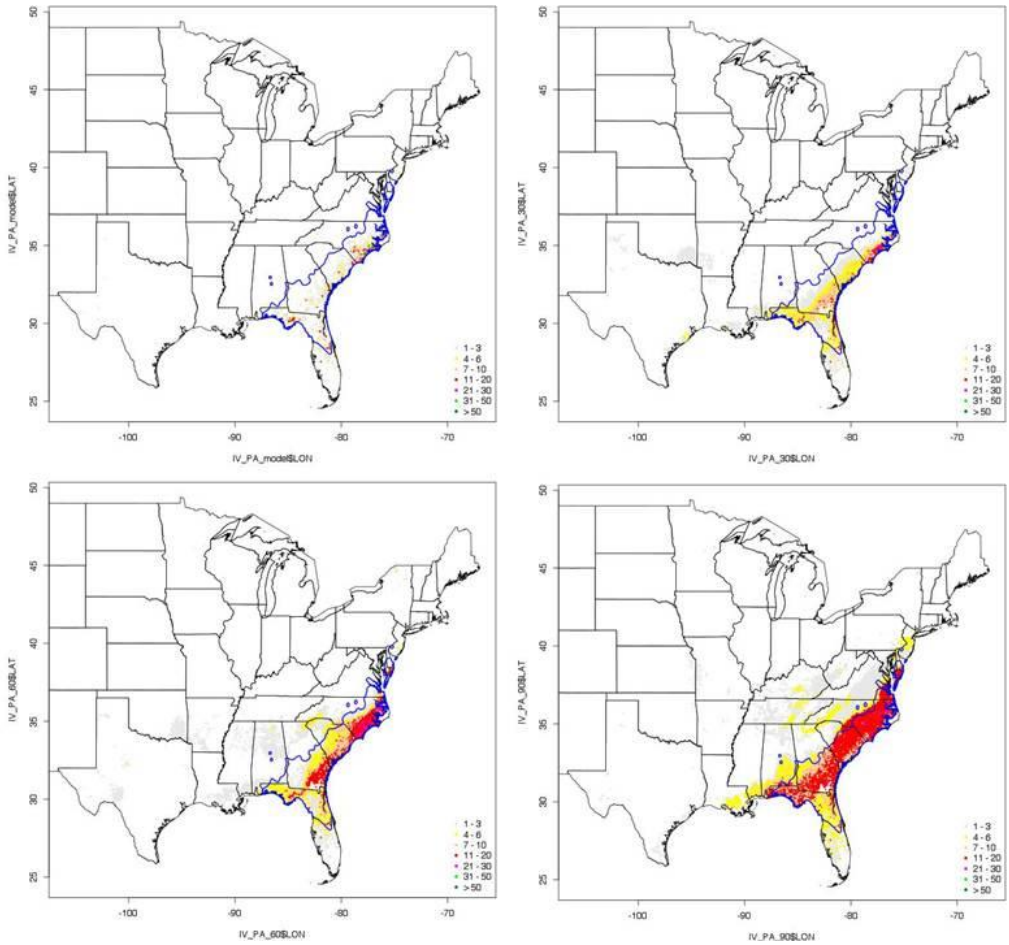


Figure A - 84. Combined IV and PA Pond pine HADCM3 A2 models (top-left), 2030 (top-right), 2060 (bottom-left), 2090 (bottom-right)

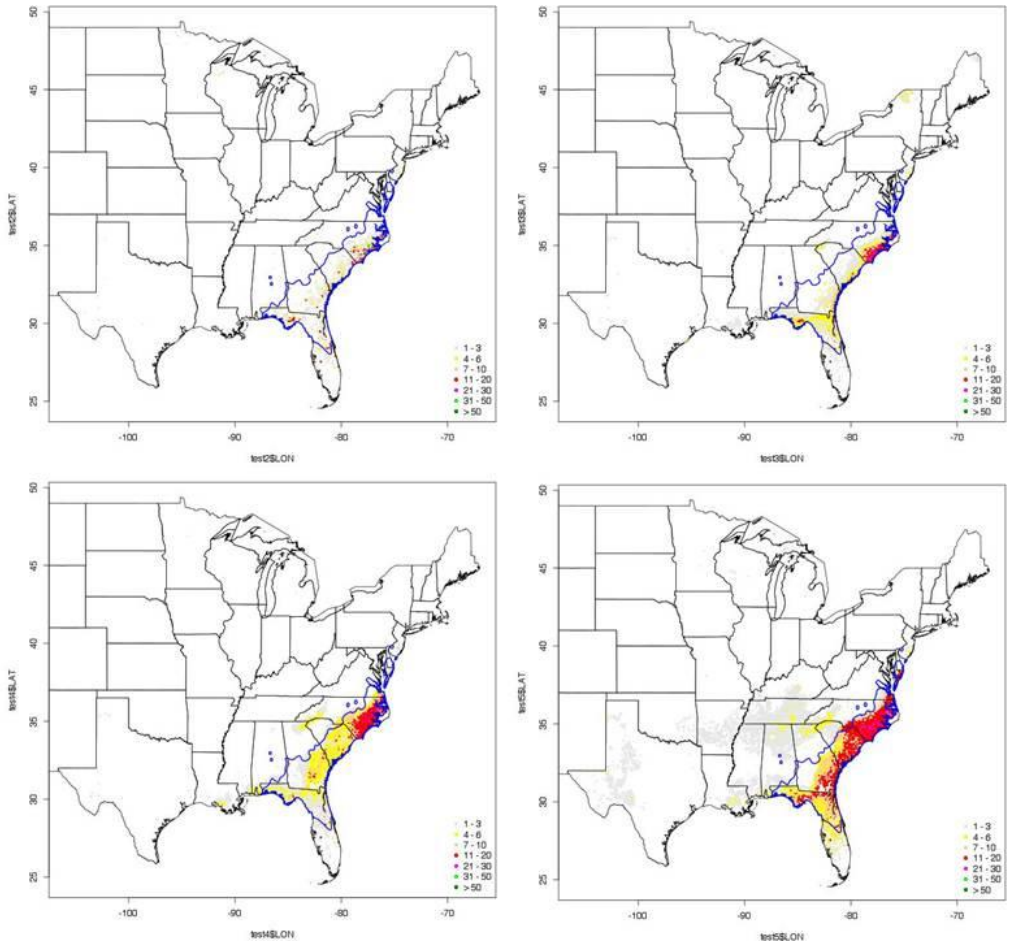


Figure A - 85. IV Pond pine HADCM3 B2 models (top-left), 2030 (top-right), 2060 (bottom-left), 2090 (bottom-right)

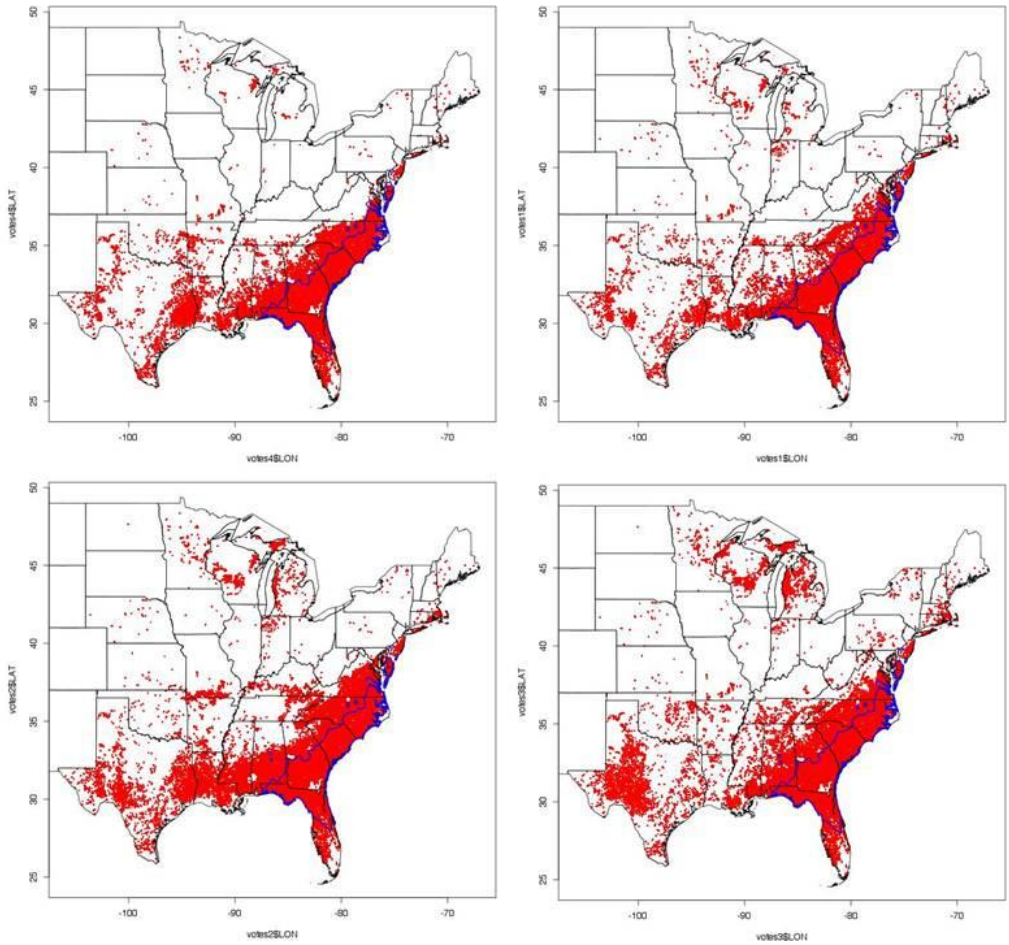


Figure A - 86. PA Pond pine HADCM3 B2 models (top-left), 2030 (top-right), 2060 (bottom-left), 2090 (bottom-right)

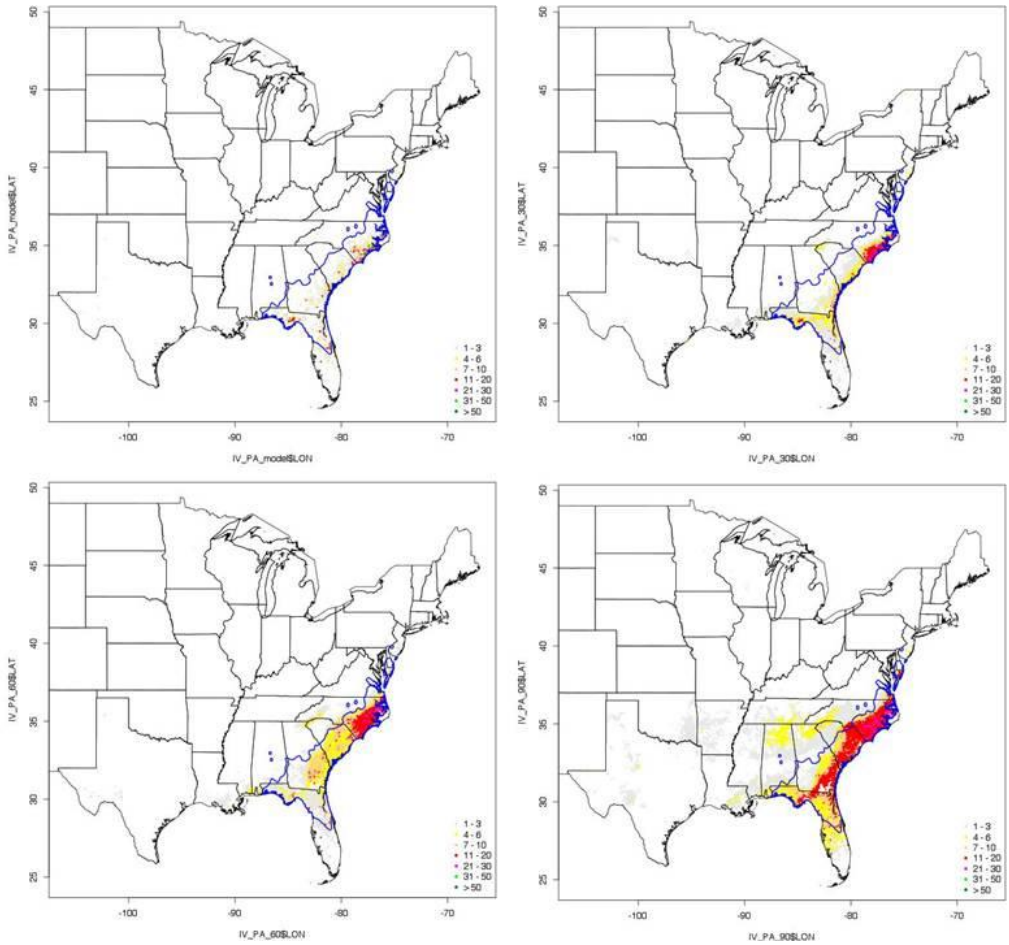


Figure A - 87. Combined IV and PA Pond pine HADCM3 B2 models (top-left), 2030 (top-right), 2060 (bottom-left), 2090 (bottom-right)

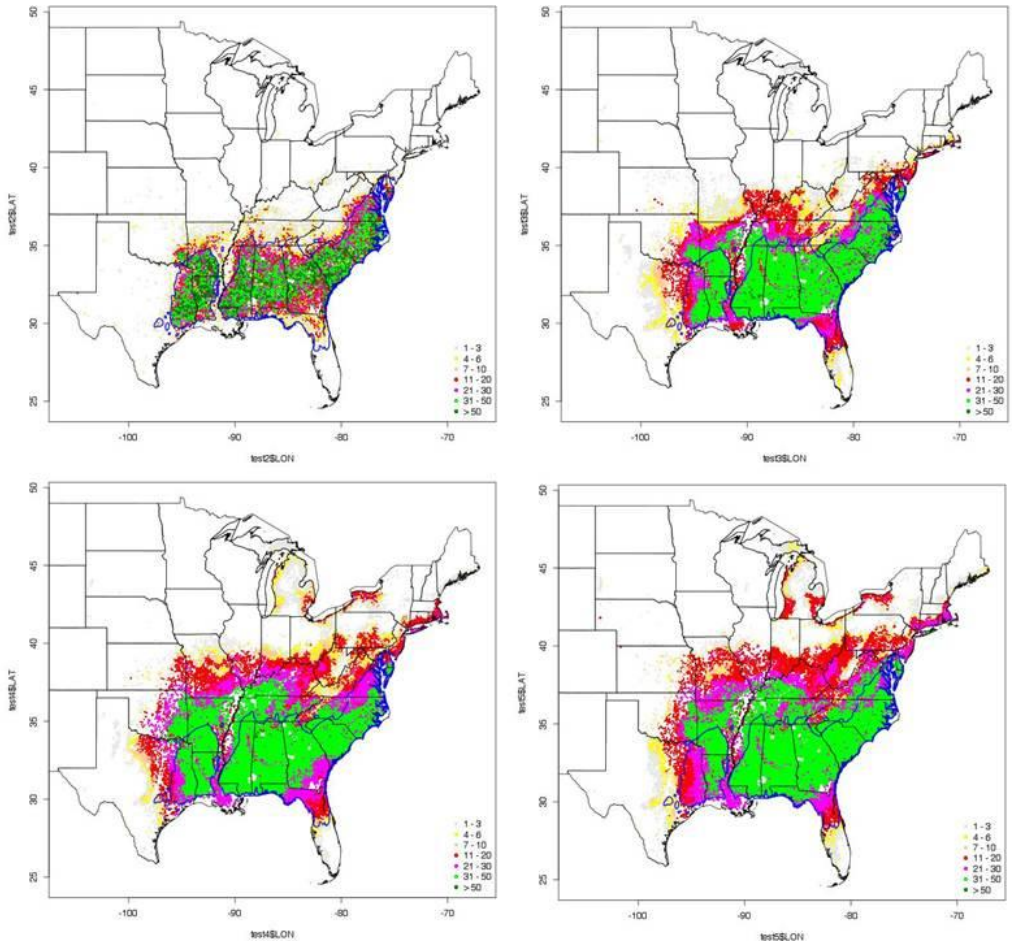


Figure A - 88. IV Loblolly pine CGCM3 A1B models (top-left), 2030 (top-right), 2060 (bottom-left), 2090 (bottom-right)

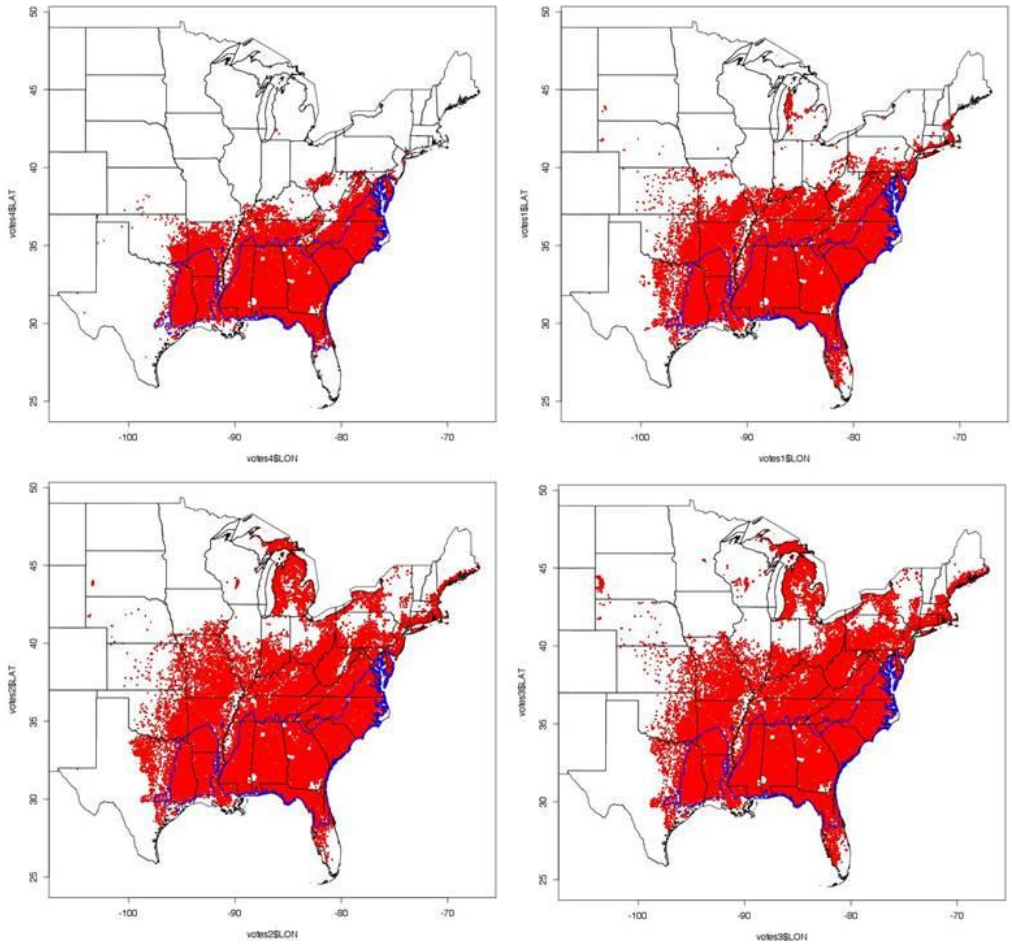


Figure A - 89. PA Loblolly pine CGCM3 A1B models (top-left), 2030 (top-right), 2060 (bottom-left), 2090 (bottom-right)

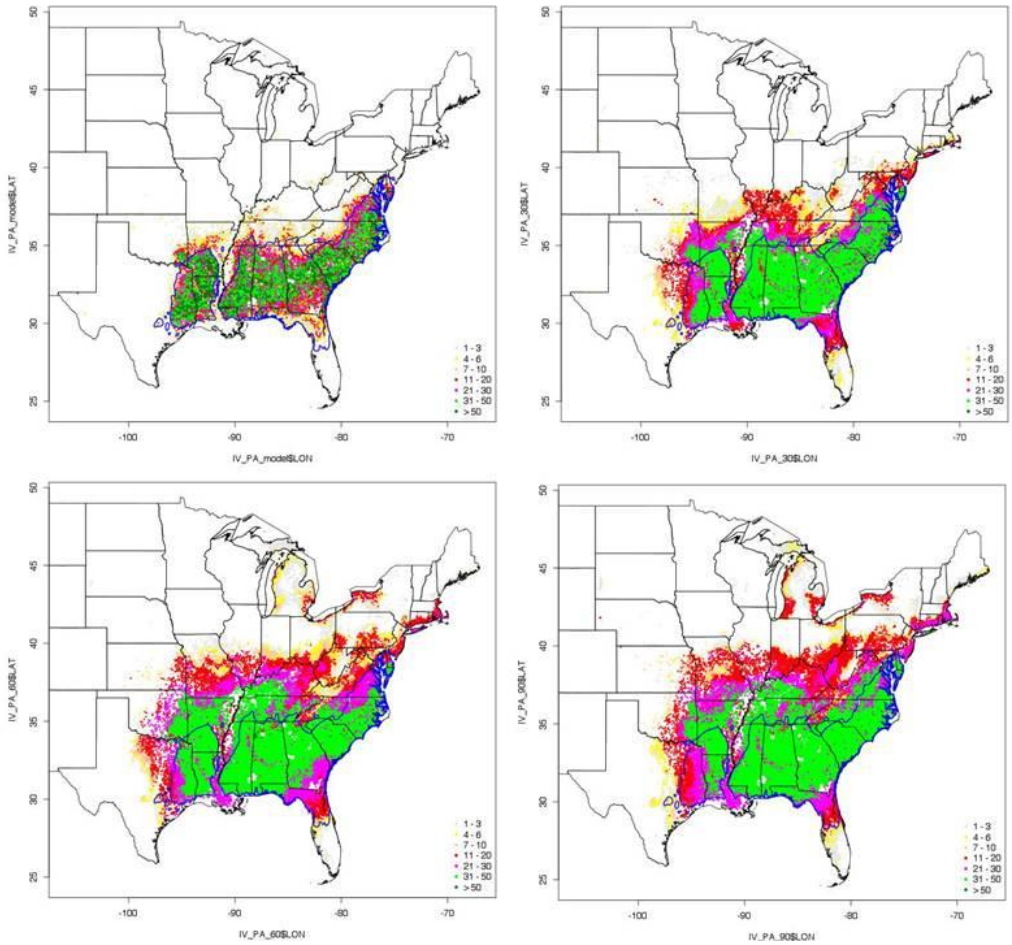


Figure A - 90. Combined IV and PA Loblolly pine CGCM3 A1B models (top-left), 2030 (top-right), 2060 (bottom-left), 2090 (bottom-right)

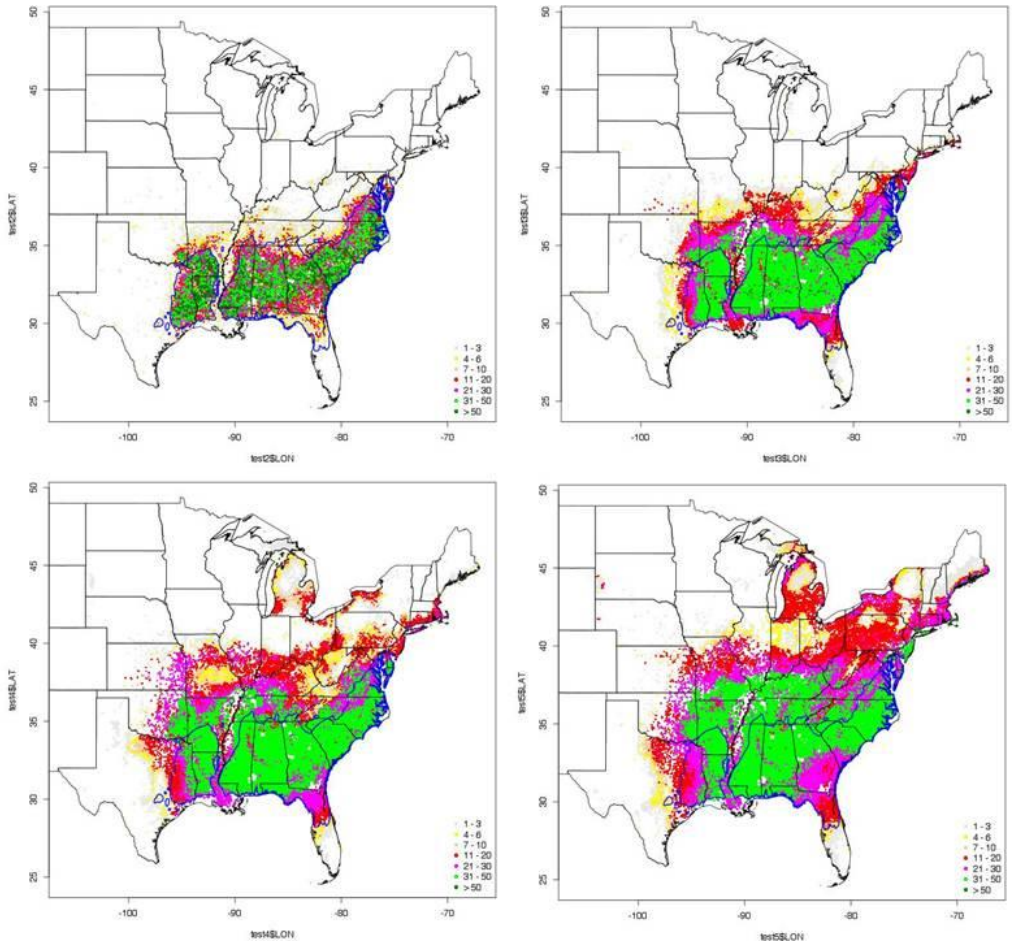


Figure A - 91. IV Loblolly pine CGCM3 A2 models (top-left), 2030 (top-right), 2060 (bottom-left), 2090 (bottom-right)

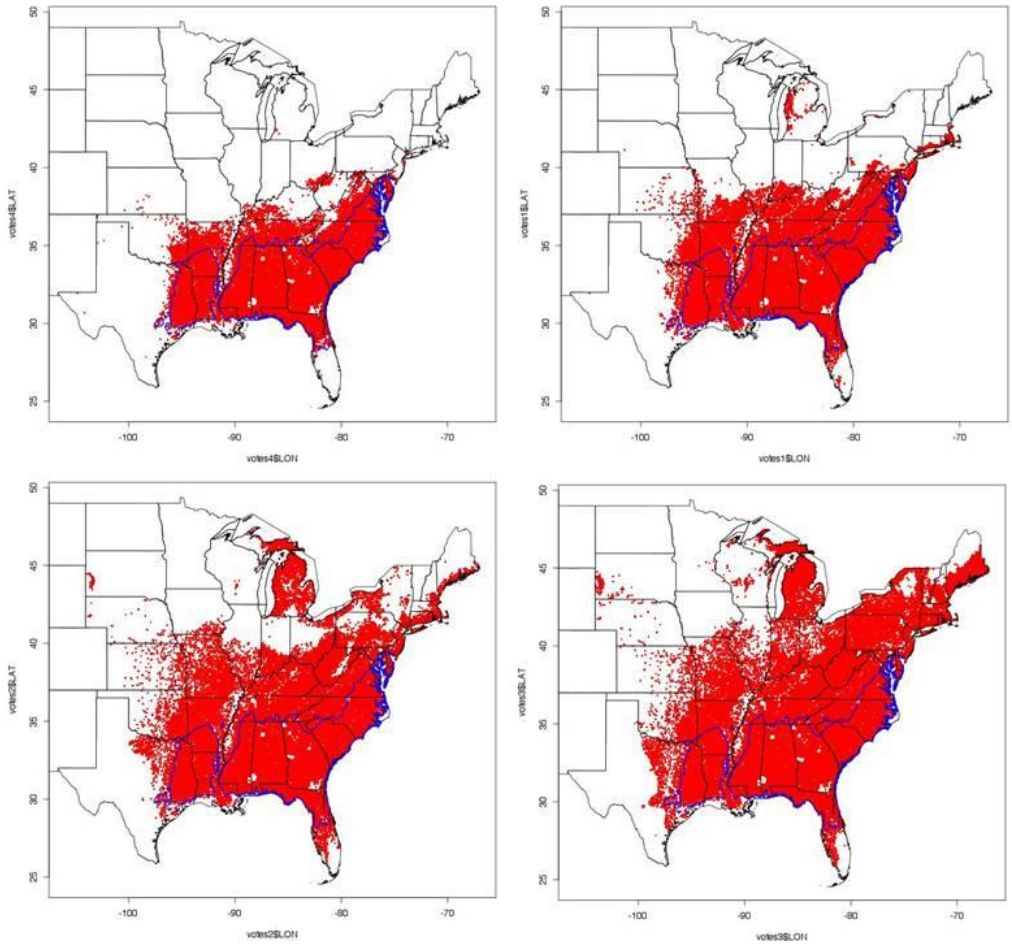


Figure A - 92. PA Loblolly pine CGCM3 A2 models (top-left), 2030 (top-right), 2060 (bottom-left), 2090 (bottom-right). Also referenced as **Error! Reference source not found.** in Results.

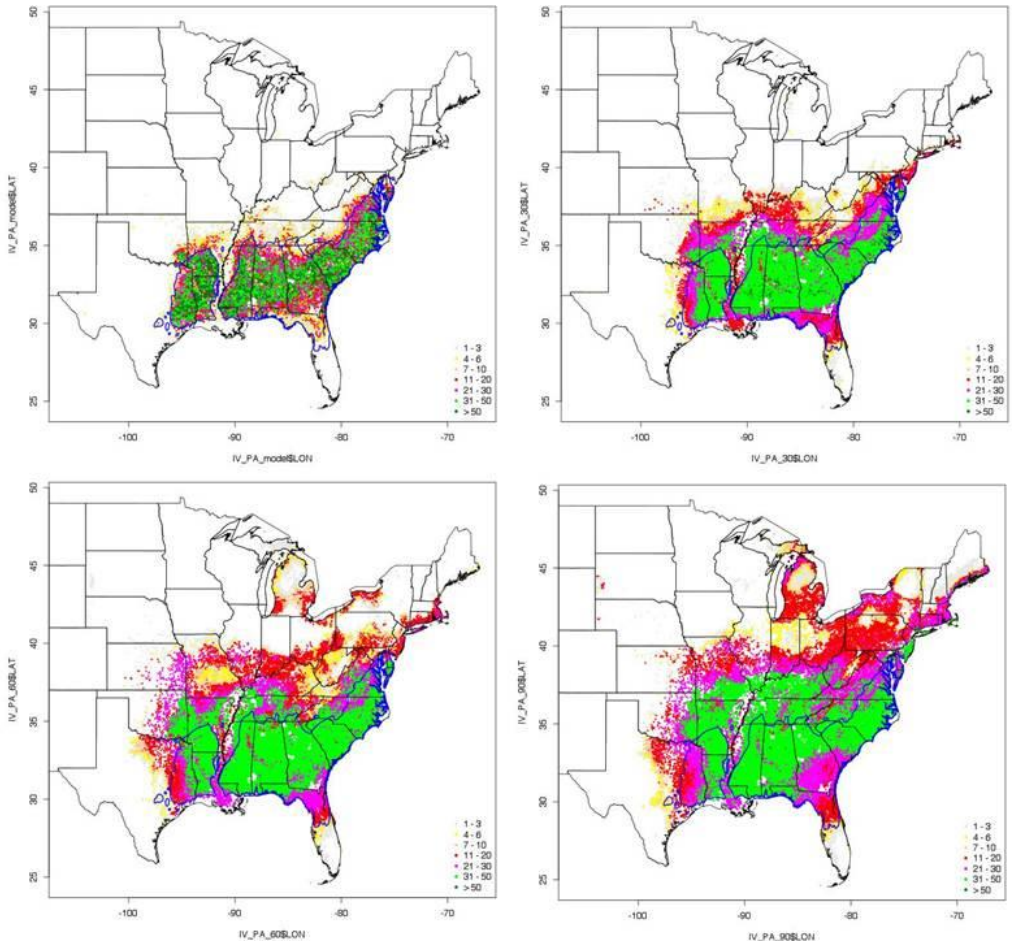


Figure A - 93. Combined IV and PA Loblolly pine CGCM3 A2 models (top-left), 2030 (top-right), 2060 (bottom-left), 2090 (bottom-right)

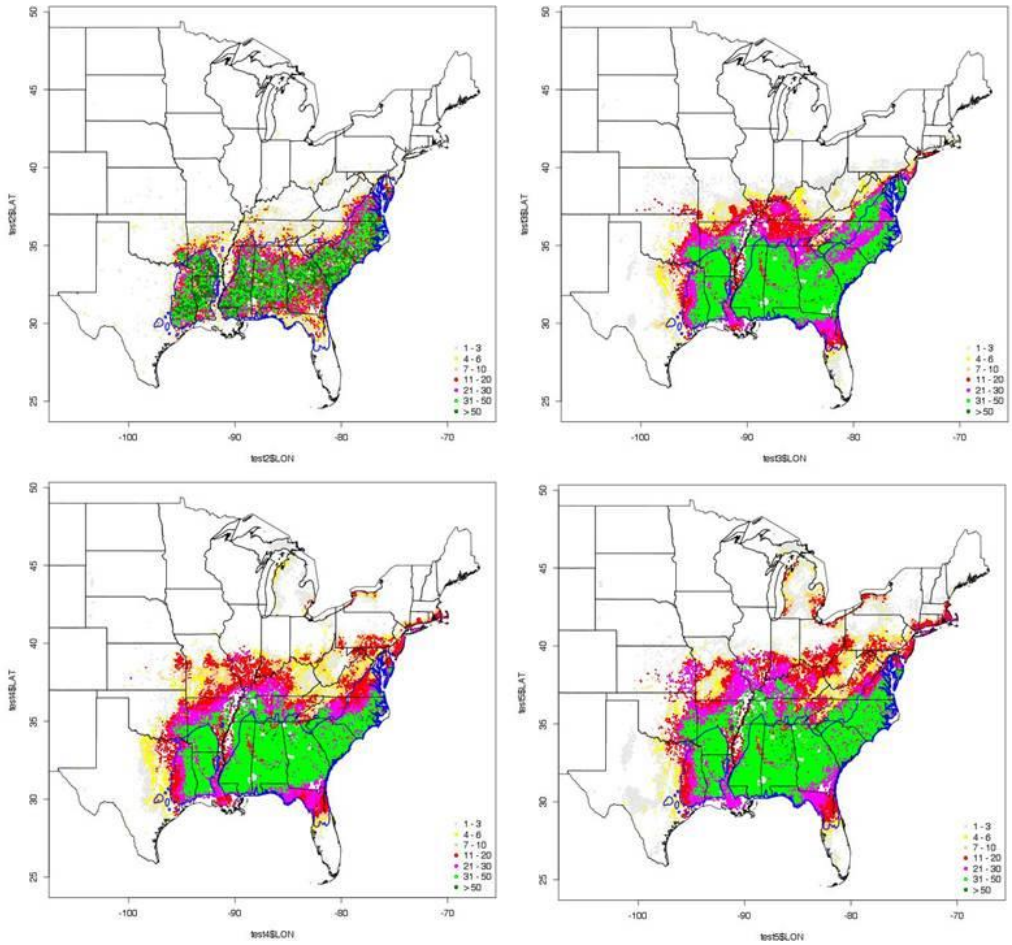


Figure A - 94. IV Loblolly pine CGCM3 B1 models (top-left), 2030 (top-right), 2060 (bottom-left), 2090 (bottom-right)

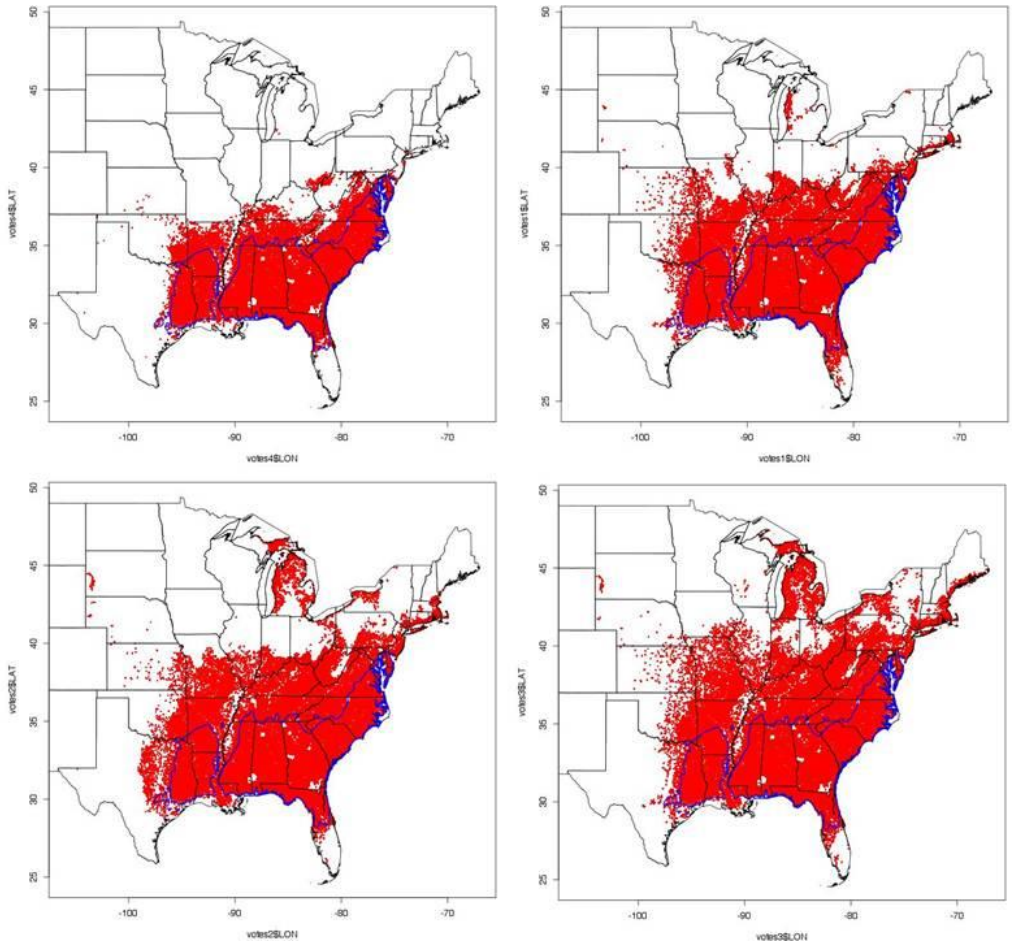


Figure A - 95. PA Loblolly pine CGCM3 B1 models (top-left), 2030 (top-right), 2060 (bottom-left), 2090 (bottom-right)

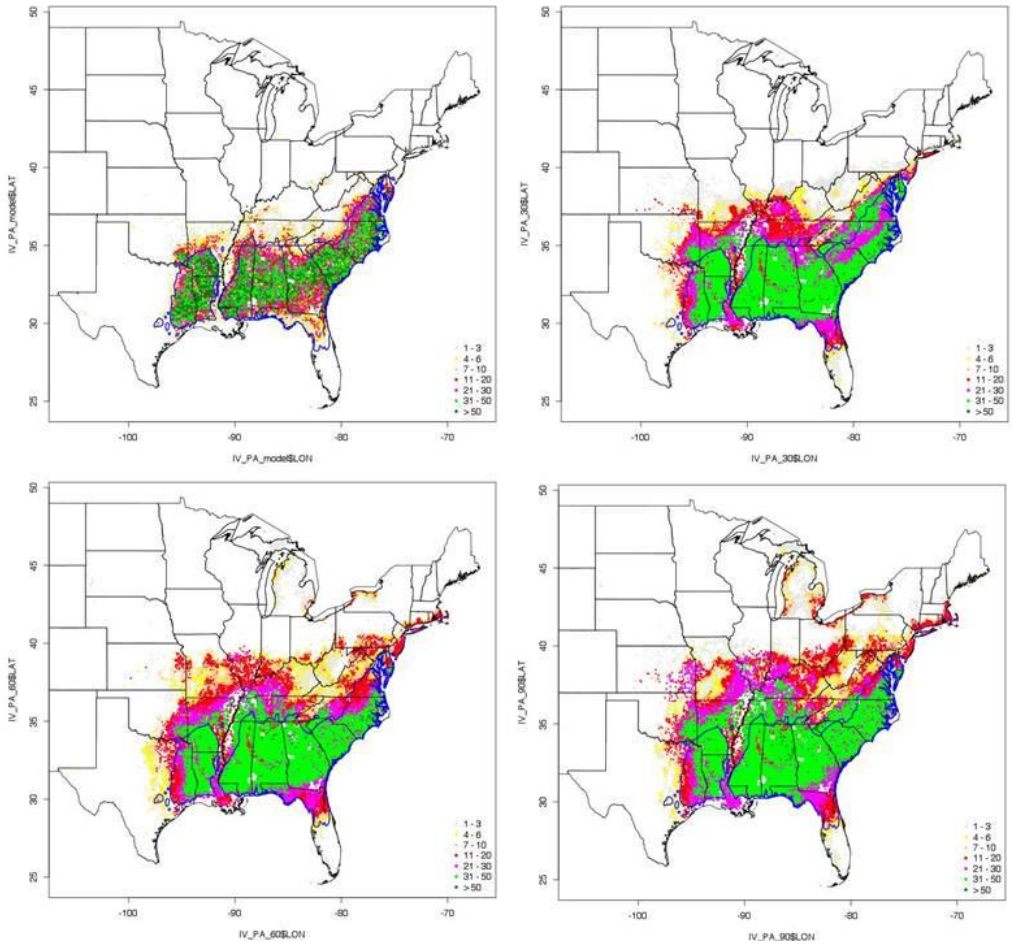


Figure A - 96. Combined IV and PA Loblolly pine CGCM3 B1 models (top-left), 2030 (top-right), 2060 (bottom-left), 2090 (bottom-right)

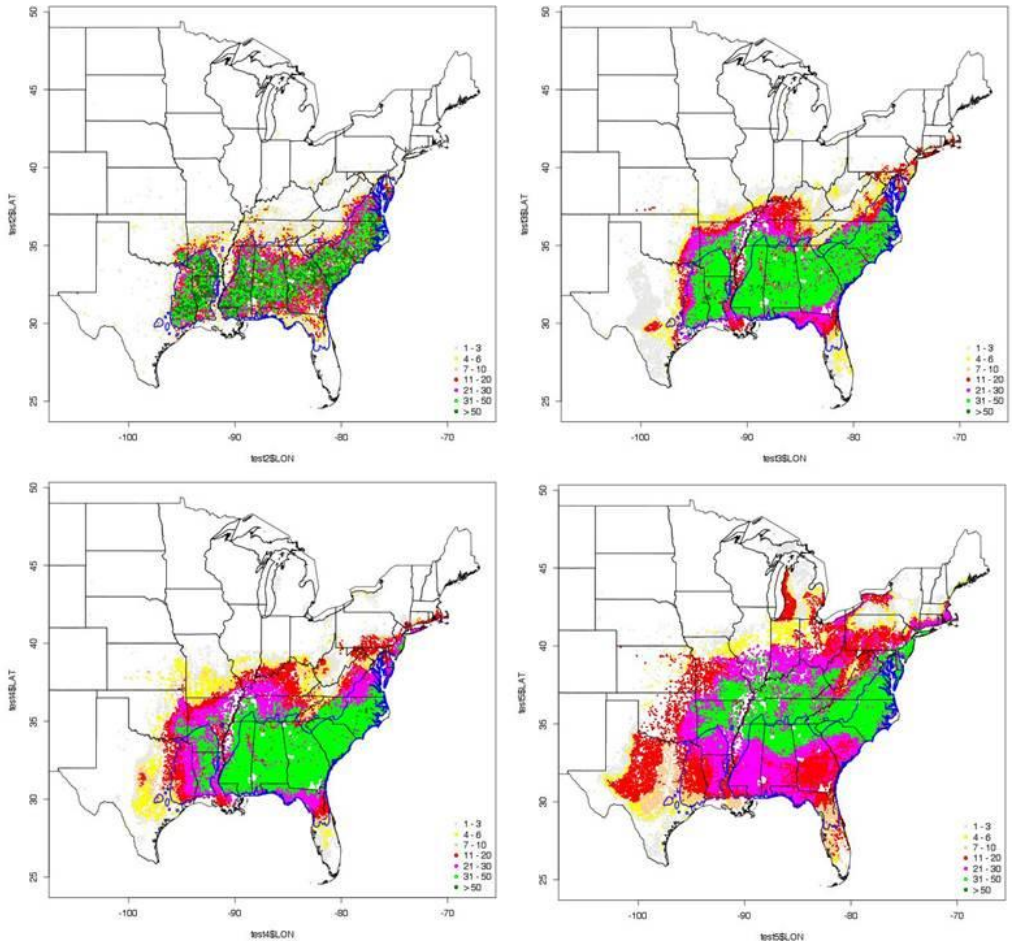


Figure A - 97. IV Loblolly pine GFDLCM21 A2 models (top-left), 2030 (top-right), 2060 (bottom-left), 2090 (bottom-right)

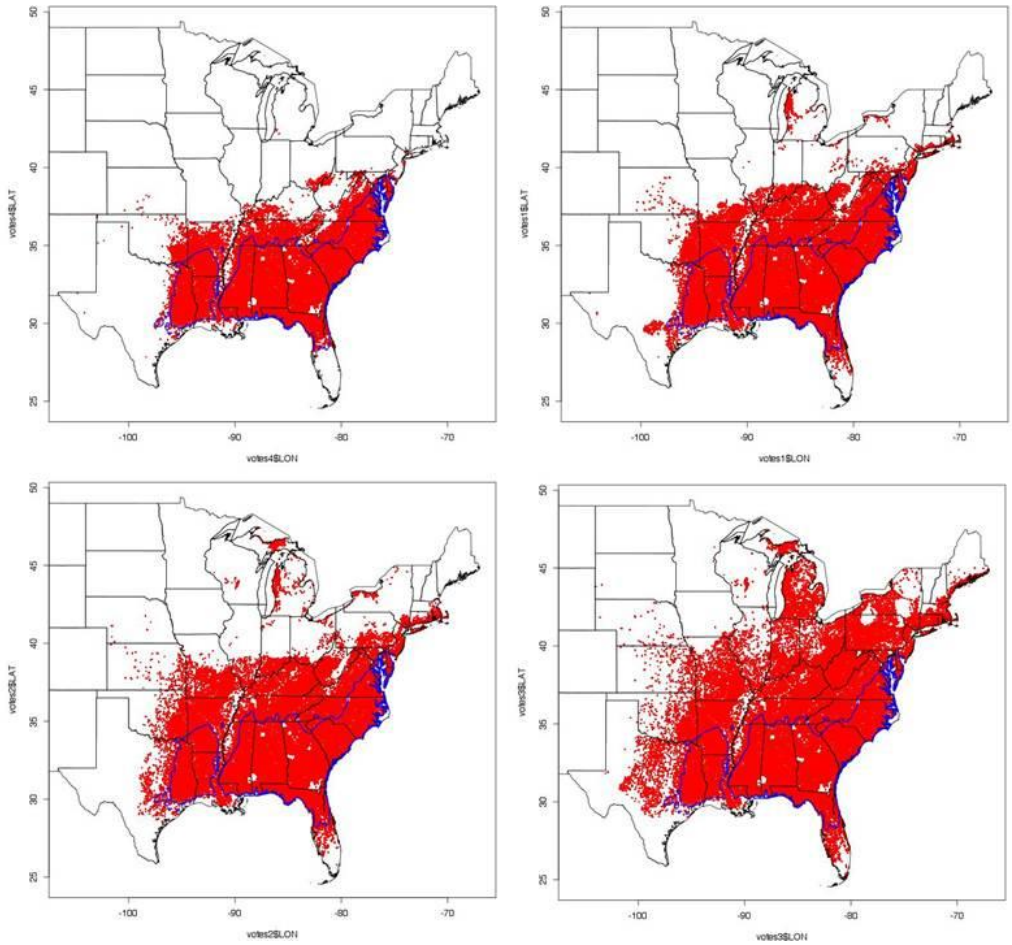


Figure A - 98. PA Loblolly pine GFDLCM21 A2 models (top-left), 2030 (top-right), 2060 (bottom-left), 2090 (bottom-right)

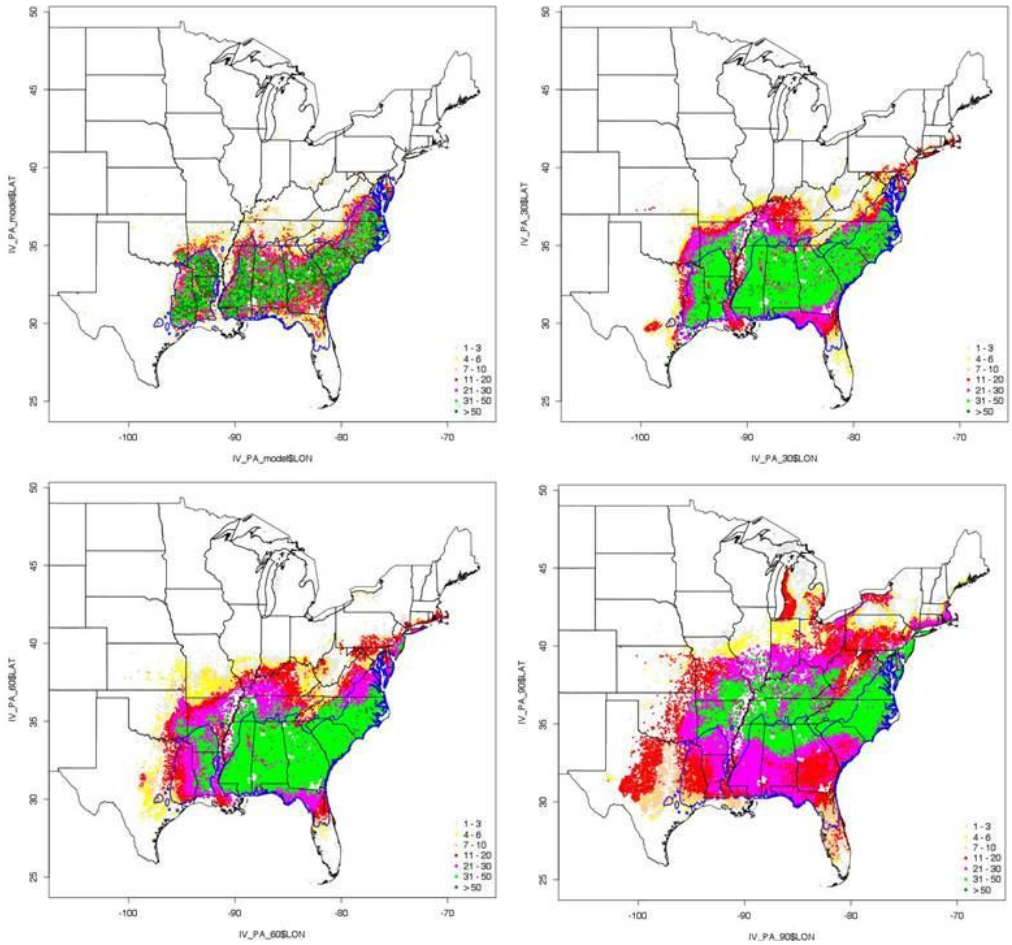


Figure A - 99. Combined IV and PA Loblolly pine GFDLCM21 A2 models (top-left), 2030 (top-right), 2060 (bottom-left), 2090 (bottom-right)

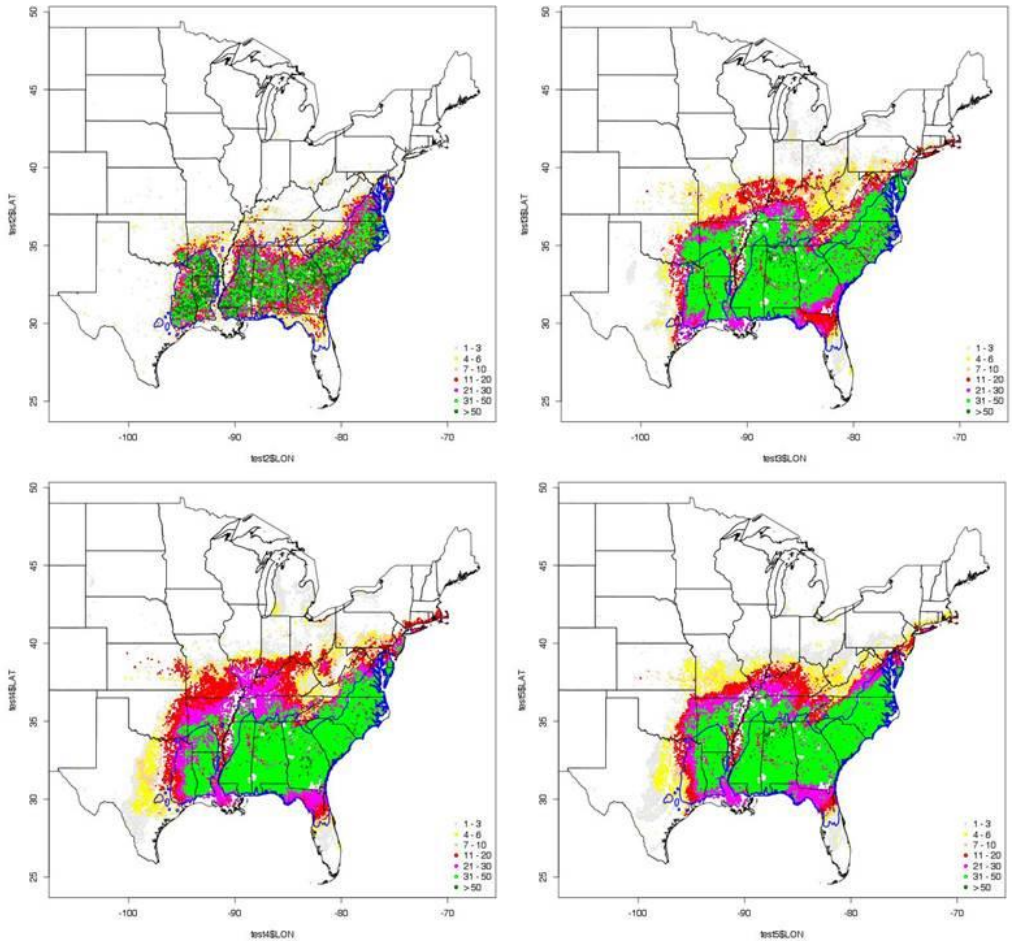


Figure A - 100. IV Loblolly pine GFDLCM21 B1 models (top-left), 2030 (top-right), 2060 (bottom-left), 2090 (bottom-right)

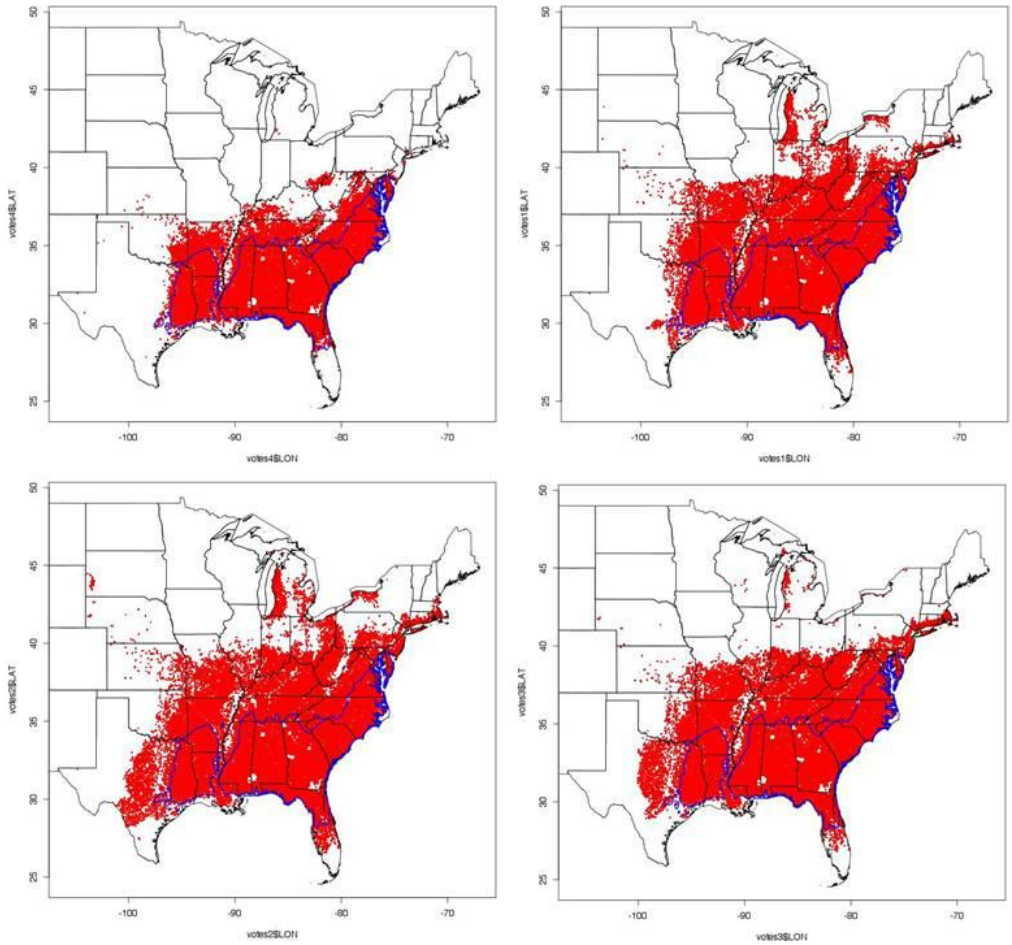


Figure A - 101. PA Loblolly pine GFDLCM21 B1 models (top-left), 2030 (top-right), 2060 (bottom-left), 2090 (bottom-right). Also referenced as **Error! Reference source not found.** in Results.

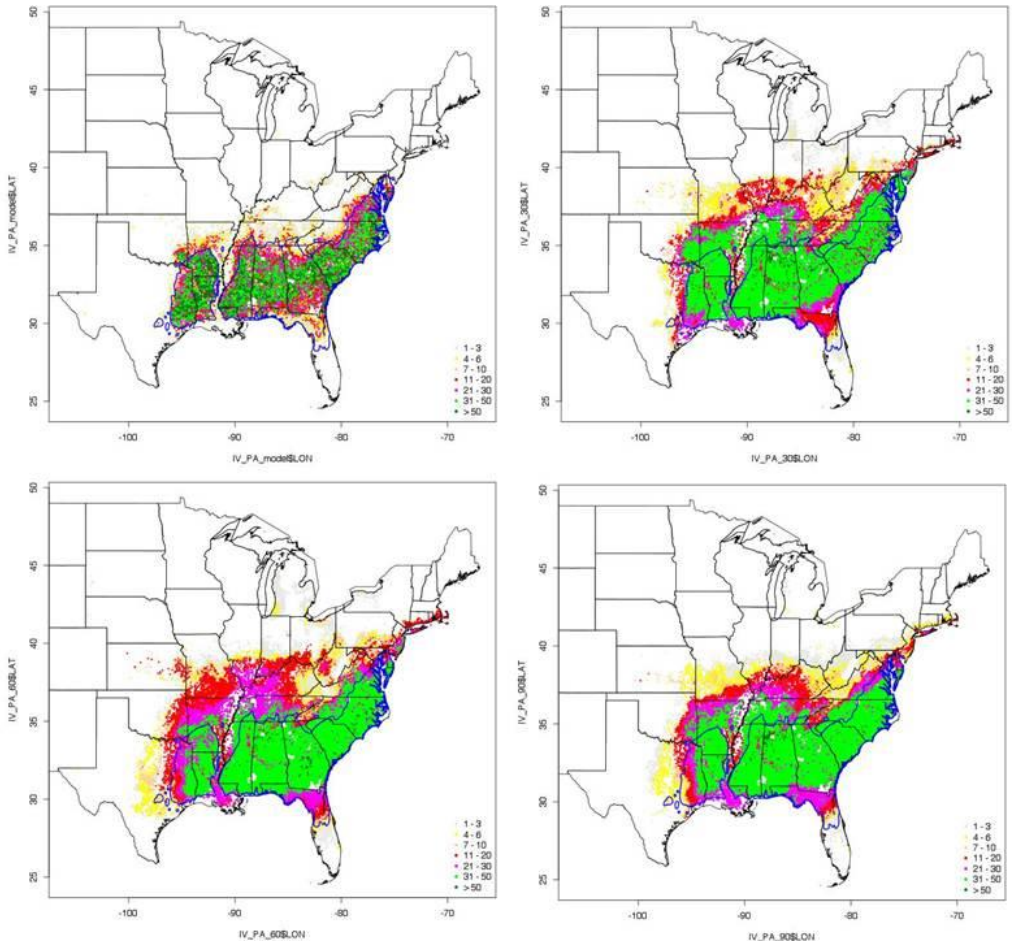


Figure A - 102. Combined IV and PA Loblolly pine GFDLCM21 B1 models (top-left), 2030 (top-right), 2060 (bottom-left), 2090 (bottom-right)

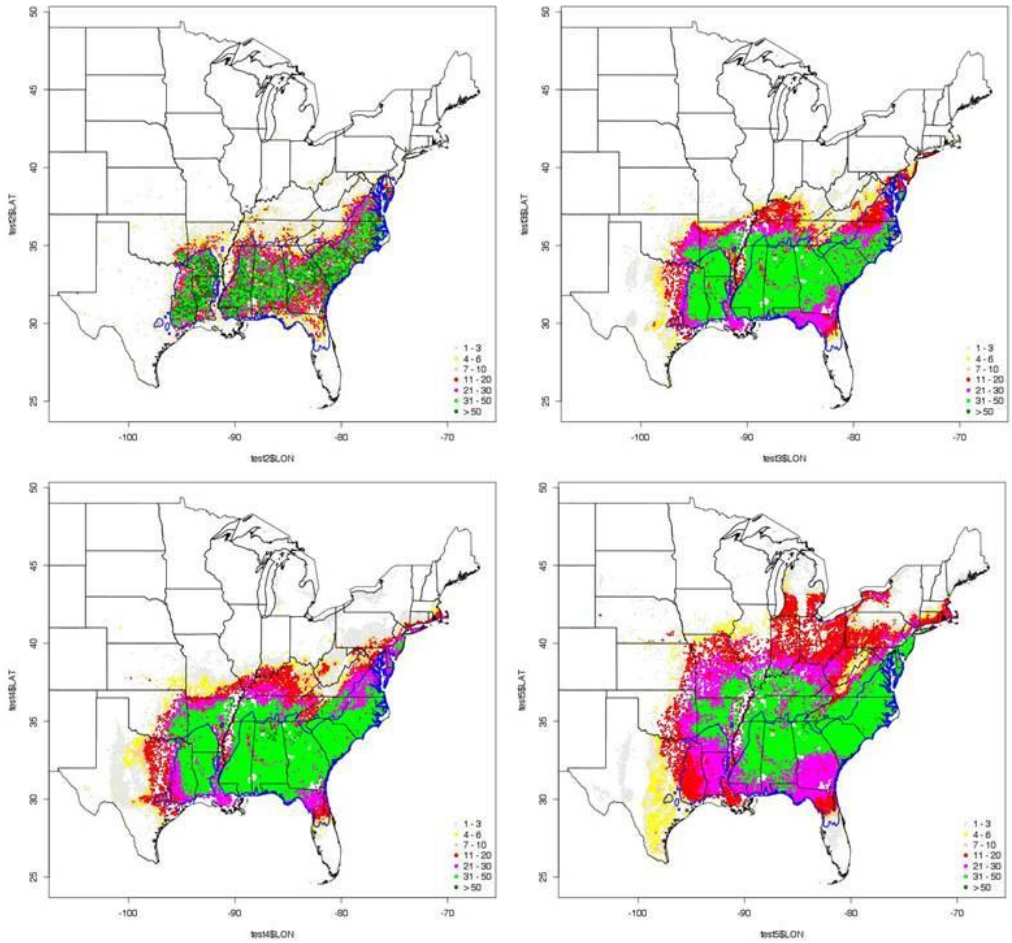


Figure A - 103. IV Loblolly pine HADCM3 A2 models (top-left), 2030 (top-right), 2060 (bottom-left), 2090 (bottom-right)

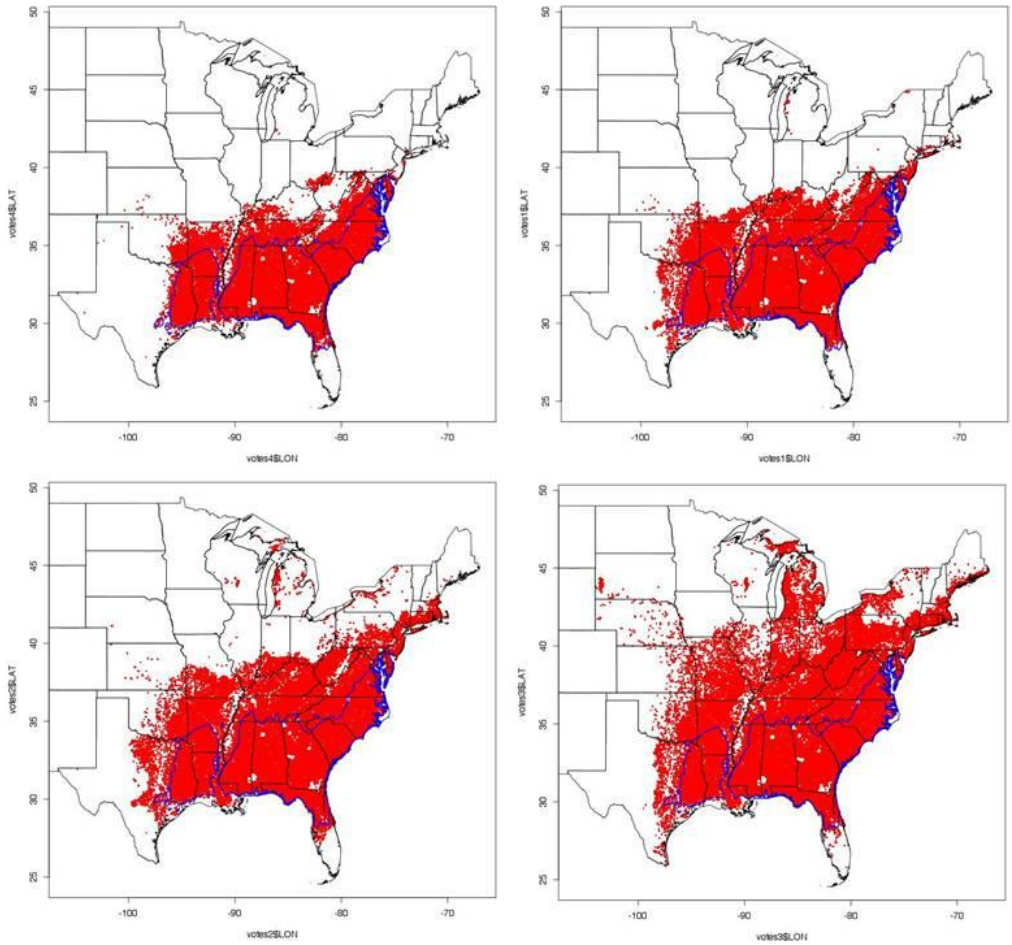


Figure A - 104. PA Loblolly pine HADCM3 A2 models (top-left), 2030 (top-right), 2060 (bottom-left), 2090 (bottom-right)

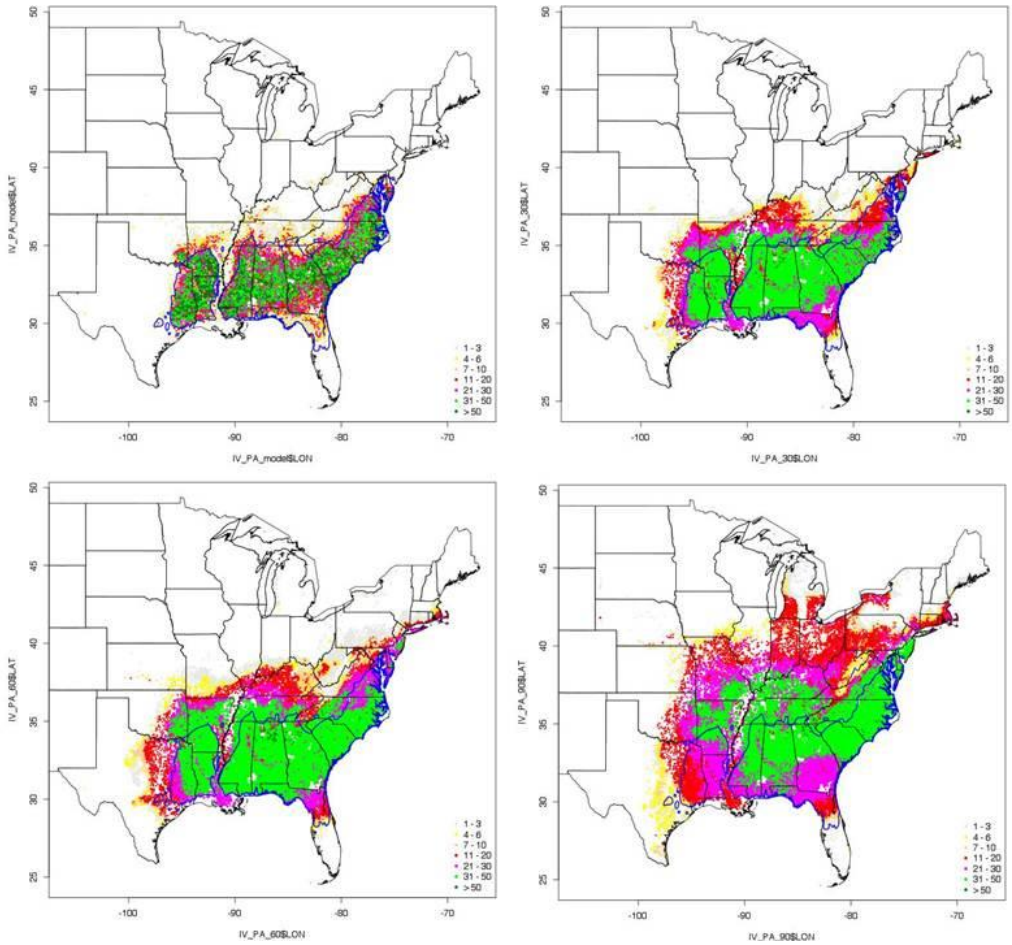


Figure A - 105. Combined IV and PA Loblolly pine HADCM3 A2 models (top-left), 2030 (top-right), 2060 (bottom-left), 2090 (bottom-right)

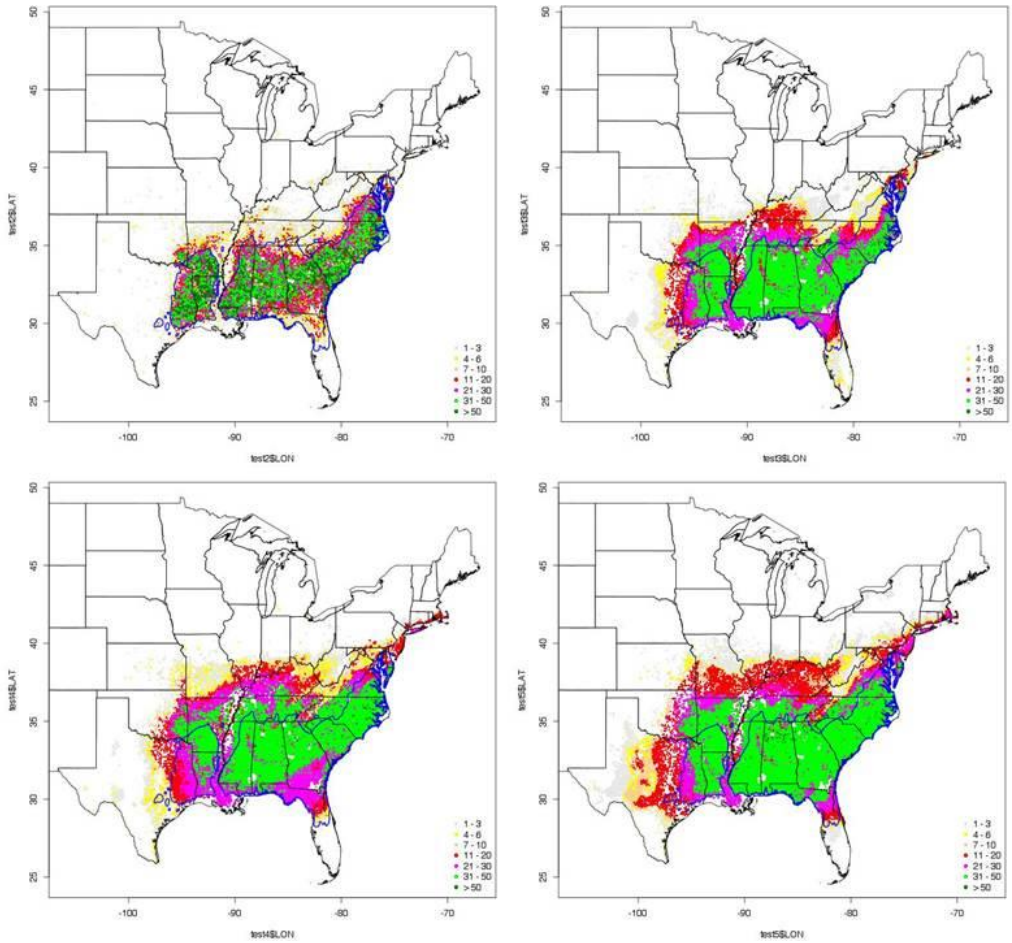


Figure A - 106. IV Loblolly pine HADCM3 B2 models (top-left), 2030 (top-right), 2060 (bottom-left), 2090 (bottom-right)

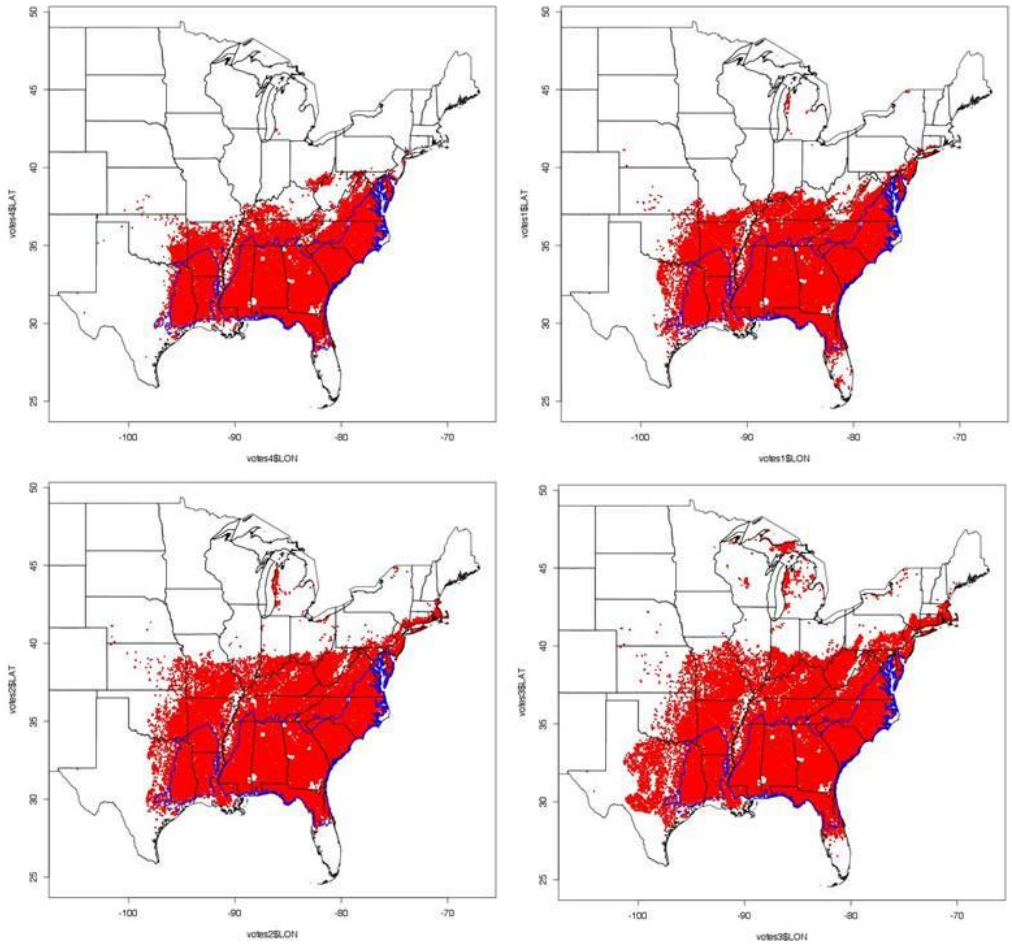


Figure A - 107. PA Loblolly pine HADCM3 B2 models (top-left), 2030 (top-right), 2060 (bottom-left), 2090 (bottom-right)

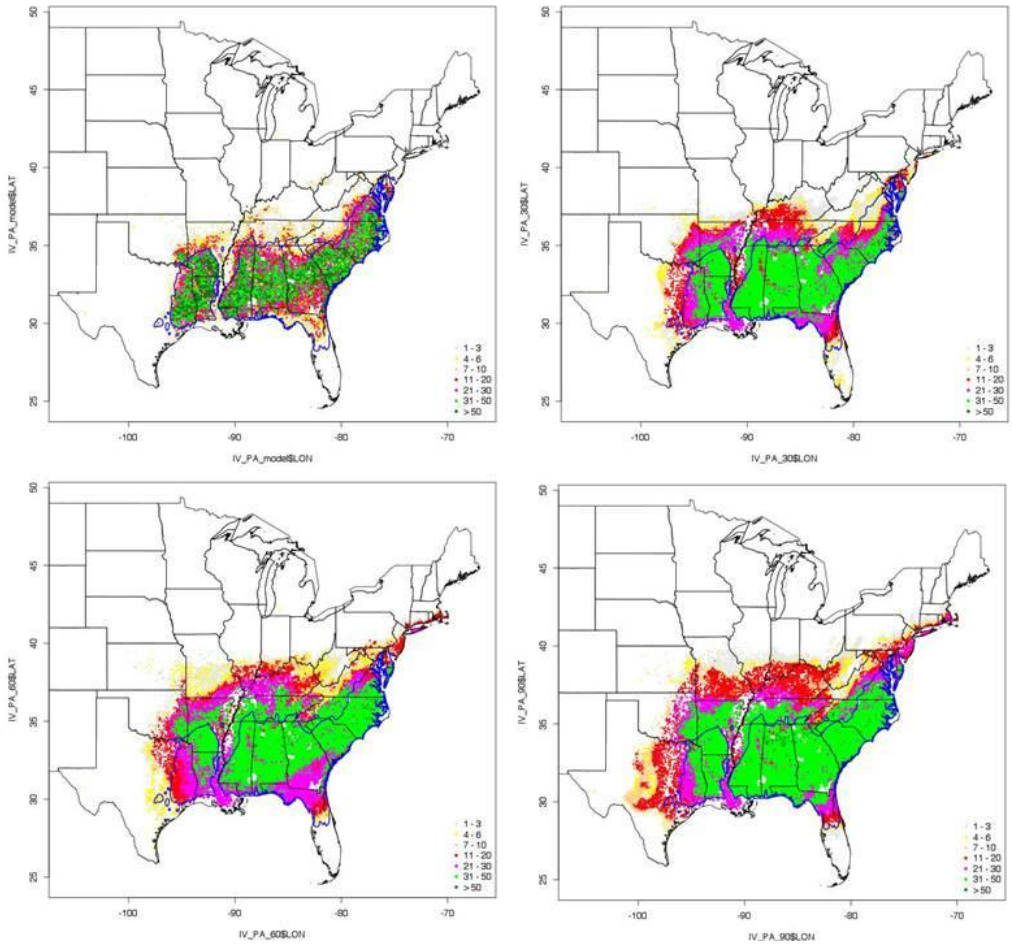


Figure A - 108. Combined IV and PA Loblolly pine HADCM3 B2 models (top-left), 2030 (top-right), 2060 (bottom-left), 2090 (bottom-right)

Figure A - 109. Loblolly pine PA vote count threshold movie

Figure A - 110. Longleaf pine PA vote count threshold movie

Figure A - 111. Pond pine PA vote count threshold movie

Figure A - 112. Shortleaf pine PA vote count threshold movie

Figure A - 113. Slash pine PA vote count threshold movie