Optimal Seed Deployment under Climate Change using Spatial Models: Application to Loblolly Pine in the Southeastern US

Alfredo Farjat¹ ², Brian J. Reich¹, Joseph Guinness¹, Ross Whetten², Steven McKeand², and Fikret Isik²

¹Department of Statistics, North Carolina State University
²Department of Forestry & Environmental Resources, North Carolina State University

Abstract

Provenance tests are a common tool in forestry designed to identify superior genotypes for planting at specific locations. The trials are replicated experiments established with seed from parent trees collected from different regions and grown at several locations. In this work a Bayesian spatial approach is developed for modeling the expected relative performance of seed sources using climate variables as predictors associated with the origin of seed source and the planting site. The proposed modeling technique accounts for the spatial dependence in the data and introduces a separable Matérn covariance structure that provides a flexible means to estimate effects associated with the origin and planting site locations. The statistical model was used to develop a quantitative tool for seed deployment aimed to identify the location of superior performing seed sources that could be suitable for a specific planting site under a given climate scenario. Cross-validation results indicate that the proposed spatial models provide superior predictive ability compared to multiple linear regression methods in unobserved locations. The general trend of performance predictions based on future climate scenarios suggest an optimal assisted migration of loblolly pine seed sources from southern and warmer regions to northern and colder areas in the southern USA.

Keywords: Spatial statistics, Bayesian statistics, provenance tests, seed source movements, assisted migration.
1 Introduction

There is evidence that the global climate is changing rapidly relative to historical trends [Stocker et al., 2014]. As a result, there is considerable interest in predicting the biological response of forest trees to changes in the environment. Forest trees in temperate regions are likely to experience accelerated rates of climate change in the coming century, relative to the average rate of change since the last glacial maximum [Mann et al., 2008]. The equatorial edge of the current range of forest tree species (e.g. the southern edge of the range of Northern Hemisphere species) may face the earliest deleterious impacts of this accelerated rate of change, while the polar edge may undergo range expansion, at least in terms of temperature adaptation.

Tree populations must respond to adverse changes in environmental conditions either by migrating to new locations where conditions are more favorable, or adapting to the new conditions in the current location [Aitken et al., 2008]. The changes in climate that took place during past ice ages were slow and some tree species were able to migrate as the glaciers advanced [Schmidtling et al., 2007]. However, the changes in climate that the earth is currently facing are likely to occur at a faster pace, and hence many tree species will be threatened.

In a climate change context, locally-adapted tree species will experience environmental conditions to which they are not well adapted [Skelly et al., 2007]. As a result, planting seedlings predicted to be adapted to future climate conditions has been proposed as a forest management strategy aimed to mitigate the potential negative impacts of global climate change [Ledig and Kitzmiller, 1992]. The risks of adaptation of future forest tree plantations can be mitigated with an assisted migration of suitable genetic material to ensure that plantations are adapted to future climate conditions [Aitken et al., 2008].
Loblolly pine (*Pinus taeda* L.) is a native pine species in the southeastern United States. It is the most hardy and versatile of all southern pines [Schultz, 1997] and is the major commercial species in the Southeast [McKeand et al., 2007]. It is used in a broad range of products, such as pulpwood, plywood, and construction lumber. Hence it is planted widely for commercial timber production and subjected to an intense breeding program. In addition to supply resources for human use, loblolly forests provide watershed protection and vital organic infrastructure for countless species in the region. Therefore, the species has vital commercial and environmental value to the southeastern United States.

It is very likely that climate change will have significant impacts on pine plantations across the southeastern United States. The Southeast is expected to experience an increase in the rate of warming through the end of the century, with a rise in average temperature of 2.5 to 5°C by the 2080s [Karl et al., 2009]. Precipitation predictions for the region are less certain, but the trend indicates a decline of 10% to 30% of summertime precipitation [Solomon, 2007]. The expected precipitation decrease combined with elevated temperatures will likely increase vapor pressure and soil water deficits causing stress to the loblolly plantations and a decrease in productivity of at least 10% [Schmidtling, 1994].

The establishment and analysis of provenance tests for investigating the genetic variation among forest trees has a long tradition in forestry [Langlet, 1971]. Provenance tests are common garden experiments generally intended to identify superior genotypes for planting at specific environments or across environments. Such tests are usually replicated experiments established with seed from parent trees collected from different geographical regions within the species distribution and grown at several locations (sometimes outside of the natural distribution). In addition to its traditional use, provenance tests provide meaningful information for assessing the response of tree populations to environmental change [Schmidtling, 1994, Matyas, 1994]. The geographic location from where the seed parents
were collected, the location of the test sites where the progeny trees were grown, and the
cclimate of these two sites can be used to make predictions about the relative performance
of tree populations under different environmental conditions.

Predictive models to assess the effect of climate variables on tree phenotypes using
provenance test data have been developed by different investigators. For instance, the pop-
ulation response function approach models the performance of a single population across
different planting sites in terms of the climate at those locations. This approach was used
by several researchers to predict the impact of climate change on a particular population
[Schmidtling, 1994, Matyas, 1994, Rehfeldt et al., 1999, Wang et al., 2006, O’Neill et al.,
2008]. Transfer functions using provenance test data have been also developed. A transfer
function gives the performance of multiple populations as a function of the climatic transfer
distance, which is the difference between the climate of the original location of the individ-
uals and the climate of the planting site [Matyas, 1994, Rehfeldt et al., 1999, Andalo et al.,
2005]. The Universal Response Function [Wang et al., 2010] approach integrates both the
population response and transfer functions into a single model. It makes use of all available
information from provenance tests to estimate a joint response of any population growing
in any climate [McLane et al., 2011, Kapeller et al., 2013, Farjat et al., 2015]. All of these
models are based on multiple linear regression analysis that ignore the spatial dependence
structure of provenance trials and utilize climate and geographic variables as predictors.
The other limitation of these approaches is that the joint response is constrained to a
parametric linear function that may fail to characterize the underlying process.

In this work a spatial Bayesian approach is developed for modeling the relative perform-
ance of seed sources of loblolly pine using climate and geographic variables as predictors
associated with the origin of the seed and the planting site. The underlying assumption
is that both genetic and environmental effects are captured through these variables. The
proposed modeling approach generalizes previous universal response function methodology by accounting for complex non-linear climate relationships and spatial dependence. Other authors have proposed regression approaches to model tree growth while accounting for the spatial dependence [Fox et al., 2007, Nathoo, 2010a,b].

We model the response function as a Gaussian process over climate variables and the spatial coordinates of the origin and planting sites. This poses two types of climate/spatial dependence: for a given planting site we would expect similar relative performance for trees from nearby seed sources; and for a given seed source, we would expect similar relative performance at two nearby planting sites. Also, since the goal of this study is to predict performance under different climate scenarios, ideally dependence between planting sites and seed sources would be explained by differences in climate variables. This motivates our covariance model which is a function of both spatial and climate variables from the planting site and seed sources. To account for all of these features in a computationally-feasible way, we employ a separable covariance function that is the product of the covariance between planting sites and the covariance between seed origins. Under the proposed framework, we compared models based on geographic distances, climate distances, and a combination of both geographic and climate distances.

The objectives of this work are to 1) develop a statistical model to predict the relative response of loblolly pine seed sources to climate change, 2) compare the predictive ability of the proposed model against multiple linear regression methods, and 3) use the model to create a quantitative tool for seed deployment aimed to identify the location of superior performing seed sources that are suitable for specific regions under a given climate scenario. To demonstrate the use and extent of the proposed model, application examples are presented and discussed concerning the predicted performance of loblolly pine seed sources for current and future climatic scenarios across the southeastern United States.
2 Description of the data

2.1 Genetic Material

The Cooperative Tree Improvement Program at North Carolina State University\(^1\) conducted an intensive selection effort in unimproved loblolly pine plantations between 1975 and 1981 aimed to increase the size of breeding populations. Trees were selected based on phenotypic attributes such as good growth, straight stems, and absence of fusiform rust galls (caused by the fungus *Cronartium quercuum f. sp. fusiforme*). In 1994, some of these plantation selections were used as parents to establish the Plantation Selection Seed Source Study (PSSSS), a large replicated series of provenance-progeny tests in the southeastern United States. The experiment was designed to evaluate the actual patterns of geographic genetic variation of plantation selections and to assess genotype by environment interactions [Farjat et al., 2016].

In early plantations of loblolly pine across the Southeast, it was not uncommon to observe seed movements up to 500 km from their original location. As a result, the exact origin of the selections used as parents in the PSSSS is not known. Nevertheless, based on the knowledge of foresters who established these plantations, we are confident that most of the sources of the seed are local or within a 100 to 200 km distance from the origin. Therefore, throughout this work, instead of the original geographic source of seed (provenance) we will refer to seed sources to indicate the locality and climate where a source of seed was collected.

The trees selected for the PSSSS were chosen from a pool of plantation selections identified in seven geographic regions within the natural range of loblolly pine in the southeastern United States.

\(^1\)http://www.treeimprovement.org
United States. Figure 1a shows the location of the pine plantations from where the female parents were selected. The plot displays the seven geographic regions of the natural range of loblolly pine defined by the Cooperative Tree Improvement Program of NCSU, namely: Virginia (VA), North Carolina Coastal Plain (NC), South Carolina Coastal Plain (SC), Georgia-Florida Coastal Plain (GF), Lower Gulf Coastal Plain (LG), Upper Gulf Coastal Plain (UG), and Piedmont (PD). These seven regions were delineated based on the observed geographic variation within the species as well as easily distinguishable state and physiographic boundaries. In each of seven regions, 20 selections were randomly chosen as female parents. These selections were then mated with a pollen mix of 40 pollen parents from the same region. As a result, a total of 140 maternal families were employed for the study, and a total of 280 male parents were used in the pollen mixes. The female and pollen parents were unrelated to avoid inbreeding depression in the progeny.

2.2 Experimental Design

Field trials were established using a randomized complete block design. The experiment was originally replicated at 25 locations across the southeastern United States, however only 16 sites covering a substantial area of the natural range of loblolly pine were analyzed in this work (Figure 1b). Four northern sites were not included in the analysis because many of the pine families present in the southern sites were not originally planted at these sites. In the other sites the corresponding eight-year measurements were never taken.

Each site had 24 blocks comprising a single tree from each family and four genetically unimproved local seedlings used as control (mixed seedlots) randomly distributed within blocks. Test sites were identically managed following standard progeny test protocols that included site preparation before planting and weed and vegetation control during the experiment. No fertilization treatment was applied at any site. The experiment was intended to
Figure 1: Geographic regions of loblolly pine grouped by counties, location from where the seed sources were collected (dots), test site locations (triangles), and natural range of the species. The geographic regions are Virginia (VA), North Carolina Coastal Plain (NC), South Carolina Coastal Plain (SC), Georgia-Florida Coastal Plain (GF), Lower Gulf Coastal Plain (LG), Upper Gulf Coastal Plain (UG), and Piedmont (PD).

be balanced, yet southern seed sources from South Carolina Coastal, Georgia and Florida Coastal, and the Lower Gulf were not planted in four northern sites because of space limitations and expected poor adaptability due to the low temperatures [Schmidtling, 2001, Farjat et al., 2016].

2.3 Standardized Height

Growth (height and diameter at breast height), stem straightness, fusiform rust disease incidence, forking, and survival of over 43,000 trees were assessed at age 8 years after planting. Although several traits were measured, the primary trait under consideration in this work is height because it is a heritable trait subjected to natural selection [McKeand et al., 2008] very likely to respond to changes in climate.
In order to quantify the performance of a specific seed source planted at a given site, the 8-year standardized height was obtained as follows: 

\[ Y(s, t) = \frac{\bar{X}(s, t) - \hat{\mu}_t}{\hat{\sigma}_t}, \]

where \( \bar{X}(s, t) \) represents the mean height of trees from the seed source \( s \) planted at test site \( t \), \( \hat{\mu}_t \) denotes the sample mean across seed sources obtained at site \( t \), and \( \hat{\sigma}_t \) is the sample standard deviation of seed source means from site \( t \). For a given test site, \( Y(s, t) \) measures the performance of a specific seed source relative to the others present in the same site. The standardized height was used as the response variable for the proposed model instead of the absolute height because the goal was to identify superior genotypes while accounting for environmental factors. Standardization removes site effects (environment) that may have influenced the growth and forces the response at different sites to have zero mean and variance equal to one. The absolute and standardized heights as a function of climate variables are displayed in the Supplementary Materials, section 1.

2.4 Climate Variables

The historical climate variables associated with planting site and seed source were estimated using the Parameter-elevation Regressions on Independent Slopes Model (PRISM) climate mapping system [Daly et al., 1994, 2008]. The time resolution of the PRISM mapping system was restricted to monthly estimates. Yearly average estimates for the climate variables were constructed from these monthly estimates.

Yearly average total precipitation (PPT, average of the yearly totals), yearly average minimum temperature (TMIN, average of the yearly minimums), and yearly average maximum temperature (TMAX, average of the yearly maximums) of both test site and seed source locations were used as explanatory variables to develop prediction models. The hypothesis underlying the development of the model is that the relative variation in phenotypes is influenced by climate and occurs geographically; therefore, this variation can
be predicted from climate descriptors. Throughout this work climate variables will be described by subscripts $s$ or $t$, denoting seed source and test site, respectively.

The period 1994 to 2006 was used to estimate the yearly average of the climate variables associated with the 16 test sites, whereas the period 1970 to 1990 was used for the seed source climate variables. The centroids of the counties from where the selections were collected were used as the reference location for the seed sources.

The climate variables used in this study are long-term averages with fairly small associated uncertainty (see Supplementary Materials, section 2). As a result, we acknowledge that some small attenuation bias might occur in the coefficient estimates but since our main goal is prediction, we do not pursue bias correction.

2.5 Climate Models

In order to assess future climate change impacts on loblolly pine performance, we employed projections of climate variables from 20 climate modeling groups coordinated by the Coupled Model Inter-Comparison Project 5 (CMIP5)\(^2\). The CMIP5 experiments include different trajectories of greenhouse gas concentration based on human actions called Representative Concentration Pathways (RCPs). These pathways are used in climate modeling to establish a range of changes in radiative forcing (additional energy trapped by the earth-atmosphere system relative to 1850s climate) from which possible consequences can be inferred.

We used the Multivariate Adaptive Constructed Analogs (MACA) [Abatzoglou and Brown, 2012] statistical downscaling method to obtain the climate model outputs from the 20 CMIP5 modeling groups across the southeastern United States. We used the historical

\(^2\)http://cmip-pcmdi.llnl.gov/cmip5/index.html
simulation for the period 1970-1990 as the baseline reference and considered the climate model outputs from the greenhouse concentration scenarios RCP 4.5 and RCP 8.5 for the period 2050-2099. RCP 4.5 [Clarke et al., 2007, Wise et al., 2009] refers to the scenario where an additional 4.5 \( W/m^2 \) is trapped in the earth-atmosphere system by 2100 compared to pre-industrial conditions and represents a moderate increase of temperature. On the other hand, RCP 8.5 [Riahi et al., 2007] refers to the scenario where an additional 8.5 \( W/m^2 \) is trapped by 2100 and is considered a worst-case scenario characterized by increasing greenhouse gas emissions over time. The second half of the 21st century was chosen as the future time period for the analysis because, based on climate projections, considerable changes in climate are not likely to occur in the near-term. Furthermore, from an operational and tree breeding perspective, it is very important to have a long-term plan to distribute genetic material that is better adapted to future climate.

At each location, the average across models (ensemble model) was used to estimate the projected and historical values for the climate variables. The calibration (adjustment to correct for climate model biases) between the observed and simulated historical and projected climate variables was done according to the following expression: \( \hat{X}_{FUT}(s) = X_{HIST}(s) + \Delta_{SIM}(s) \), where \( \hat{X}_{FUT}(s) \) is the projected future climate at location \( s \), \( X_{HIST}(s) \) is the observed historical climate at \( s \), and \( \Delta_{SIM}(s) \) represents the difference between the ensemble of models for future and historical simulated climate models at location \( s \), that is:

\[
\Delta_{SIM}(s) = \frac{1}{20} \sum_{j=1}^{20} \hat{X}_{FUT,j}(s) - \frac{1}{20} \sum_{j=1}^{20} \hat{X}_{HIST,j}(s)
\]

where \( \hat{X}_{FUT,j}(s) \) is the simulated future climate at \( s \) by the \( j \)th climate model and \( \hat{X}_{HIST,j}(s) \) is the simulated historical climate at \( s \) by the \( j \)th climate model, for \( j = 1, \ldots, 20 \).

11
3 Model formulation

3.1 Spatial seed source transport model

This section describes the statistical approach developed to model the relative performance of seed sources planted at different locations using geographical and climate variables as predictors. Let \( Y(s, t, u_s, u_t) \) be the standardized height for seeds from geographical location \( s \in \mathbb{R}^2 \) and planted at site \( t \in \mathbb{R}^2 \), with associated climate variables from site \( t \) and location \( s \) denoted as \( u_t = (u_{t1}, \ldots, u_{td})^T \in \mathbb{R}^d \) and \( u_s = (u_{s1}, \ldots, u_{sd})^T \in \mathbb{R}^d \), respectively. The standardized height is defined at every point over the region of interest \( D \subset \mathbb{R}^{d+2} \times \mathbb{R}^{d+2} \), and modeled as a random process \( Y(\cdot) = \{Y(s, t, u_s, u_t) : s, u_s \in \mathbb{R}^{d+2}; t, u_t \in \mathbb{R}^{d+2}\} \).

The proposed statistical model is inspired by classical geostatistics [Cressie, 1993] and takes the form:

\[
Y(s, t, u_s, u_t) = \mu(u_s, u_t) + e(s, t, u_s, u_t) \quad (2)
\]

where \( \mu(u_s, u_t) = E[Y(s, t, u_s, u_t)] \) is the mean function assumed to be deterministic, and \( e(s, t, u_s, u_t) \) is the mean-zero error. The mean term captures the general trend or large-scale spatial variations of the process under study using only climate variables. Geographical variables were not included in the mean function because the proposed model is aimed to identify optimal planting locations for a given seed source based on climatic conditions rather than geographical locations. We considered a linear mean model:

\[
\mu(u_s, u_t) = x(u_s, u_t)^T \beta \quad (3)
\]

where \( x(u_s, u_t) \) is a \( p \)-vector of covariates formed by the climate variables and \( \beta \) is a \( p \)-vector of unknown regression parameters to be estimated.

The second term in (2) accounts for the spatial dependence and is a function of both
geographic locations and climate variables. The error process \( e(\cdot) \) also accounts for the presence of an inherent measurement error in the observations. The measurement error is considered as an independent process without any spatial dependence. We decompose the error process into two additive contributions:

\[
e(s, t, u_s, u_t) = \eta(s, t, u_s, u_t) + \varepsilon(s, t, u_s, u_t)
\]

where \( \eta(s, t, u_s, u_t) \) is the spatially-dependent process and \( \varepsilon(s, t, u_s, u_t) \) is the measurement error process. The inclusion of the spatially-dependent process is based on the assumption that for a given planting site, performance of seed from neighboring origins would be similar; and conversely, performance of a given seed source at two neighboring sites would be also similar. Furthermore, dependence between planting sites and seed sources would be explained by differences in climate variables.

The spatial covariance function of \( \eta(s, t, u_s, u_t) \) is assumed to be separable, i.e., the product of the covariance for the planting sites, \( C_t \), and the covariance for the seed sources, \( C_s \). The implications of the separability assumption are that the covariance across seed sources is the same for all planting sites, and the covariance across planting sites is the same for all seed sources. This separation is mathematically convenient and computationally advantageous because it allows the full covariance matrix to be written as the Kronecker product of two covariance matrices, which facilitates algebraic manipulations and the use of numerical methods to speed up the model fitting algorithm.

We considered a zero-mean Gaussian process for the spatially-dependent process \( \eta(\cdot) \)

---

3: seed source location, \( u_s \): climate variables from location \( s \)
4: test site location, \( u_t \): climate variables from location \( t \)
with covariance function:

\[
\text{cov}[\eta(s, t, u_s, u_t), \eta(s', t', u'_s, u'_t)] = \tau^2 C_s(r_{ss'}, \theta_s) C_t(r_{tt'}, \theta_t)
\]  

where \(\tau^2\) is the spatial variance, and \(\theta_s\) and \(\theta_t\) are the covariance parameters for seed sources and test sites, respectively; \(r_{ss'}\) and \(r_{tt'}\) denote generalized distances for seed sources (\(s\) and \(s'\)) and test sites (\(t\) and \(t'\)), respectively. These generalized distances were defined to include the climate variables for seed source and test site as follows:

\[
r_{ss'} = r_{ss'}(s, s', u_s, u'_s) = \sqrt{\left(\frac{||s - s'||}{\varphi_s}\right)^2 + \sum_{i=1}^{d} \left(\frac{u_{s_i} - u'_{s_i}}{\varphi_{u_{s_i}}}\right)^2}
\]

\[
r_{tt'} = r_{tt'}(t, t', u_t, u'_t) = \sqrt{\left(\frac{||t - t'||}{\varphi_t}\right)^2 + \sum_{i=1}^{d} \left(\frac{u_{t_i} - u'_{t_i}}{\varphi_{u_{t_i}}}\right)^2}
\]

where \(\varphi_s\) and \(\varphi_t\) are unknown scale parameters associated with the seed source and test site geographical locations; \(\varphi_{u_{s_i}}\) and \(\varphi_{u_{t_i}}\) for \(i = 1, \ldots, d\) are unknown scale parameters for the climate variables observed at the location where seeds were collected and at the sites.

The spatially dependent error process \(\eta(s, t, u_s, t_t)\) is random but continuous in \(s, t, u_s,\) and \(u_t\). Valid isotropic covariance functions \(C_s\) and \(C_t\) were generated from the Matérn parametric family of correlation functions [Matérn, 1960, 1986]. For instance, the correlation between observations from sites \(t\) and \(t'\) with generalized distance as in (7), is given by:

\[
C_t(r_{tt'}, \theta_t) = \frac{2^{1-\kappa_t}}{\Gamma(\kappa_t)} (r_{tt'})^{\kappa_t} K_{\kappa_t}(r_{tt'})
\]

where \(\theta_t = (\varphi_t, \varphi_{u_{t_1}}, \ldots, \varphi_{u_{t_d}}, \kappa_t)\) groups the scale and smoothness parameters associated with the geographic and climate variables for the test sites, \(K_{\kappa_t}\) is a modified Bessel function.
of the second kind of order $\kappa_t$, $\Gamma(\cdot)$ denotes the gamma function, and $\kappa_t$ is the smoothness parameter for the sites, which directly controls the smoothness of the random field. The scale parameters that control the range of the correlation are in the generalized distance $r_{tt'}$. Similarly for the seed sources, the correlation function between observations from $s$ and $s'$ is denoted as $C_{s}(r_{ss'}, \theta_s)$, where $r_{ss'}$ is the generalized distance defined in (6), and $\theta_s = (\varphi_s; \varphi_{u_{s1}}, \ldots; \varphi_{u_{sd}}, \kappa_s)$ is the vector of scale and smoothness parameters associated with the geographic and climate variables for the seed sources.

### 3.2 Spatial and Climate Predictions

The proposed parametric model (2) can be used to provide spatially explicit estimates of seed source performance across the Southeast for current and future climates. Genotype by environment interactions are assumed to be fairly small, which indeed tend to be minor for loblolly pine [McKeand et al., 2006]. For instance, to estimate the performance of a given seed source across the region of interest under current climate, we hold constant the variables associated with the seed, and set the planting site climate variables equal to current climate values. Similarly, to predict the performance under projected future climate values, the planting site climate variables are replaced by the site projected future climate values. Hence, the performance of seed sources at a given planting site can be compared for current and projected future climates, leading to optimal planting strategies.

The model is fit using current climate data and then predictions are based on projected future climate values. The underlying assumption is that the model estimated based on current climate data also holds over the future climate scenarios that we trying to extrapolate over.
4 Bayesian Formulation

Let $Y$ denote the vector of standardized heights ordered as sources within sites of dimension $n = n_s n_t$, where $n_s$ is the number of seed sources and $n_t$ is the number of test sites. Let $X$ represent the $n \times p$ design matrix of covariates created from climate variables observed at the test sites and at the seed sources. Let $\Sigma = \Sigma(\theta)$ denote the variance-covariance matrix of the response vector, where $\theta$ is a $q$-vector of unknown parameters that belong to a parameter space $\Theta \subset \mathbb{R}^q$, within which the covariance function is valid. Since the error process is Gaussian, the joint distribution of $Y|\beta, \theta$ is multivariate normal with mean $X\beta$ and covariance matrix $\Sigma(\theta)$, denoted hereafter as: $Y|\beta, \theta \sim N(X\beta, \Sigma(\theta))$. The spatial covariance matrix can be written as follows:

$$\Sigma(\theta) = \tau^2 \Gamma_t(\theta_t) \otimes \Gamma_s(\theta_s) + \sigma^2 I$$

where $\tau^2$ denotes the spatial variance, $\Gamma_t(\theta_t) \otimes \Gamma_s(\theta_s)$ is the spatial correlation matrix expressed as a Kronecker product of the test site and seed source correlation matrices, and $\sigma^2$ is the measurement error. All the unknown parameters that define the spatial covariance matrix are grouped in the vector $\theta = (\tau^2, \sigma^2, \theta_s^T, \theta_t^T)^T$, where $\theta_s$ and $\theta_t$ contain the scale and smoothness parameters from seed source and test site, respectively.

The Bayesian model is completed by assigning a prior distribution to all parameters. The regression coefficients $\beta$ was assigned a weakly-informative prior: $\beta \sim N(0, \Sigma_\beta)$, where $\Sigma_\beta = \sigma_\beta^2 I$. The hyper-parameter $\sigma_\beta^2$, that defines the prior knowledge about the coefficients was chosen corresponding to a fairly vague prior ($\sigma_\beta^2 = 5000$). The parameters from the covariance matrix $\theta$ were assigned a log-normal prior to enforce their positiveness, that is $\log \theta_i \sim N(\mu_{\theta_i}, \sigma_{\theta_i}^2)$, where the location $\mu_{\theta_i}$ and scale $\sigma_{\theta_i}^2$ hyper-parameters were chosen to obtain fairly vague priors ($\mu_{\theta_i} = 0$ and $\sigma_{\theta_i}^2 = 3$), except for the smoothness parameters $\kappa_s$.
and $\kappa_t$ for which the priors were somewhat informative around the exponential covariance $(\mu_{\kappa_{s,t}} = \log 0.5$ and $\sigma^2_{\kappa_{s,t}} = 0.15)$. These hyper-parameters were chosen relatively centered around the exponential covariance to simplify the analysis and to avoid identification issues associated with these parameters [Gelfand et al., 2010].

The response vector of standardized heights $Y$ had missing values that impeded the necessary calculations of the assumed balanced configuration. Thus, to account for the missing values, an iterative data augmentation approach was used in the Gibbs sampler [Tanner and Wong, 1987, Li, 1988] (see Supplementary Materials, section 5).

### 4.1 Bayesian Computation

Monte Carlo methods [Carlin and Louis, 2011, Robert and Casella, 2013] were used for estimation and inference of the model parameters. Specifically, a Gibbs sampler [Gelfand and Smith, 1990] was developed to draw samples from the posterior distribution of the model parameters given the data. The Metropolis algorithm [Metropolis et al., 1953] was used to sample from the full conditional distributions in cases where the prior and the likelihood did not form a conjugate pair, and hence, a closed form was not available. The regression coefficients $\beta$ were sampled from the closed form full conditional $p(\beta|\cdot) = N(\tilde{\mu}_\beta, \tilde{\Sigma}_\beta)$, where $\tilde{\Sigma}_\beta = (X^T \Sigma^{-1} X)^{-1}$, and $\tilde{\mu}_\beta = \tilde{\Sigma}_\beta X^T \Sigma^{-1} Y$. Samples from the full conditional distribution $p(\theta|\cdot)$ were obtained from the Metropolis algorithm using a multivariate normal $N(\theta^{(t-1)}, \tilde{\Sigma}_\theta)$ proposal density, where $\theta^{(t-1)}$ denotes the previous sample and $\tilde{\Sigma}_\theta$ is the proposal covariance matrix. The choice of $\tilde{\Sigma}_\theta$ was thoroughly explored to accelerate the convergence of the MCMC chains. To avoid the issue of posterior correlations among the elements of $\theta$, the proposal covariance-matrix was set to be diagonal, that is $\tilde{\Sigma}_\theta = \text{diag}(\tilde{\sigma}^2_{\theta_1}, \ldots, \tilde{\sigma}^2_{\theta_q})$, and its elements adaptively chosen during the burn-in period following a pilot adaptation approach [Gilks et al., 1998]. The convergence of the MCMC
The value of the process at the unobserved point \( x = (s, t, u_s, u_t) \in D \) was predicted using:

\[
\hat{Y}(x) = \mathbb{E}[Y(x)|Y(x_1) = z_1, \ldots, Y(x_n) = z_n] = \mathbb{E}[Y|Z],
\]

that is the conditional expectation given the observations, where \( Z = (z_1, \ldots, z_n)^T \) denotes the observed values of process \( Y \) at the points \( x_1, \ldots, x_n \in D \). This estimator is optimal in the sense that minimizes the expected squared prediction error [Ferguson, 1967].

The posterior predictive distribution has density that can be expressed as:

\[
P(Y|Z) = \int P(Y,Z|\beta, \theta)P(\beta, \theta|Z)d\beta d\theta,
\]

where \( Y \) is the vector of unobserved values, \( Z \) is the vector of observed values, and \( \beta \) and \( \theta \) are the model parameters. Assuming a joint multivariate normal distribution of the form

\[
\begin{pmatrix}
Y \\
Z
\end{pmatrix}
\sim
N
\begin{pmatrix}
\mu_y \\
\mu_z
\end{pmatrix},
\begin{pmatrix}
\Sigma_{yy} & \Sigma_{yz} \\
\Sigma_{zy} & \Sigma_{zz}
\end{pmatrix}
\]

the full conditional distribution of the unobserved values is:

\[
Y|Z, \beta, \theta \sim N \left( \mu_y + \Sigma_{yz} \Sigma_{zz}^{-1} (Z - \mu_z), \Sigma_{yy} - \Sigma_{yz} \Sigma_{zz}^{-1} \Sigma_{zy} \right).
\]  

Predictions were obtained using two approaches. The first method is the standard approach for Bayesian prediction which consisted of generating MCMC samples from \( Y|Z \) to make predictions from the posterior mean and to form predictions intervals. This approach allows to assess the predictive variability of the process by averaging over uncertainty in model parameters. However, this can be time consuming, particularly if the performance of several origins needs to be calculated. The second method is aimed to approximate the
outcomes from the first approach and to reduce the computational burden. To this end, the model parameters ($\beta$ and $\theta$) are first estimated using the sample mean of the MCMC chains generated from their posterior distributions. Then, the posterior means are evaluated in the Gaussian analytical expression of $Y|Z, \hat{\beta}, \hat{\theta}$ (Equation 10) to obtain the predicted values $\hat{Y} = \mathbb{E}(Y|Z, \hat{\beta}, \hat{\theta})$ and their estimated covariance matrix $\hat{\Sigma} = \text{Var}(Y|Z, \hat{\beta}, \hat{\theta})$. We acknowledge that a consequence of this plug-in approach is that the predictive variability could be underestimated, however for our application the obtained predicted values are similar to the fully Bayesian prediction method with the advantage that the computational cost is greatly reduced. A detailed comparison between the above described prediction methods is given in the Supplementary Materials, section 6.

The approximate (plug-in) method was used only for creating the arrow plots shown in section 5.3. These plots are meant to indicate the direction of optimal seed movement based on the performance of multiple origins at a given location, for which case the fully Bayesian method would be impractical due to the computational cost and processing time. Furthermore, the potential underestimation of the predictive variability in creating these plots does not represent a limitation because the performance of the origins is assessed only through their predicted values.

5 Results

5.1 Model Comparisons

The proposed formulation offers the flexibility to model the spatial dependence of the data in terms of combinations of geographic and climate variables related to test sites and seed sources. Four models were derived from the proposed framework and compared against each
other through cross-validation. The explored models are all special cases of the model given in Section 3 with identical predictors in their mean function but with different covariance functions.

1. The first model (IID) represents the benchmark from which the models were assessed. For this model the observations were assumed to be independent with the same measurement error variance. This non-spatial model is analogous to a Bayesian multiple linear regression (i.e., $\tau^2 = 0$).

2. The second model (GEO) assumes that the correlation between observations is a function of solely geographic variables and does not consider any climate variable in the spatial covariance function (that is, $\varphi_{u_i} \to \infty$ and $\varphi_{u,t_i} \to \infty$ for $i = 1, \ldots, d$, in equations (6) and (7)).

3. The third model (CLIM) ignores the geographic variables in the spatial covariance and uses climate variables only (that is, $\varphi_s \to \infty$ and $\varphi_t \to \infty$, in equations (6) and (7)). The climate variables used for this correlation model were minimum temperature and precipitation.

4. The fourth model (GEO/CLIM) is the full model (2) that combines both the geographic and climate variables. So, the correlations between the observed values are expressed in terms of the generalized distances of equations (6) and (7). The climate variables utilized to define the generalized distances were minimum temperature and precipitation.

The prior distributions were the same for all four models. We considered different values for the prior distribution of the regression coefficients $\beta$ and posteriors were not sensitive to these choices. The hyper-parameters for the covariance parameter $\theta$ represented fairly
vague priors (except for the smoothness parameters $\kappa_s$ and $\kappa_t$) and posteriors were not very sensitive to these choices either (see Supplementary Materials, section 8).

The models were compared using $K$-fold cross-validation. The data were partitioned into $K$ equally-sized subsamples. Each of the $K$ subsamples was used exactly once as the validation set for testing the model, while the remaining $K - 1$ subsamples were treated as the training data for fitting the model. Then, the average of the mean squared errors (MSE) from the $K$ folds was used as a measure of the predictive ability of the model. The sample was partitioned in two different ways for conducting the cross-validation. To assess the predictive ability of the model based on unobserved sites, the data were partitioned using the test sites as the folds. This type of cross-validation will be referred as leave-site-out, and is arguably the most relevant given our goal of predicting performance under different climate scenarios for planting sites. In addition, to quantify the capacity of the models to predict the performance of unobserved seed sources, the data were randomly split by grouping the seed sources into 19 folds of 5 sources each. This type of cross-validation will be referred as leave-source-out.

The cross-validation results are presented in Figure 2. The box-plots summarize the distribution of the MSEs for the explored models and for the two types of partitions of the data between training and validation sets. The leave-site-out cross-validation (left) shows that the spatial models (GEO, CLIM, and GEO/CLIM) provide a superior predictive ability than the non-spatial IID model in unobserved locations. On the other hand, the relative predictive powers of the three spatial models are essentially the same. The climate variables used as covariates in the models (yearly averages of temperatures and precipitation) vary smoothly over space with strong North/South and East/West gradients. Reich et al. [2006] and Hodges and Reich [2010] discuss how adding spatial covariates that are smoothly varying over space do not help with spatial prediction. Therefore it is not surprising
that using either geographic or climate variables would produce similar predictions. The MSEs from the leave-source-out cross-validation (right) are very similar to each other for all four models. Although the range of MSE values is smaller for the spatial models, overall these models provide a comparable predictive ability to the standard IID model in the leave-source-out cross-validation. This result is indeed surprising. It may be that the standardization of the data, which removes site effects and forces the response to be comparable across sites, attenuates the spatial dependence among sites and makes it difficult to interpolate across seed sources.

Based on the cross-validation analysis results, we decided to select the CLIM model for making predictions about the relative performance of seed sources under current and
future climate conditions. Although the predictions are similar to the GEO model in the current climate, we believe that the CLIM model is more scientifically justifiable and will thus give more reliable predictions under different climate regimes.

The separability of the covariance function was a major assumption that led to important computational savings. Separability holds if and only if the correlation across sites is the same for all origins and the correlation across origins is the same for all sites. This characteristic was assessed using a Bayesian spatial analysis separate by site and found that the chosen covariance model was suitable for our data set (see Supplementary Materials, section 4).

5.2 Summarizing the posterior of the selected model

The posterior distributions for the CLIM parameters are summarized in Table 1. The covariates that define the mean were chosen from a pool of variables using the Akaike Information Criterion [Akaike, 1974] in a stepwise algorithm. Although ignoring spatial correlation can influence variable selection results [Hoeting et al., 2006], in order to facilitate targeted comparisons against the IID model (benchmark), and the more complex spatial models, we chose to use the same covariates throughout the tested models.

The error term includes eight parameters, namely the measurement error variance ($\sigma^2$), the spatial variance ($\tau^2$), the site and seed source scale parameters for precipitation and minimum temperature ($\varphi_{P_t}, \varphi_{P_s}, \varphi_{T_{t}}, \varphi_{T_{s}}$), and the smoothness parameters ($\kappa_{s}, \kappa_{t}$). The results suggest that climate variables $T\text{MIN}_{t}$, $T\text{MIN}_{s}$, $T\text{MIN}_{T_{t}}$, $P\text{PT}_{s}$, and $T\text{MIN}_{s}P\text{PT}_{s}$ have a significant effect on the relative performance of loblolly pine seed sources. Also, it can be inferred that there is spatial dependence between observations given the magnitude of the error term parameter estimates. The scale parameter for the site precipitation $\varphi_{P_t}$ is the exception with a extremely large posterior mean value that eliminates site precipitation
from the spatial correlation function. The measurement error variance ($\sigma^2$) is slightly larger than the variance of the spatial process ($\tau^2$), indicating that there is considerable inherent randomness in tree height and potential local variation due to unmeasured factors such as management practices (e.g., tree measurements were taken in the fields by different crews).

The posteriors of the scale parameters for the minimum temperature ($\varphi_{T_i}, \varphi_{T_s}$) indicate a stronger spatial dependence between observations based on minimum temperature differences for the sites than for the seed sources. The estimated posterior mean of the smoothness parameter for the seed source ($\kappa_s$) is close to $\frac{1}{2}$ (i.e. the exponential correlation) and the posterior mean of the smoothness parameter for the site ($\kappa_t$) is greater than 1, suggesting that the process associated with the site is smoother than for the seed source.

The MCMC samples from the posterior distributions of the CLIM model parameters were summarized using trace plots, histograms, auto-correlation functions (ACF), and Geweke’s convergence diagnostic (see Supplementary Materials, section 7). Visual inspection of the trace plots revealed that the chains moved successfully throughout the parameter space. The Geweke’s diagnostic suggested that convergence of the MCMC chains to the stationary distribution was reached after about 20,000 samples.

5.3 Projections and Seed Source Movements

In this section we present examples to illustrate the capacity of the proposed model to make predictions about the performance of specific seed sources under current and future climate. In addition, examples of optimal seed source movements to specific sites are presented.

An exploratory analysis of the climate variables indicates that there is shift in the projected future climate values relative to current climate data (see Supplementary Materials, section 3). TMIN and PPT exhibit a moderate overall mean increase with a large
Table 1: Posterior summaries for the parameters of CLIM model. The mean coefficients \( (\beta) \) are represented by capital letters for their corresponding covariates, and the covariance parameters \( (\theta) \) are denoted with greek letters.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>MEAN</th>
<th>SD</th>
<th>PERCENTILE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2.5%</td>
<td>50%</td>
<td>97.5%</td>
</tr>
<tr>
<td>INTERCEPT</td>
<td>-0.007</td>
<td>0.056</td>
<td>-0.117 -0.007 0.102</td>
</tr>
<tr>
<td>TMIN ( t )</td>
<td>0.163</td>
<td>0.039</td>
<td>0.086 0.164 0.240</td>
</tr>
<tr>
<td>TMIN ( s )</td>
<td>2.413</td>
<td>0.563</td>
<td>1.307 2.415 3.514</td>
</tr>
<tr>
<td>TMIN ( s ) TMIN ( t )</td>
<td>0.260</td>
<td>0.033</td>
<td>0.195 0.260 0.325</td>
</tr>
<tr>
<td>TMAX ( t )</td>
<td>0.011</td>
<td>0.029</td>
<td>-0.046 0.011 0.069</td>
</tr>
<tr>
<td>TMAX ( s )</td>
<td>0.124</td>
<td>0.087</td>
<td>-0.048 0.124 0.294</td>
</tr>
<tr>
<td>PPT ( t )</td>
<td>0.000</td>
<td>0.020</td>
<td>-0.038 0.000 0.039</td>
</tr>
<tr>
<td>PPT ( s )</td>
<td>-0.278</td>
<td>0.076</td>
<td>-0.427 -0.278 -0.130</td>
</tr>
<tr>
<td>TMIN ( s ) PPT ( s )</td>
<td>-2.056</td>
<td>0.542</td>
<td>-3.116 -2.058 -0.990</td>
</tr>
<tr>
<td>( \sigma^2 )</td>
<td>0.410</td>
<td>0.018</td>
<td>0.377 0.410 0.446</td>
</tr>
<tr>
<td>( \tau^2 )</td>
<td>0.296</td>
<td>0.044</td>
<td>0.221 0.293 0.392</td>
</tr>
<tr>
<td>( \varphi_P ) ( t )</td>
<td>2481</td>
<td>2122</td>
<td>679 1884 7845</td>
</tr>
<tr>
<td>( \varphi_P ) ( s )</td>
<td>0.690</td>
<td>2.270</td>
<td>0.017 0.254 3.341</td>
</tr>
<tr>
<td>( \varphi_T ) ( t )</td>
<td>6.508</td>
<td>2.774</td>
<td>3.071 5.862 13.741</td>
</tr>
<tr>
<td>( \varphi_T ) ( s )</td>
<td>2.147</td>
<td>12.403</td>
<td>0.020 0.324 14.916</td>
</tr>
<tr>
<td>( \kappa ) ( t )</td>
<td>1.074</td>
<td>0.348</td>
<td>0.552 1.022 1.888</td>
</tr>
<tr>
<td>( \kappa ) ( s )</td>
<td>0.499</td>
<td>0.197</td>
<td>0.215 0.467 0.973</td>
</tr>
</tbody>
</table>

overlap between current and projected values. However, TMAX displays a considerable increase of projected values with a small overlap with current values. Therefore, we reiterate that projections for future climate rely on the assumption that extrapolating the relationship between climate and growth are valid.

Figures 3 and 4 show the predicted relative performance (standardized height) and associated standard deviations of seed sources from Nassau, FL and Appomattox, VA under current climate using the fully Bayesian prediction method. The climate scenario was created from historical data assuming that future climate conditions would be the same
as the average of the last 20 years. The prediction locations correspond to the centroids of
the counties across the Southeast which were then interpolated to facilitate visualization
of results. The predicted standardized heights indicate that seed sources form Nassau
(Figure 3a) will have a superior performance in the south, especially in Florida, and a poor
performance in the northern regions. Not surprisingly, the predicted relative performance
of seed sources from Appomattox (Figure 4a) exhibits a radically different pattern. In this
case, the spatial trend indicates that seeds from Appomattox will only be superior in the
most northern regions with a decreasing performance towards the south. These results are
in agreement with previous studies [Wells, 1983, Schmidtling, 2001, Farjat et al., 2016].
The standard deviations of the predictions (Figures 3b and 4b) show a pattern of smaller
uncertainty on a broad central region that include most of the observations and larger
standard deviations in northern and southern areas.

Figure 5 shows the predicted relative performance (standardized height) of all seed
sources from the southeastern US, planted in Sussex County (Virginia) for three climate
scenarios using the plug-in prediction method. The first scenario (Figure 5a) assumes that
future climate conditions would be the same as the average of the last 20 years (current
climate). From the plot, the location of the top performing seed sources can be identified.
The arrows show the direction of the optimal seed source movement based on the top
1% standardized heights. Seed source movements from specific locations from northern
Georgia, central South Carolina, and North Carolina are indicated as optimal. The other
two scenarios are based on Representative Concentration Pathways (RCPs) for greenhouse
concentration levels for the period 2050-2099. The moderate RCP 4.5 scenario (Figure 5b),
which assumes a stabilization of the CO₂ concentration level towards the end of the century,
suggests an optimal seed source movement from specific locations from central Georgia to
Sussex county. The third scenario is based on RCP 8.5 for which emissions are assumed to
Figure 3: a) Predicted performance (standardized height) for seed sources from Nassau, FL (black point) assuming that future climate conditions would be the same as the average of the last 20 years (current climate). The colormap indicates that seeds from Nassau county will have a superior performance in the southern region. b) Standard deviation for the predictions.

continue rising throughout the 21st century. This is the worst case scenario because it could result in important changes in future climate due to global warming. Under this scenario, the southern seed sources from central Georgia regions and central Florida improve their relative performance in Sussex county (Figure 5c).

Figure 6 portrays an extension to the previously presented performance map as a function of seed source for a given location by considering movements of seed to a group of counties instead of just one. The performance of the seed sources was estimated with the plug-in prediction method. The plot shows the optimal seed source movement to all the counties from southern Virginia based on the top 1% performing locations for the three
future climate scenarios above described, that is (1) current climate, (2) RCP 4.5, and (3) RCP 8.5. The arrows indicate the direction of the optimal seed movement and the dots on the map show the location of the top performing seeds. To facilitate visualization, the color of the arrows were chosen depending on the state of origin. The pattern under the most conservative scenario indicates that the Piedmont region of North Carolina, South Carolina, and Georgia as well as the Virginia region contain the top performing seed sources. The plot suggests that seed sources from the north Georgia are suitable genetic material for deployment to a broad region in Virginia. Also, movements of seed from the South Carolina Piedmont area to the Coastal Plains of Virginia are indicated. The RCP 4.5 sce-
Figure 5: Predicted performance in Sussex County, Virginia as a function of seed source for three future climate scenarios: a) future climate would be the same as the average of the last 20 years (current climate); and for two climate scenarios derived from Representative Concentration Pathways (RCPs) based on greenhouse concentration levels for the period 2050-2099: b) RCP 4.5 (moderate emission level), and c) RCP 8.5 (high emission level). The maps indicate the location of top performing seeds. The arrows show the direction of the optimal seed source movement based on the top 1% standardized height.

Scenario indicates that the Piedmont region of Georgia, the Upper Gulf region of Alabama as well as the Georgia-Florida Coastal Plain region contain the top performing seed sources. Movements of seed from central Florida to the Coastal Plains of Virginia are also indicated as optimal under this moderate scenario. In the most extreme climate scenario, seed movements predominately from the Georgia-Florida Coastal Plain region and also from Central Florida are indicated as optimal.
Figure 6: Optimal seed deployment to Virginia under future climate scenarios. The arrows show the direction of the optimal seed source movement to Southern Virginia based on the top 1% locations (dots) assuming 1) future climate conditions would be the same as the average of the last 20 years; and two future climate scenarios derived from Representative Concentration Pathways (RCPs) of greenhouse concentration levels for the period 2050-2099: 2) RCP 4.5 (moderate emission level), and 3) RCP 8.5 (high emission level). The dots on the maps indicate the location of the top performing seeds.

6 Discussions

In this paper we have developed a Bayesian spatial model to use data from a provenance test to predict the performance of different seed sources under climate change scenarios. We demonstrated that there is strong spatial dependence between both sites and seed sources, and that exploiting this fact results in better predictions than multiple linear regression analysis. We used the predictive model to map optimal seed source deployment, and develop graphical tools for visualization. The proposed statistical model can be used as a quantitative tool for designing forest management strategies oriented to mitigate the negative impacts of climate change. The model can be used as an interactive migration
tool aimed to identify relatively superior seed sources within the species distribution that will best allow trees to thrive under future climate conditions.

The proposed Bayesian spatial model exploits the spatial dependence between planting sites and seed sources from a provenance test to produce more accurate predictions than multiple linear regression analysis. The spatial dependence in the data is accounted for by fitting a non-parametric response surface and the separable Matérn covariance structure provides a flexible means to describe and estimate effects associated with the source of the seed and planting site locations.

The results indicate that the spatial dependence between observations cannot be ignored, and therefore predictions about the performance of different seed sources under future climate change scenarios can be considerably improved relative to traditional methods that constrain the response surface to a parametric linear function.

We found that the separability assumption of the covariance function was computationally convenient as well as adequate for the data set under study. Employing other approaches such as the linear model of coregionalization [Schmidt and Gelfand, 2003] would require new developments because rather than a finite number of variables (e.g. temperature and precipitation) we require a model for an infinite number of both sites and origins.

The proposed model assumes stationarity since the covariance between two locations of the region of interest depends only on their relative locations. We acknowledge that the data were collected over a large and diverse geographic domain so that non-stationarity is likely present here, as it is in almost any large-scale spatial analysis. However, with only 16 sites available it is very difficult to assess, let alone estimate, a non-stationary covariance function. The model with climate variables in addition to geographical variables (GEO/CLIM) could be viewed as an effort to address non-stationarity in the spatial-only model (GEO). But this model does not improve cross-validation prediction error, suggesting
that non-stationarity is not a major issue for these data. Furthermore, Parker et al. [2016] for example, show through simulation studies that predictions from a stationary model remain fairly efficient even in the presence of non-stationarity.

The model uses Bayesian analysis because it provides a convenient and flexible setting to estimate the model parameters and to deal with the problem of the missing data. MCMC methods make computation tractable for a wide range of parametric models and inferences about the model parameters are exact and conditional on the data, independently of the sample size and without using asymptotic approximations. Implementing a Bayesian analysis requires a higher computational cost compared to classical methods; however, we believe that the flexibility and better predictive ability of the resulting models outweigh the computational burden.

The proposed model assumes that the environmental and genetic effects are accounted for by climate variables associated with the planting site and seed sources, respectively. However, among and within seed source variation could be affected by other factors related to the composition and evolution of the population under study. For instance, considerable genetic differences created from past events such as migration or gene flow can result in spatial patterns that are independent of climate.

The model was developed under the assumption that phenotypic variation of pine trees occurs geographically, which was documented by different studies of loblolly pine [Wells, 1983, Farjat et al., 2016]. The second important assumption is that geographic locations can be characterized by a set of climate variables, as a result, climate variables are used as predictors to identify optimal locations for planting. One of the caveats of identifying seed sources solely by a set of climate variables is that the predicted relative response of genetically different seed sources coming from the same location and planted in the same place will be identical. That is, the model gives the average relative performance of
genetically diverse groups based on seed source similarities.

Future work could be aimed to extend the proposed model to account for the existing genetic differences within seed sources from a given location. Specifically, modeling individual pine family performance is important, because most commercial plantations consist of only a few families. Another potentially interesting avenue to extend the current model would be to include the susceptibility of loblolly pine to fusiform rust infection (Cronartium quercuum f. sp. fusiforme). Moreover, since tree infection and tree height are intertwined, perhaps these two traits could be considered jointly. For instance, Nathoo [2010a] proposed a Bayesian spatial model relating infection and growth applicable for the analysis of forestry data. This extension will help to tackle the challenging task of utilizing genetic variation within loblolly pine provenances to permit development of planting stock for establishment of plantations resilient to the expected pace of climate change.

References


