# Mapping and imputing potential productivity of Pacific Northwest forests using climate variables

**Gregory Latta, Hailemariam Temesgen, and Tara M. Barrett**

Abstract: Regional estimation of potential forest productivity is important to diverse applications, including biofuels supply, carbon sequestration, and projections of forest growth. Using PRISM (Parameter-elevation Regressions on Independent Slopes Model) climate and productivity data measured on a grid of 3356 Forest Inventory and Analysis plots in Oregon and Washington, we evaluated four possible imputation methods to estimate potential forest productivity: nearest neighbour, multiple linear regression, thin plate spline functions, and a spatial autoregressive model. Productivity, measured by potential mean annual increment at culmination, is explained by the interaction of annual temperature, precipitation, and climate moisture index. The data were randomly divided into 2237 reference plots and 1119 target plots 30 times. Each imputation method was evaluated by calculating the coefficient of determination, bias, and root mean square error of both the target and reference data set and was also tested for evidence of spatial autocorrelation. Potential forest productivity maps of culmination potential mean annual increment were produced for all Oregon and Washington timberland.

Resume: L'estimation regionale de la productivite forestiere potentielle est importante pour diverses applications, y compris les stocks de biocarburants, la sequestration du carbone et les projections de la croissance forestiere. A l'aide des donnees climatiques PRISM *(Parameter-elevation Regressions on Independent Slopes Model)* et des donnees de productivite mesurees sur une grille de 3356 placettes du programme d'analyse et d'invcntaire forestiers dans les Etats de l'Oregon et de Washington, nous evaluons quatre methodes d'imputation pour estimer la productivite forestiere potentielle :le plus proche voisin, la regression lineaire multiple, les fonctions dinterpolation spline et un modele spatial autoregressif. Mesuree par l'accroisscmeru annuel moyen potentiel (AAMP) maximum, la productivite est expliquee par l'interaction de la temperature annuelle, des precipitations et de l'indice d'humidite du climat. Les donnees ont etc repartics au hasard dans 2237 placettes de reference et 1119 placettes cibles une trentaine de fois. En plus d'etre testee pour la presence d'autocorrelation spatiale, chaque methode d'imputation a ete evaluee a l'aide du coefficient de determination, du biais et de l'erreur quadratique moyenne calculecs pour les ensembles de donnees cibles et de reference. Les cartes de productivite forestiere potentielle de l' AAMP maximum ont ete produites pour l'ensemble du territoire forestier exploitable des Etats de l'Oregon et de Washington.

[Traduit par la Redaction]

## Introduction

Forest productivity is of interest in a wide variety of applications and, correspondingly, is measured in a wide variety of ways. Net primary productivity of forests, i.e., the rate at which energy is produced by plants through photosynthesis in excess of energy used in respiration and maintenance, is a fundamental measure of productivity in ecology (Krebs 1985). Biomass productivity in forests is of increasing interest for biofuels production and is closely related to the rate of carbon sequestration, which is of importance to understanding climate change (Marland and Schlamadinger 1997). The potential for forested lands to produce wood over time, which is the focus of this paper, has long been of interest to foresters and is closely related to these other measures of forest productivity (Gower et al. 2001).

The temporal pattern of productivity for an even-aged stand after initiation appears as (i) a period of slow but increasing productivity as trees are established, (ii) an inflection point of maximum rate of productivity, and (iii) a period of decreasing productivity as the stand fully occupies the site (Moller 1947; McArdle et al. 1949)_Maximum long-term wood productivity in managed even-aged forests is realized when rotation length is set to the time of culmination of mean annual increment (Davis and Johnson 1987). The potential mean annual increment (PMAI) of wood volume that occurs at this time of culmination is a measure of the maximum potential forest productivity at a particular site. It reflects the average annual productivity of wood volume that would be realized over the long term, if rotation lengths were chosen to maximize wood production. Maximum potential forest productivity is affected by species mix and silvicultural practices, such as thinning, fertilization, or brush treatment, as well as by the basic factors that determine site, such as soils and climate.

In the western US, potential forest productivity is assessed by the Forest Inventory and Analysis (FIA) program whereby site trees are selected to produce a site index for forested field plots. The site index is then used in combina-

Received II July 2008. Accepted 9 March 2009. Published on the NRC Research Press Web site at cjfr.nrc.ca on 17 June 2009.

G. Latta' and H. Temesgen. Department of Forest Engineering, Resources and Management, Oregon State University, Corvallis, OR 97331, USA.
T.M. Barrett. Forestry Sciences Laboratory, Pacific Northwest Research Station, Anchorage, AK 99503, USA.

'Corresponding author (e-mail: greg.latta@oregonstate.edu).

tion with normal yield tables to determine PMAI, i.e., the maximum potential cubic metre volume per hectare per year that would be produced over the long term at a given site for a normally stocked stand (Hanson et al. 2002).

While this method is useful in many cases, alternative methods of estimating potential forest productivity are needed. Users often prefer forest productivity maps with continuous spatial coverage rather than the point-level estimates provided by plots. In addition, it is common to be unable to find valid site trees at a plot. In the Pacific West region (Alaska, Oregon, Washington, and California), approximately 20% of forested plots do not have PMAI estimates because valid site trees cannot be found. There are a number of reasons for the lack of site trees on plots. (i) The trees present may be too young and, thus, boring would damage the tree, or the available site index equations do not cover this age range. (ii) Trees may be unsuitable for the available site index equations because of poor form, suppression, disease, or old age. (iii) Western hardwoods are often impractical to bore because of rot, the difficulty of reading rings, and a propensity for breaking borers.

Alternatives to using site index trees to estimate productive potential have been described in the published literature. In the Rocky Mountain states, it is common to use habitat typing to describe potential vegetation in the absence of disturbance (Pfister et al. 1977), which is frequently crosswalked to a site index or site class. The assumptions underlying habitat typing may limit general applicability (Cook 1996), and it is only available for portions of western North America. For these reasons, habitat typing was eliminated as a source for PMAI estimates for this study.

With recent advances in geographic information systems, mapping forest productivity over large geographic regions has become more common. Monserud et al. (2008) estimated a linear model of site index explained by growing degree days and mapped potential future changes in site index under climate change. Wang et al. (2005) compared site index predictions for a variety of methods (least-squares regression, generalized additive model, tree-based model, and neural network model), using stem-analysis of lodgepole site trees in Alberta. Nigh et al. (2004) created a suite of site index models, using weather data, PRISM data for precipitation, and site index data. Studies using multiple linear regression (MLR) to relate Douglas-fir productivity to geographic and climatic variables have been completed in France (Curt et al. 2001), Portugal (Fontes et al. 2003), and central Italy (Corona et al. 1998). Swenson et al. (2005) predicted site index for all of Oregon using a process-based 1 km$^2$ resolution growth model with FIA plot data (fuzzed coordinates), DAYMET weather inputs, and soil nitrogen.

As much of this past work uses nonspatial models, there is the potential for a spatial component in prediction error, as can be seen in errors such as those reported by Swenson et al. (2005). In contrast, Monserud et al. (2006) created a spatial model of site index for lodgepole pine *(Pinus contorta* Dougl. ex. Loud.) in Alberta using Hutchinson's thin-plate smoothing splines with inputs of latitude, longitude, elevation, and site index. Expanding on the work of these earlier studies, we developed an autoregressive model to reduce the spatial component of prediction error and compared the model with thin-plate splining, such as those used by Monserud et al. (2006); nearest neighbour imputation; and the nonspatial multiple linear regression used by most previous researchers.

The objective of this study was to examine and identify the best practical method for using climatic variables to estimate PMAI for forested plots in Oregon and Washington. We also produced productivity maps for forested areas in both states that can aid forest rnanagers in management decisions. Four different imputation approaches were evaluated: nearest neighbour (NN), multiple linear regression (MLR), thin plate splines (TPS), and a simultaneous autoregressive model (SAR).

## Methods

### Data

Data for this study were obtained from the FIA databases for Oregon and Washington. The FIA databases are part of the national inventory of forests for the United States. A tessellation of hexagons, each approximately 2400 ha in size, is superimposed across the nation, with one field plot randomly located within each hexagon. Approximately the same number of plots is measured each year; each plot has the same probability of selection. In the western US, plots are remeasured every 10 years. Each field plot is composed of four subplots, with each subplot composed of three nested fixed-radius areas used to sample trees of different sizes. Forested areas that are distinguished by structure, management history, or forest type, are mapped as unique polygons (also called condition-classes) on the plot and correspond to stands of at least 0.4047 ha in size. PMAI is calculated from the stand's site index, which is itself calculated from the age and height of site trees (Hanson et al. 2002). For our study area, there were 4557 forested FIA plots measured between 2001 and 2006, with PMAI values calculated based on 27 different tree species. Of the 4557 plots with PMAI values, 74% were estimated from either Douglas-fir *(Pseudotsuga menziesii* (Mirb.) Franco), ponderosa pine *(Pinus ponderosa* C. Lawson), or western hemlock *(Tsuga heterophylla* (Raf.) Sarg.). PMAI, elevation, and species identifier for the 3356 forested FIA plots of these three primary species in Oregon and Washington plots were obtained from the FIA annual database.

Graphical analysis of plots with site trees of at least two different species from the 3356 FIA plots indicated an upward shift in the PMAI on plots where western hemlock site tree measurements were used, while PMAI showed no discernable trend on plots with both ponderosa pine and Douglas-fir site trees. This upward shift in PMAI on plots for which western hemlock trees were used is most likely due to the shade tolerance of the species. To account for this, an indicator variable for shade tolerance (ST) was used for plots with PMAI calculated from western hemlock trees.

We used normal monthly temperature and precipitation data for the period 1971-2000 produced by the parameter-elevation regressions on independent slopes model (PRISM). The PRISM data is provided on an 800 m grid, which produced differences between measured plot elevation and overlaid PRISM grid elevation of up to 350 m in the mountainous areas of Oregon and Washington. To account for changes in climate due to these elevation differences, we

**Table 1.** Summary statistics for Forest Inventory and Analysis plot data.

| Variable | Mean | Median | Maximum | Minimum | SD |
|---|---|---|---|---|---|
| PMAI ($m^3 \cdot ha^{-1} \cdot year^{-1}$) | 7.4 | 6.8 | 23.8 | 0.2 | 4.08 |
| **Geographic** | | | | | |
| Latitude (°) | 45.4 | 45.3 | 49.0 | 42.0 | 1.98 |
| Longitude (°) | –121.6 | –122.1 | –116.5 | –124.7 | 2.02 |
| Elevation (m a.s.l.) | 808.2 | 777.4 | 2171.4 | 4.9 | 503.97 |
| **Climatic** | | | | | |
| Temperature (°C) | 8.4 | 8.5 | 13.7 | 0.8 | 2.16 |
| Precipitation (cm) | 143.7 | 127.0 | 593.9 | 25.7 | 90.62 |
| CMI (cm) | –30.3 | –34.2 | 83.8 | –80.0 | 19.61 |

**Note:** PMAI, potential mean annual increment; CMI, climate moisture index; SD, standard deviation.

**Table 2.** The correlation matrix for the data set.

| | PMAI | Latitude | Longitude | Elevation | Temperature | Precipitation | CMI | Shade tolerance |
|---|---|---|---|---|---|---|---|---|
| PMAI | 1.00 | | | | | | | |
| Latitude | 0.12 | 1.00 | | | | | | |
| Longitude | –0.60 | 0.30 | 1.00 | | | | | |
| Elevation | –0.68 | –0.20 | 0.63 | 1.00 | | | | |
| Temperature | 0.55 | –0.25 | –0.67 | –0.83 | 1.00 | | | |
| Precipitation | 0.62 | 0.10 | –0.63 | –0.50 | 0.39 | 1.00 | | |
| CMI | 0.60 | 0.29 | –0.47 | –0.45 | 0.22 | 0.84 | 1.00 | |
| Shade tolerance | 0.46 | 0.22 | –0.21 | –0.19 | 0.05 | 0.38 | 0.41 | 1.00 |

**Note:** PMAI, potential mean annual increment; CMI, climate moisture index.

utilized a process similar to that of Wang et al. (2006), where we created a scale-free interpolation process using a 90 m digital elevation model and PRISM temperature and elevation gradients of the larger 800 m grid. The result was a 90 m monthly climate grid. Like Wang et al. (2006), we used this procedure for temperature ($T$) only and used a simple distance weighting method for precipitation ($P$).

As a measure of moisture availability, we used climate moisture index (CMI), which is a measure of precipitation in excess of evapotranspiration (ET). Hourly shortwave incoming solar radiation (SR) was calculated based on Coops et al. (2000), utilizing latitude, longitude, slope, aspect, elevation, and the PRISM monthly maximum and minimum temperatures. These hourly values were then used to create monthly averages. The daily evapotranspiration in month $m$, $ET_m$, was calculated using the Hargreaves method as presented in Narongrit and Yasuoka (2003) as

$$[1] \quad \mathrm{ET}_m = 0.0135(T_m + 17.78)\mathrm{SR}_m\left(\frac{238.8}{595.5 - 0.55T_m}\right)$$

CMI is then calculated using the following equation:

$$[2] \quad \mathrm{CMI} = \sum_{m=1}^{\mathrm{mgs}}[P_m - (\mathrm{days}_m \mathrm{ET}_m/10)]$$

where mgs is the months in the growing season, $P_m$ is the precipitation in month $m$, $\mathrm{days}_m$, is the number of days in month $m$, and $ET_m$ is the daily evaporation in month $m$. In this study, the growing season months are those that have growing degree days above 10°C.

Descriptive statistics for the geographic variables along with the climatic variables used in the models are given in Table 1 and maps are provided in Fig. 1. Correlation coefficients between selected variables are given in Table 2.

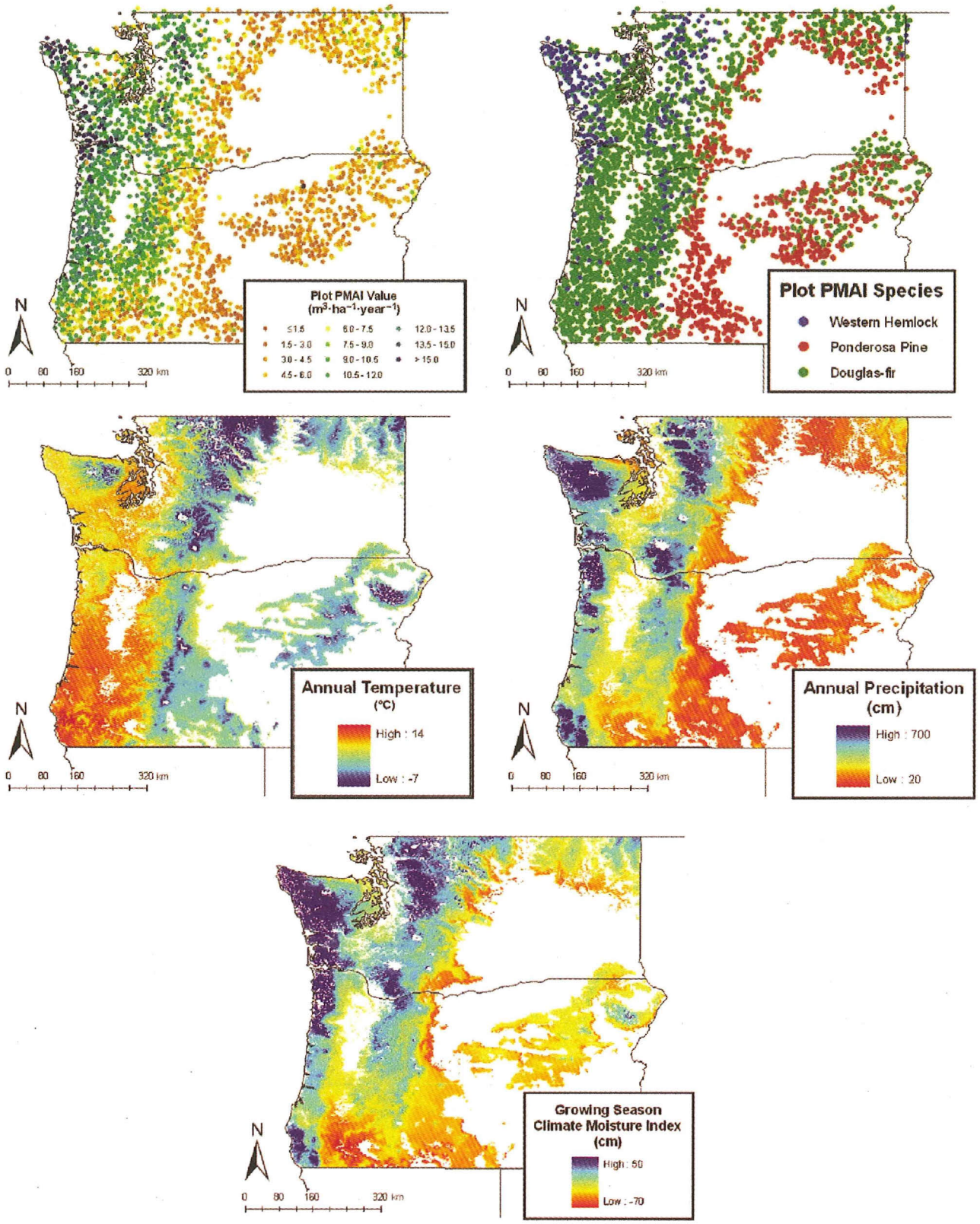
### Nearest neighbor

Unlike the parametric prediction approaches, nearest neighbour (NN) methods can retain both spatial and attribute variance structures of the data (Moeur and Stage 1995; Temesgen et al. 2003), do not restrict the form or shape of the underlying distribution, and will always result in projections within the bounds of biological reality (Moeur and Stage 1995). This approach is therefore feasible for imputing and mapping potential productivity of forests as an alternative to parametric approaches.

In this study, the 3356 sample FIA plots ($n$) were randomly divided as 2237 reference and 1119 target plots. Reference plots refer to sampled plots that had both calculated PMAI values and climate attributes, while target plots refer to unsampled plots that only had climate attribute data. Reference plots formed the pool of potential data that could be selected to impute site productivity of target polygons by each method described below. The target plots were assumed to be unsampled plots (missing site productivity data) and were used to validate the accuracy of each imputation approach by comparing the observed PMAI with the imputed PMAI. The random division of the data set was replicated 30 times.

The NN approach linked a target plot to a reference plot by selecting the reference plot with the smallest squared Euclidian distance between the climatic variables on the reference plot and target plot, respectively. Let $x_i$ be the vector

**Fig. 1.** Maps of the plot, climatic variables, and potential mean annual increment (PMAI).



of climatic variables ($T_i$, $CMI_i$, $TP_i$, $T^2_i$, $P^2_i$, $ST_i$) on plot $i$. The squared Euclidian distance, dist., between target plot $j$ and reference plot $i$ was calculated as

$$[3] \quad \text{dist}_{ji} \sqrt{\sum_{p=1}^{\text{No. of var}} \left( \frac{x_{ip} - x_{jp}}{\overline{x_p}} \right)^2}$$

where $x_{ip}$ is the value of climatic variable $p$ on plot $i$, $\overline{x_p}$ is the mean value of climatic variable $p$, and No. of var is the total number of climatic variables. One of the best aspects of the NN approach is that the PMAI imputed to the target plot did actually occur on a reference plot with similar climatic attributes.

## Multiple linear regression

The linear regression model solves an equation for a set of parameters that minimize the squared errors of a dependent variable and its predicted value from the equation. Imputing PMAI using MLR consists of solving a linear regression for the reference plots and then using that equation to calculate a predicted value for the target plots. Unlike NN, the MLR approach may produce a PMAI value for the target plot that is inconsistent with both the target plot climatic attributes and any reference plot climatic attributes. The PMAI on references plot $i$, $PMAI_i$, are given by the MLR model:

$$[4] \quad \text{PMAI}_i = \beta_1 + \beta_2 T_i + \beta_3 \text{CMI}_i + \beta_4 T_i P_i + \beta_5 T_i^2 + \beta_6 P_i^2 + \beta_7 \text{ST}_i + e_i$$

where $\beta_1, \ldots, \beta_7$ are regression parameters, $i$ is the set of reference plots, and $e_i$ is a stochastic error on plot $i$. The model is solved using ordinary least squares (OLS), and the equation for predicting imputed PMAI for target plot $j$, $PMAI_j$, was

$$[5] \quad \widehat{\text{PMAI}}_j = \beta_1 + \beta_2 T_j + \beta_3 \text{CMI}_j + \beta_4 T_j P_j + \beta_5 T_j^2 + \beta_6 P_j^2 + \beta_7 \text{ST}_j$$

## Thin plate splines

TPS has been widely used for interpolating spatial data. TPS minimizes prediction error by a generalization of the standard MLR in which the parametric model is replaced by a suitable smooth nonparametric model (Hutchinson 2006). This is done by the process of generalized cross validation (GCV) in which each data point is singularly removed and the error for that point is estimated from a surface fitted to the remaining data (Hutchinson and Gessler 1994). These GCV estimated errors are then utilized to minimize the error in the final fitted surface. We used the ANUSPLIN version 4.36 (Hutchinson 2006) package to impute missing site productivity data. ANUSPLIN uses a set of data points, or knots, to create its interpolated surface. While typically ANUSPLIN will determine the best points in a data set to serve as knots, for our 30 replications, the randomly chosen reference plots were given to ANUSPLIN as the knots to use. We used five independent spline variables (latitude, longitude, $T$, CMI, and $P$) to generate a third-order spline over one surface. The output from ANUSPLIN is a grid, or raster map, which can then be overlaid with data points to impute values. Given that our data points were not on a grid, we used ANUSPLIN to predict PMAI for a one-cell grid at each of our data points. This eliminates any differences that may arise if the climate of the larger grid cell differs from that of the individual point to be imputed.

## Simultaneous autoregressive model

Spatial autocorrelation is a frequent occurrence in spatial vegetative modeling, as nearby observations are more similar than if they had been selected at random. The result is that the OLS parameter estimates of the original models, while still unbiased, are no longer the most efficient. While testing for and correcting autocorrelation has been prevalent in the econometrics literature for decades, it is just recently gaining acceptance as a method for solving spatial models. In a SAR model, the error term is composed of two components: a stochastic error term and an error term that is a function of the neighboring data error terms. The MLR model described above was tested for the presence of spatial autocorrelation, and a SAR model was generated, where the PMAI on reference plot $i$, $PMAI_i$, is given by

$$[6] \quad \text{PMAI}_i = \beta_1 + \beta_2 T_i + \beta_3 \text{CMI}_i + \beta_4 T_i P_i + \beta_5 T_i^2 + \beta_6 P_i^2 + \beta_7 \text{ST}_i + \rho u_i + e_i$$

The model contains eq. 6 with the addition of the term $\rho u_i$, where $\rho$ is the autocorrelation correction parameter, and $u_i$ is the spatially autocorrelated error term (which would be the $e_j$ in eq. 6). The autocorrelated error term is a function of the lagged or neighbouring error terms, and for plot $i$, $u_i$ is given as

$$[7] \quad u_i = \text{PMAI}_{L_i} - (\beta_1 + \beta_2 T_{L_i} + \beta_3 \text{CMI}_{L_i} + \beta_4 T_{L_i} P_{L_i} + \beta_5 T_{L_i}^2 + \beta_6 P_{L_i}^2 + \beta_7 \text{ST}_{L_i})$$

Let $XLiP$ be the vecto~ of lagged terms for the variables $(hi'$ $CMI_{Li}$, $TP_{L;}$, $7]''$ $P_{L;}$, $ST_{L)}$ on plot $t$. The lagged term for the PMAI for plot $i$, $PMAI_{L;}$, is calculated as

$$[8] \quad \text{PMAI}_{L_i} = \frac{\sum_{k \in \text{NW}_i} w_k \text{PMAI}_k}{\sum_{k \in \text{NW}_i} w_k}$$

and

$$[9] \quad x_{L_i p} = \frac{\sum_{k \in \text{NW}_i} w_k x_{kp}}{\sum_{k \in \text{NW}_i} w_k}$$

where $NW_i$ is the neighboring window of observation (or plot) $i$, PMAI is the PMAI value of observation (or neighbor, plot) $k$ within the neighboring window, $X_{kp}$ is the value of climatic variable $p$ of neighboring plot $k$, and $w_k$ is the weighting term for observation (or neighbor, plot) $k$, given as the inverse Euclidian distance between observation $i$ and its neighbor $k$. This distance is calculated using longitude and latitude decimal degrees, and the neighboring window is set to $0.5^\circ$ (in all directions). The model is solved using nonlinear least squares.

### Comparison of approaches

For each of the four imputation methods, the random separation of the data into target versus reference stands was repeated 30 times. Fit statistics commonly used by other authors are based on comparing observed with estimated values in the simulated target data set, particularly, the squared correlation between the actual and predicted values. The average difference, often called bias, and root mean squared error (square root of the average squared difference; RMSE) are often calculated. We calculated the fit statistics for each imputation method for the reference and target plots.

To evaluate the results for each simulation, coefficient of determination ($r^2$), bias (average difference), and RMSE were calculated separately for both the reference and target data sets for each replicate as follows:

$$[10]\quad r^2 = \left( \frac{\sum_{i=1}^{n}(\mathrm{PMAI}_i - \overline{\mathrm{PMAI}})(\widehat{\mathrm{PMAI}}_i - \overline{\widehat{\mathrm{PMAI}}})}{\sqrt{\sum_{i=1}^{n}(\mathrm{PMAI}_i - \overline{\mathrm{PMAI}})^2 \sum_{i=1}^{n}(\widehat{\mathrm{PMAI}}_i - \overline{\widehat{\mathrm{PMAI}}})^2}} \right)^2$$

$$[11]\quad \mathrm{bias} = \frac{\sum_{i=1}^{n}(\mathrm{PMAI}_i - \widehat{\mathrm{PMAI}}_i)}{n}$$

$$[12]\quad \mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(\mathrm{PMAI}_i - \widehat{\mathrm{PMAI}}_i)^2}{n}}$$

where $n$ is the number of plots in either the reference or target data set.

The mean, minimum, maximum, and standard deviation of each of these three statistics were summarized over the 30 sampling replications.

### Analysis of spatial autocorrelation

Spatial autocorrelation is the correlation between the model's error term and the error term of nearby observations. The presence of spatial autocorrelation violates one of the assumptions of the classical linear regression model - that the disturbance term relating to any observation is not influenced by the disturbance term relating to any other observation. Some of the possible causes of spatial autocorrelation include measurement error, omitted variables, incorrect functional form, and incorrect data transformations. Autocorrelation in an OLS regression will still result in unbiased coefficient estimates; however, they may no longer be the most efficient unbiased estimators. This potential inefficiency would lead to standard errors that are no longer appropriate, and any hypothesis testing using these standard errors may be misleading. A commonly used measure of global spatial autocorrelation is Moran's I. Like a correlation coefficient, the values of Moran's I vary from -1 to 1, depending on the direction and magnitude of the spatial autocorrelation. Negative values indicate dispersal and positive values indicate clustering.

The equation used to calculate Moran'sI is given as follows:

$$[13]\quad I = \frac{n\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(E_i - \overline{E})(E_j - \overline{E})}{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\sum_{i}^{n}(E_i - \overline{E})^2}$$

whereIs Moran'sItatistic, $E_i$ and $E_j$ are the observed PMAI minus the predicted PMAI on plots $i$ and $j$, respectively, with $\overline{E}$ as its mean, and $w_{ij}$ is the inverse Euclidian distance between observations $i$ and $j$ if the distance is less than 0.5°, otherwise 0. The standard error for Moran's 1 with a symmetric weighting matrix assuming normality of the variance is calculated by

$$[14]\quad \mathrm{SE}(I) = \sqrt{\frac{\frac{n^2}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}(2w_{ij})^2 - n\sum_{i=1}^{n}(2\sum_{j=1}^{n}w_{ij})^2 + 3(\sum_{i=1}^{n}\sum_{j=1}^{n}w_{ij})^2}{(\sum_{i=1}^{n}\sum_{j=1}^{n}w_{ij})^2(n^2-1)} - \frac{-1}{(1-n)}}$$

A Z score can then be calculated to represent the statistical significance of the I value. The Z score is given as

$$[15]\quad Z = \frac{\left(I - \frac{-1}{(1-n)}\right)}{\mathrm{SE}(I)}$$

## Results and discussion

### Evaluation of imputation methods

Table 3 has the $r^2$, RMSE, and bias values for each of the modeling methods. The NN model performed the worst on this data set, with $r^2$ values of 0.50 and 0.49 and RMSE values of 3.14 and 3.15 $m^3 \cdot ha^{-1} \cdot year^{-1}$ for the reference and target data sets, respectively. The MLR model had very similar performance on both the reference and target data sets with RMSE values of 2.45 and 2.46 $m^3 \cdot ha^{-1} \cdot year^{-1}$, respectively, and an $r^2$ of 0.64 for both. Both the TPS and SAR models, which use additional information regarding each plot's error term, outperformed the simpler NN and MLR models. The TPS model had an $r^2$ of 0.71 for the reference data, while its $r^2$ for the target data fell by 2.6% to 0.69 for the target data. None of the other three models exhibited a drop in $r^2$

**Table 3.** Average statistics from 30 replications.

| Imputation approach | Reference data set | | | | Target data set | | | |
|---|---|---|---|---|---|---|---|---|
| | Average | Minimum | Maximum | SD | Average | Minimum | Maximum | SD |
| $r^2$ | | | | | | | | |
| NN | 0.50 | 0.47 | 0.52 | 0.017 | 0.49 | 0.45 | 0.54 | 0.021 |
| MLR | 0.64 | 0.62 | 0.66 | 0.008 | 0.64 | 0.61 | 0.67 | 0.016 |
| TPS | 0.71 | 0.69 | 0.72 | 0.008 | 0.69 | 0.65 | 0.72 | 0.015 |
| SAR | 0.74 | 0.72 | 0.75 | 0.007 | 0.73 | 0.70 | 0.77 | 0.013 |
| **RMSE** | | | | | | | | |
| NN | 3.14 | 3.03 | 3.26 | 0.062 | 3.15 | 2.99 | 3.29 | 0.083 |
| MLR | 2.45 | 2.36 | 2.50 | 0.026 | 2.46 | 2.35 | 2.62 | 0.052 |
| TPS | 2.21 | 2.15 | 2.29 | 0.030 | 2.28 | 2.11 | 2.43 | 0.066 |
| SAR | 2.09 | 2.04 | 2.14 | 0.024 | 2.11 | 2.01 | 2.22 | 0.049 |
| **Bias** | | | | | | | | |
| NN | –0.005 | –0.090 | 0.058 | 0.036 | 0.004 | –0.206 | 0.252 | 0.107 |
| MLR | 0.000 | 0.000 | 0.000 | 0.000 | 0.009 | –0.176 | 0.129 | 0.073 |
| TPS | –0.008 | –0.049 | 0.044 | 0.023 | 0.015 | –0.089 | 0.098 | 0.046 |
| SAR | 0.000 | 0.000 | 0.001 | 0.000 | 0.017 | –0.132 | 0.109 | 0.067 |

**Note:** NN, nearest neighbour; MLR, multiple linear regression; TPS, thin plate splines; SAR, simultaneous autoregressive model.

from the reference to the target data of more than 0.5%. Likewise, the TPS model had a reduction of precision of 3.2%, as RMSE increased from 2.21 to 2.28 $m^3 \cdot ha^{-1} \cdot year^{-1}$ from the reference to the target data sets. The highest loss in precision among the other models was 0.7%. The top performing model for both the reference and target data sets was the SAR model with $r^2$ values of 0.74 and 0.73 and RMSE values of 2.09 and 2.11 $m^3 \cdot ha^{-1} \cdot year^{-1}$, respectively.

Among the four imputation methods examined, the NN method resulted in the widest range of bias for both the reference and target data sets (Table 3). The wide range of bias resulting from using the NN method has also been reported in estimating other variables, including stand density (LeMay and Temesgen 2005), stand tables (Eskelson et al. 2008), and abundance of cavity trees (Temesgen et al. 2008).

## Predicted error maps

While most studies present statistics supporting how well their models fit the data, very few present the spatial structure of the model error. Figure 2 maps out the observed PMAI minus the predicted PMAI to give a spatial representation of the overestimation and underestimation of the four models. The NN, MLR, and TSP models all tend to underestimate PMAIs for productivity on the western side of the border between Oregon and Washington. These three models also tend to overestimate productivity in the far southwestern end of Oregon. The NN and MLR methods also tend to overestimate productivity through the west side of the Cascade Mountains in west central Oregon. The overestimation of productivity in the Klamath Mountains of southwestern Oregon is consistent with the error map produced by Swenson et al. (2005), which also showed a tendency to overestimate productivity there as well. The SAR model errors are more dispersed, not displaying any trends or patterns.

## Spatial autocorrelation

Spatial autocorrelation is a relationship between the error terms of nearby observations. We investigated the spatial relationship of the residuals of the four models, as presented in Fig. 2. We used Morans's *I* to test for the presence of spatial autocorrelation. An important component of Moran's I is the weighting of the observations. Weights are typically defined as a geographic window of influence, some sort of inverse distance function, or a combination of both. We employed both the commonly used inverse distance function as well as a window of influence, which we set at 0.5° of latitude and longitude. The window of influence was determined by examining the residuals from the NN, MLR, and TPS models. In particular, the 0.5° represents the approximate radius of the clusters of underestimation errors found on the far western border of Oregon and Washington and also that of the overestimation cluster found in the far southwest of Oregon in the three models.

Table 4 gives the Moran's *I* value and Z scores for the four models for both the reference and target data sets. The NN, MLR, and TPS models all display significant clustering of the errors terms, corroborating what Fig. 2 showed. The MLR model has both the highest magnitude and most significant levels of spatial autocorrelation. The SAR model has significant but low levels of dispersion of the error term, as indicated by the -0.01 value of Moran's I.

## Regional mean annual increment maps

Figure 3 contains the PMAI maps for the four imputation methods along with the actual plot PMAI values given in Fig. I. As seen on the map, in general, the regression techniques of the MLR and SAR methods produced much smoother transitions between the PMAI classes. The NN map has the least smooth PMAI class transitions, as each individual pixel is assigned a class independent of the neighbouring pixel values. The overestimation of PMAI in the southwest of Oregon (as seen in Fig. 2) for both the NN

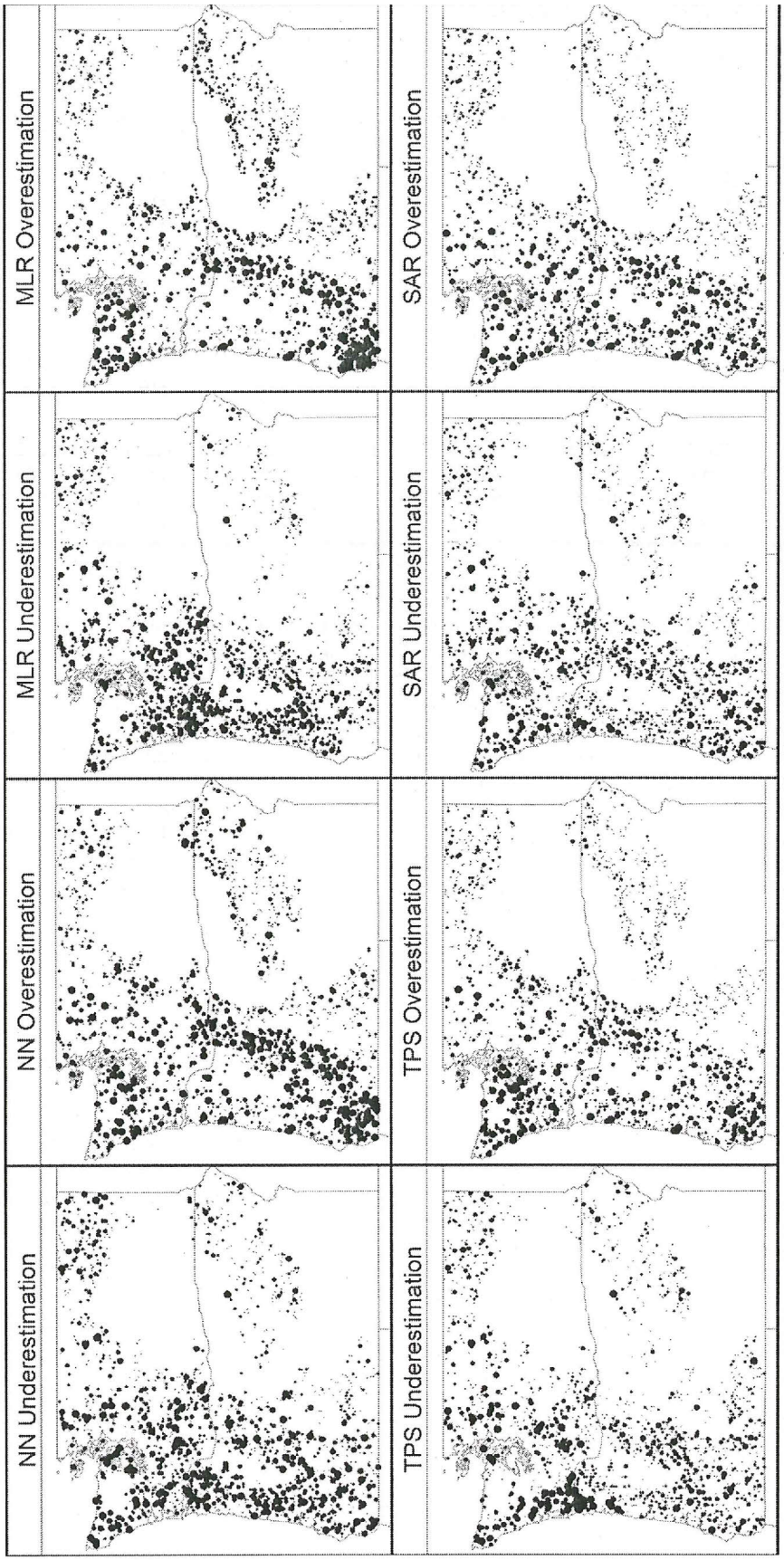


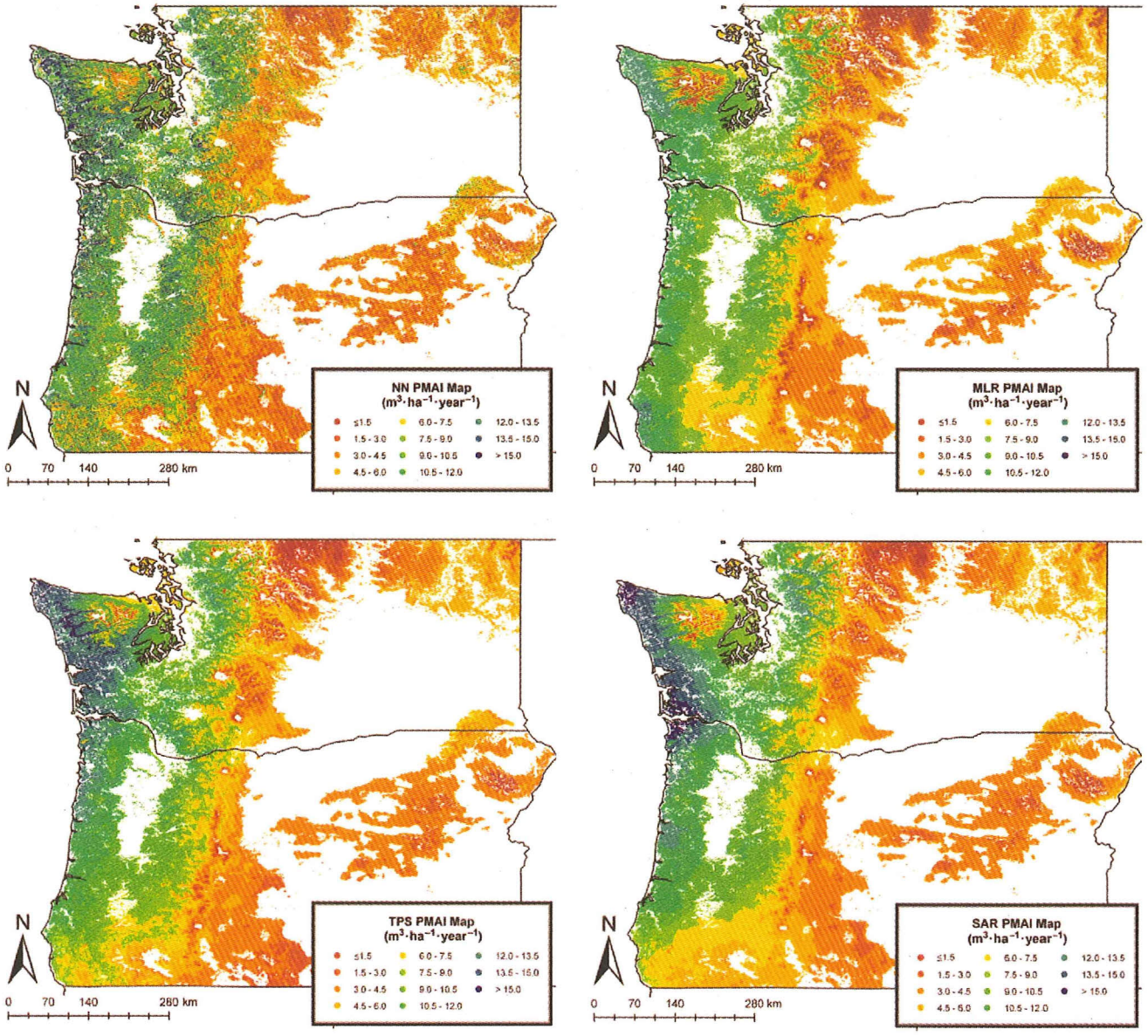

**Fig. 2.** Model prediction error maps.

**Table 4.** Average spatial autocorrelation statistics for 30 replications.

| | Reference data set | | | | Target data set | | | |
|---|---|---|---|---|---|---|---|---|
| | Average | Minimum | Maximum | SD | Average | Minimum | Maximum | SD |
| **Moran's *I*** | | | | | | | | |
| NN | 0.05 | 0.03 | 0.06 | 0.007 | 0.05 | 0.03 | 0.06 | 0.007 |
| MLR | 0.23 | 0.22 | 0.24 | 0.006 | 0.23 | 0.18 | 0.25 | 0.015 |
| TPS | 0.10 | 0.08 | 0.10 | 0.005 | 0.09 | 0.08 | 0.11 | 0.008 |
| SAR | –0.01 | –0.01 | –0.01 | 0.002 | –0.01 | –0.01 | –0.01 | 0.001 |
| **Z score** | | | | | | | | |
| NN | 10.3 | 7.0 | 14.1 | 1.7 | 10.0 | 6.8 | 13.7 | 1.7 |
| MLR | 51.1 | 48.8 | 54.4 | 1.3 | 49.1 | 39.8 | 54.3 | 3.2 |
| TPS | 21.6 | 18.6 | 23.5 | 1.2 | 19.8 | 16.7 | 24.2 | 1.8 |
| SAR | –2.3 | –3.1 | –1.7 | 0.4 | –2.2 | –2.8 | –1.5 | 0.3 |

**Note:** NN, nearest neighbour; MLR, multiple linear regression; TPS. thin plate splines; SAR, simultaneous autoregressive model.

**Fig. 3.** Potential mean annual increment (PMAI) maps for each imputation method.



and MLR models appears to be the result of high coastal PMAI values, which are not found in the TPS and SAR maps. Conversely, the underestimation of PMAI (from Fig. 2) at the mouth of the Columbia River on the western border of Oregon and Washington by the NN, MLR, and TPS models is abated by higher PMAI values in the region predicted by the SAR model.

## Conclusions and recommendations

This study compared the performance of four techniques to impute site productivity based on climatic variables in the Pacific Northwest. Based on 30 random divisions of the data into reference and target data, the SAR model outperformed the NN, MLR, and TPS models in both $r^2$ and RMSE, while at the same time it had the lowest bias of the four modeling techniques. The SAR model also had the least problem with spatial autocorrelation, which was significant and visible in the error maps of the other three methods.

As concern over the impact of climate change leads to increased efforts to model climate change effects across the landscape, spatial modelers will need to become more sophisticated in their techniques. Inflated standard errors of coefficients estimated by OLS with spatial autocorrelation present in the model could lead to errors in interpreting the impacts on productivity of changes in climatic parameters, and entire regions mapped using the model could be prone to systematic errors. Methods such as the SAR model estimated in this study could prove invaluable in their ability to potentially identify regions more susceptible to climatic change. Elasticities of changes in productivity with respect to changes in climatic variables could be mapped, and statistical significance for those elasticities could be presented. Elasticities could also be paired with climate model data, such as the Special Report on Emissions generated for the 4th Assessment Report of the Intergovernmental Panel on Climate Change, to map and quantify changes in future productivity under various scenarios.

## Acknowledgements

This research was supported by the PNW Research Station, USDA, Forest Service, Anchorage, Alaska, and by the College of Forestry, Oregon State University. We thank Robert Monserud and Darius Adams for their thoughtful insights as the project progressed. Thanks are extended to an anonymous Associate Editor and two anonymous reviewers for their helpful comments and suggestions.

## References

Cook, J.E. 1996. Implications of modern successional theory for habitat typing: a review. For. Sci. 42(1): 67-75.

Coops, N.C., Waring, R.H., and Moncrieff, J.B. 2000. Estimating mean monthly incident solar radiation on horizontal and inclined slopes from mean monthly temperature extremes. Int. J. Biorneteoro!. 44: 204-211. doi:10.1007/s004840000073. PMID: 11131293.

Corona, P., Scotti, R., and Tarchiani, N. 1998. Relationship between environmental factors and site index in Douglas-fir plantations in central Italy. For. Ecol. Manage. 110: 195-207. doi: 10.10 16/S0378-1127(98)00281-3.

Curt, 1'., Bouchaud, M., and Agrech, G. 2001. Predicting site index of Douglas-fir plantations from ecological variables in the Massif Central area of France. For. Ecol. Manage. 149: 61-74. doi: 10.1 016/S0378-1127(00)00545-4.

Davis, L.S., and Johnson, K.P. 1987. Forest management. 3rd ed. McGraw-Hill, New York.

Eskelson, B.N.I., Temesgen, H., and Barrett, T.M. 2008. Comparison of stratified and non-stratified most similar neighbour approaches for estimating stand tables. Int. J. For. Res. 81: 125-134.

Fontes, L., Tome, M., Thompson, F., Yeomand, A., Sales Luis, J., and Savill, P. 2003. Modelling the Douglas-fir *(Pseudotsuga menziesii* (Mirb.) Franco) site index from site factors in Portugal. Forestry, 76(5): 491-507. doi: 10.1 093/forestry/76.5.491.

Gower. S.T., Krankina, O., Olson, R.1., Apps, M., Linder, S., and Wand, C. 2001. Net primary production and carbon allocation patterns of boreal forest ecosystems. Ecol. Appl, 11(5): 1395-1411. doi: 10.1890/1051-0761 (2001)011 [1 395:NPPACA]2.0. CO;2.

Hanson, E.1., Azuma, D.L., and Hiserote, B.A. 2002. Site index equations and mean annual increment equations for Pacific Northwest Research Station Forest Inventory and Analysis Inventories, 1985-2001. USDA Forest Service Research Note PNW-RN-533.

Hutchinson, M.F. 2006. ANUSPLIN version 4.36. Centre for Resource and Environmental Studies, The Australian National University, Canberra, http://cres.anu.edu.auloutputs/anusplin.php.

Hutchinson, M.F . and Gessler, P.E. 1994. Splines - more than just a smooth interpolator. Geoderma, 62: 45-67. doi:10.l016/0016-706l (94)90027-2.

Krebs, C.J. 1985. Ecology: the experimental analysis of distribution and abundance. 3rd ed. Harper and Row, New York.

LeMay, Y., and Temesgen, H. 2005. Comparison of nearest neighbor methods for estimating basal area and stems per ha using aerial auxiliary variables. For. Sci. 51(2): 109-119.

Marland, G., and Schlamadinger, B. 1997. Forests for carbon sequestration or fossil fuel substitution? A sensitivity analysis. Biomass Bioenergy, 13(6): 389-397. doi:10.1016/S0961-9534(97)00027-5.

McArdle, R.E., Meyer, W.H., and Bruce, D. 1949. The yield of Douglas-fir in the Pacific Northwest. USDA For. Servo Tech. Bull. 201. Washington, D.C.

Moeur, M., and Stage, A.R. 1995. Most similar neighbour: an improved sampling inference procedure for natural resource planning. For. Sci. 41: 337-359.

Moller, C.M. 1947. The effect of thinning, age, and site on foliage, increment and loss of dry matter. J. For. 45: 393-404.

Monserud, R.A., Huang, S., and Yang, Y. 2006. Predicting lodgepole pine site index from climatic parameters in Alberta. For. Chron. 82(4): 562-571.

Monserud, R.A., Yang, Y., Huang, S., and Tchebakova, N. 2008. Potential change in lodgepole pine site index and distribution under climatic change in Alberta. Can. J. For. Res. 38: 343-352. doi: 10.1139/X07-166.

Narongrit, C., and Yasuoka, Y. 2003. The use of terra-MODIS data for estimating evapotranspiration and its change caused by global warming. Env. Inf. Arch., l: 505-511.

Nigh, G.D., Ying, C.C., and Qian, H. 2004. Climate and productivity of major conifer species in the interior of British Columbia, Canada. For. Sci. 50(5): 659-671.

Pfister, R.D., Kovalchik, B.L., Arno, S.F., and Presby, R.D. 1977. Forest habitat types of Montana. USDA Gen. Tech. Rep. INT-34, USDA Forest Service Intermountain Forest and Range Experiment Station, Ogden, Utah.

Swenson, J.J., Waring, R.H., Fan, W., and Coops, N. 2005. Predict-

ing site index with a physiologically based growth model across Oregon, USA. Can. J. For. Res. 35: 1697-1707. doi:l0.1139/x05-089.

Temesgen, H., LeMay, V.M., Marshall, P.L., and Froese, K. 2003. Imputing tree-lists from aerial attributes for complex stands of British Columbia. For. Ecol. Manage. 177: 277-285. doi:l0.1016/S0378-1127(02)00321-3.

Temesgen, H., Barrett, T., and Latta, G. 2008. Estimating cavity tree abundance using nearest neighbor imputation methods for western Oregon and Washington forests. Silva Fenn. 42(3): 337-354.

Wang, Y., Raulier, F., and Ung, C.H. 2005. Evaluation of spatial predictions of site index obtained by parametric and nonparametric methods - a case study of lodgepole pine productivity. For. Ecol. Manage. 214: 201-211. doi:10.1016/j.forcco.2005.04.025.

Wang, T., Hamann, A., Spittlehouse, D., and Aitken, S. 2006. Development of scale-free climate data for western Canada for use in resource management. Int. J. Climatol. 26: 383-397. doi:l0.I002/joc.1247.

## List of symbols

| | |
|---|---|
| CMI | growing season precipitation in excess of evapotranspiration, em |
| $E$ | prediction error of the models |
| ET | potential daily evapotranspiration, mm-day$^{-1}$ |
| $L$ | lagged value indicator for SAR model |
| mgs | months of the growing season |
| n | number of plots |
| NW | neighborhood window for SAR model |
| $P$ | annual precipitation, ern |
| $P_m$ | total precipitation in month, $m$, em |
| PMAI | potential mean annual increment at culmination, m$^3$·ha$^{-1}$·ycar$^{-1}$ |
| SR | incident solar radiation, MJ·m$^{-2}$·day-1 |
| ST | shade tolerance indicator variable 0,1 |
| $T$ | average annual temperature, °C |
| $T_m$ | average temperature in month, m, °C |
| w | weight for SAR model |
| $x$ | the vector of climatic variables (*T,* CMI, *TP,* $T^2$, *p2,* ST) |