NRC Research Press

ARTICLE

Integrating forest inventory data and MODIS data to map species-level biomass in Chinese boreal forests

Qinglong Zhang, Hong S. He, Yu Liang, Todd J. Hawbaker, Paul D. Henne, Jinxun Liu, Shengli Huang, Zhiwei Wu, and Chao Huang

Abstract: Timely and accurate knowledge of species-level biomass is essential for forest managers to sustain forest resources and respond to various forest disturbance regimes. In this study, maps of species-level biomass in Chinese boreal forests were generated by integrating Moderate Resolution Imaging Spectroradiometer (MODIS) images with forest inventory data using *k* nearest neighbor (*k*NN) methods and evaluated at different scales. The performance of 630 *k*NN models based on different distance metrics, *k* values, and temporal MODIS predictor variables were compared. Random Forest (RF) showed the best performance among the six distance metrics: RF, Euclidean distance, Mahalanobis distance, most similar neighbor in canonical correlation space, most similar neighbor computed using projection pursuit, and gradient nearest neighbor. No appreciable improvement was observed using multi-month MODIS data compared with using single-month MODIS data. At the pixel scale, species-level biomass for larch and white birch had relatively good accuracy (root mean square deviation < 62.1%), while the other species had poorer accuracy. The accuracy of most species except for willow and spruce was improved up to the ecoregion scale. The maps of species-level biomass captured the effects of disturbances including fire and harvest and can provide useful information for broad-scale forest monitoring over time.

Key words: species-level biomass, MODIS, Chinese boreal forest, Random Forest (RF), *k*NN.

Résumé : Une connaissance précise et en temps opportun de la biomasse de chaque espèce est essentielle pour permettre aux aménagistes forestiers d'effectuer un aménagement durable des ressources forestières et pour s'ajuster aux divers régimes de perturbations forestières. Dans cette étude, des cartes de biomasse par espèce ont été générées dans les forêts boréales chinoises en intégrant des images MODIS (spectroradiomètre imageur à résolution moyenne) à des données d'inventaire forestier au moyen de l'approche des k plus proches voisins (*k*NN) et évaluées à différentes échelles. Les performances de 630 modèles *k*NN ont été comparées en fonction de différentes métriques de distance, valeurs de *k* et variables prédictives temporelles de MODIS. Les forêts d'arbres décisionnels (RF) ont fait ressortir les meilleures performances parmi les six métriques de distance : la distance RF, la distance euclidienne, la distance de Mahalanobis, le plus proche voisin dans l'espace de corrélation canonique, le plus proche voisin calculé par poursuite de projection, et le plus proche voisin par gradient. L'utilisation de données MODIS sur plusieurs mois n'a apporté aucune amélioration notable en comparaison de l'utilisation des données MODIS d'un seul mois. À l'échelle des pixels, la biomasse par espèce avait une précision relativement bonne pour le mélèze et le bouleau blanc (écart quadratique moyen < 62,1 %), tandis que la précision était plus faible pour les autres espèces. Pour la plupart des espèces, à l'exception du saule et de l'épinette, la précision s'est améliorée jusqu'à l'échelle de l'écorégion. Les cartes de la biomasse par espèce ont capté les effets des perturbations, y compris les feux et la récolte, et peuvent fournir des informations utiles pour la surveillance des forêts à une vaste échelle au fil du temps. [Traduit par la Rédaction]

Mots-clés : biomasse par espèce, spectroradiomètre imageur à résolution moyenne (MODIS), forêt boréale chinoise, forêts d'arbres décisionnels (RF), méthode des *k* plus proches voisins (*k*NN).

1. Introduction

The boreal forest is the second largest terrestrial biome in the world, covering 33% of forest area and holding 23% of terrestrial carbon stocks (Carlson et al. 2009; Ji et al. 2015). The Chinese boreal forest in the Great Xing'an Mountains of northeastern China is the southern extension of the eastern Siberian light coniferous forest, covering about 11.2% of forest areas in China (Xu 1998). Disturbance by fire and timber harvest have extensively altered forest structure and biomass in this region from stand to landscape scales (Luo et al. 2014; Wu et al. 2013; Xu 1998). Timely and accurate knowledge of species-level biomass in this area is essential for forest managers to design effective forest manage-

Received 14 September 2017. Accepted 2 January 2018.

Q. Zhang and C. Huang. CAS Key Laboratory of Forest Ecology and Management, Institute of Applied Ecology, Shenyang 110016, China; University of Chinese Academy of Sciences, Beijing 100049, China.
H.S. He. School of Natural Resources, University of Missouri, 203 Anheuser-Busch Natural Resources Building, Columbia, MO 65211, USA; School of Geographical Sciences, Northeast Normal University, Changchun 130024, China.
Y. Liang. University of Chinese Academy of Sciences, Beijing 100049, China.
T.J. Hawbaker and P.D. Henne. U.S. Geological Survey, Geosciences and Environmental Change Science Center, PO Box 25046, MS 980, Denver, CO 80225, USA.
J. Liu. Western Geographic Science Center, U.S. Geological Survey, Menlo Park, CA 94025, USA.
S. Huang. USDA Forest Service, Region 5, Remote Sensing Lab, 3237 Peacekeeper Way, Suite 201, McClellan, CA 95652, USA.
Z. Wu. School of Geography and Environment, Jiangxi Normal University, 99 Ziyang Road, Nanchang 330022, Jiangxi, China.
Corresponding authors: Hong S. He (email: heh@missouri.edu) and Yu Liang (email: liangyu@iae.ac.cn).

ment plans to sustain forest resources and respond to changes due to various disturbances.

Species-level forest biomass is typically derived from forest inventory data, which may be limited in space and time; therefore, remotely sensed data are increasingly used to generate spatial records of forest attributes (He et al. 1998; Shataee et al. 2012; Zald et al. 2014; Zhang et al. 2009). Whereas high-resolution optical remotely sensed images (e.g., aerial photographs, IKONOS, WorldView2) and light detection and ranging (lidar) data can derive more accurate species-level forest biomass than coarse-resolution images (e.g., Moderate Resolution Imaging Spectroradiometer, MODIS) (Pu and Landry 2012; van Ewijk et al. 2014; Zellweger et al. 2013), it is often challenging to obtain and analyze high-resolution images over large regions due to their limited spatial and temporal coverage. However, coarse-resolution sensors, with large swath widths, moderate pixel sizes, and near-daily coverage, are efficient for deriving information over large areas (Wilson et al. 2012).

Researchers often integrate field inventory data with coarse-resolution imagery through imputation models to map detailed forest attributes over large areas (Beaudoin et al. 2014; Blackard et al. 2008; Huang et al. 2017; Wilson et al. 2012; Zhang et al. 2014). One widely used approach is *k* nearest neighbor (*k*NN) imputation (Tomppo et al. 2008; Zald et al. 2014), which uses a set of predictor variables (x) to determine a number (k) of most similar reference observations (nearest neighbors or NN) to derive response variables (y) for the target pixel (Crookston and Finley 2008; McRoberts 2012; Ohmann et al. 2011). Predictor variables can include multispectral satellite imagery and other datasets (e.g., climate, topography, soil) that are spatially complete, while response variables are only available for limited sites in the study area and usually include measures of forest composition or structure derived from field plots. In *k*NN imputation, either a single reference observation ($k = 1$) or multiple reference observations ($k > 1$) can be chosen to assign response variable values to a given target pixel (Beaudoin et al. 2014; Duveneck et al. 2015; Hudak et al. 2008; McRoberts 2012; Moeur and Stage 1995; Ohmann and Gregory 2002; Ohmann et al. 2011; Wilson et al. 2012). A major strength of NN imputation where $k = 1$ is the retention of the covariance structure of multiple response variables, because each target is only linked to a single reference (McRoberts 2012; Zald et al. 2014); however, root mean square errors (RMSE) are generally higher when selecting small values of k, especially for $k = 1$, and RMSE is sometimes greater than the standard deviation of the response variable observations, meaning that the overall mean as a prediction for every target element is better at minimizing RMSE than the *k*NN predictions (McRoberts 2012). Predictive accuracy increased with k in previous studies (McRoberts 2012; Muinonen et al. 2001; Wilson et al. 2012). Muinonen et al. (2001) reported that the bias was unstable with k increasing when $k < 8$. Muinonen et al. (2001) and McRoberts (2012) suggested balancing the greater imputation accuracy with the expense of modifying the covariance structure. Therefore, selecting a reasonable k is also necessary for maximizing the accuracy and efficiency of the *k*NN imputation method. Since Tomppo and Katila (1991) first proposed the *k*NN method for applications in forestry using satellite data, it has been widely applied to impute many different forest attributes worldwide (McRoberts 2009; Ohmann and Gregory 2002; Temesgen et al. 2003; Wilson et al. 2012). Many alternative *k*NN distance metrics have been used for associating target and reference pixels, e.g., canonical correlation analysis most similar neighbor (MSN) (Moeur and Stage 1995), gradient nearest neighbor (GNN) using canonical correspondence analysis (CCA) (Ohmann and Gregory 2002), Euclidian distance (McRoberts 2009; Tomppo and Katila 1991), and metrics based on ensemble machine learning methods such as Random Forest (RF) (Breiman 2001; Hudak et al. 2008; Zald et al. 2016). In theory, all of the *k*NN distance metrics can be used to impute biomass at the tree species level. Although Hudak et al. (2008) reported that RF produced superior results compared with other distance metrics when using lidar-derived predictor variables, there is still no agreement about the best distance metric to map species-level biomass from coarse-resolution imagery.

Imputation mapping over large areas often relies on coarse-resolution satellite imagery as predictor variables, especially MODIS data, which are available twice daily and globally and are less expensive and time-consuming to produce and update compared with moderate-resolution imagery (e.g., Landsat) (Wilson et al. 2012). However, passive optical sensors such as MODIS are often limited when estimating forest biomass because of their lower sensitivity to vertical and below-canopy vegetation structure (Zhang et al. 2014). Some previous studies indicated that multi-temporal optical images have the potential to improve the accuracy of aboveground biomass (AGB) estimation (Zhu and Liu 2015). Although Wilson et al. (2012, 2013) imputed tree species and forest carbon stocks over large areas using phenology metrics derived from multi-temporal MODIS data, whether multi-temporal MODIS data improve the accuracy of species-level biomass imputation compared with single-temporal data has not been explored at large scales.

The primary objective of this study was to map species-level biomass in the Great Xing'an Mountains of northeastern China using MODIS reflectance and vegetation indices as predictors. To accomplish this objective, we investigated the performance of species-level biomass imputation models based on six *k*NN distance metrics and different k values while using single- and multi-month composites of MODIS data as predictor variables. Finally, we assessed the accuracy of the predictions from pixel to regional scales to determine what scales are most appropriate for use of our predictions.

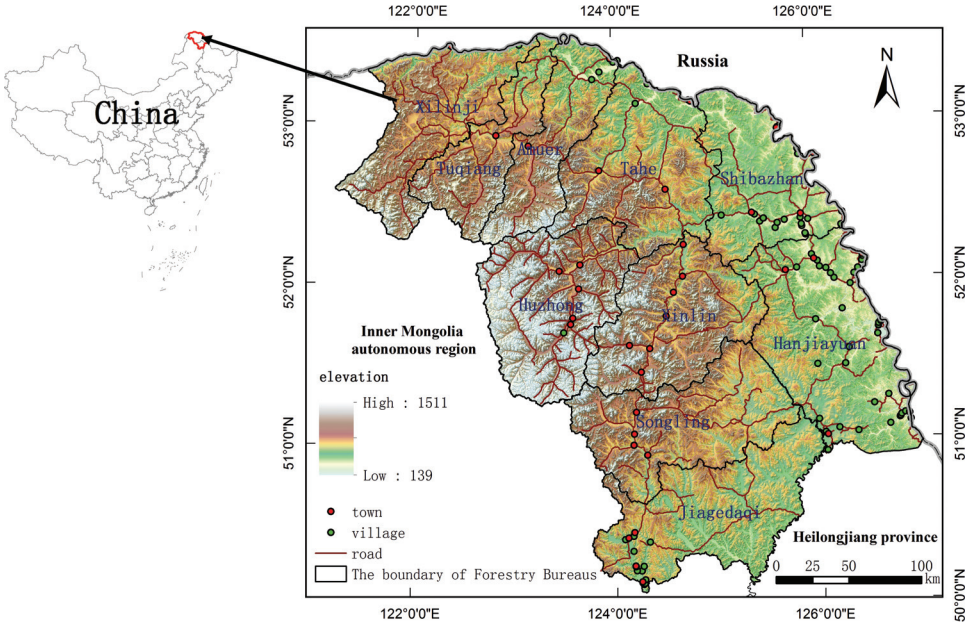
2. Study area and method

2.1. Study area

Our study area is located on the northern and eastern slopes of the Great Xing'an Mountains (121°12′–127°00′E, 50°10′–53°33′N) in northeastern China, covering about 8.46×10^4 km^2. Elevations vary from 139 m in the east to 1511 m in the west. On the whole, the terrain of the study area is gentle and over 80% of the area has a slope of less than 15° (Fig. 1). This region has a long and severe continental monsoon climate with mean annual precipitation varying in a northwestern to southeastern direction and ranging from 240 to 442 mm; 60% or more of all precipitation occurs between June and August. Mean annual temperature varies from −6 to 1 °C in a northwestern to southeastern direction; the coldest month is January, with an average temperature of −33 °C, and the hottest month is July, with an average temperature of 17.5 °C.

The eastern Siberian boreal forest reaches its southernmost extension in the Great Xing'an Mountains. Forests in this region cover about 6.56 million hectares of mostly mountainous terrain. The dominant tree species is Dahurian larch (*Larix gmelinii* (Rupr.) Kuzen, hereafter larch), which is a boreal conifer and late successional species with wide distribution, occupying moist and cool sites. White birch (*Betula platyphylla* Suk.), the second most widely distributed species, is an early successional species and occupies drier, well-drained sites (Xu 1998). Other tree species include Korean spruce (*Picea koraiensis* Nakai, hereafter spruce), Asian black birch (*Betula davurica* Pall.), Mongolian Scots pine (*Pinus sylvestris* var. *mongolica* Litv., hereafter pine), willow (*Chosenia arbutifolia* (Pall.) A. Skv.), two species of aspen (*Populus davidiana* Dode and *Populus suaveolens* Fischer), Mongolian oak (*Quercus mongolica* Fisch. ex Ledeb.), and a shrub species (*Pinus pumila* (Pall.) Regel). Great differences in forest composition, age structure, and tree density in this region result from environmental heterogeneity and disturbances such as fire and timber harvesting. Because of recent timber harvest, mid-seral, secondary forests are the main components of the forest landscape, except in natural reserves (Luo et al. 2014), which are mainly occupied by old-growth forests.

Fig. 1. Elevation (m) and location of the study area. The boundaries of the 10 forestry bureaus are shown with black lines. [Colour online.]



The growth rates of tree species in this area are strongly influenced by enviromental factors determining site conditions. Xu (1998) divided the larch forests into five forest type groups with different growth rates based on site conditions that depend on terrain and geomorphic factors. There are 10 forestry bureaus (Fig. 1), and their boundaries mainly follow watershed boundaries. In this study, 48 ecoregions were generated by integrating the 5 larch forest type groups and 10 forestry bureaus (Fig. 2*a*).

2.2. Data sets

2.2.1. Forest inventory data

The forest inventory data in this study were from the Chinese National Secondary Forest Resource Inventory. We acquired 7635 forest stand polygons (Fig. 2*b*) from the China Forestry Science Data Center (http://www.cfsdc.org/), which is part of the National Secondary Forest Resource Inventory data of the Great Xing'an Mountains from 1997 to 2001. The data contained stand age, mean diameter, stand height, and stand volume density by species in each polygon. Because growth of the boreal forest is relatively slow, we assumed that AGB during 1997–2001 in the study area had not changed sufficiently to affect our overall results. The area of each polygon ranges from several to tens of hectares with relatively homogeneous forest attributes. The polygon boundaries were generated by interpreting aerial photographs according to Chinese technical regulations for inventory for forest management planning and design. The original coordinates of the forest inventory data were in the Beijing 54 coordinate system. To match the forest inventory data with the predictor variables, we transformed coordinates of the forest inventory data and all of the predictor variables into Universal Transverse Mercator (UTM) Zone 51 north projection with the World Geodetic System (WGS) 1984 datum, using the raster package in R (R Core Team 2013).

In each polygon, stand attributes were estimated based on several angle gauge plots (Bitterlich 1948) following a mechanical sampling design. Each angle gauge plot was greater than 50 m away from the stand boundary. The distance between two angle count plots was at least 100 m. Trees with diameter at breast height (DBH) > 5 cm were counted in each angle gauge plot (basal area factor = 1). The DBH of each tree was transformed into volume according to species-specific DBH–volume relationships from the China Forestry Science Data Center (http://www.cfsdc.org/). The volume density of each angle gauge plot was derived by aggregating all single-tree volume estimates by species counted in each angle gauge plot. The stand volume density was estimated by averaging the volume density of all of the angle gauge plots in each polygon. We transformed tree species stand volume into species-level AGB ($t \cdot ha^{-1}$, hereafter species-level biomass) in each polygon using biomass–volume relationships (Fang et al. 1998). Biomass values for eight tree species (larch, white birch, pine, aspen, willow, spruce, Mongolian oak, and black birch) were selected as the response variables for use in the *k*NN methods.

2.2.2. MODIS spectral variables

In this study, seven MODIS surface reflectance bands from MOD09Q1 (b1–b2; 250 m) and MOD09A1 (b3–b7; 500 m) were used as explanatory variables; they were processed into monthly data by averaging all of the reflectance values for each month in year 2000 and resampled to 250 m resolution using a nearest-neighbor algorithm. Several vegetation indices (Table 1) were also calculated from the MODIS monthly surface reflectance. Because the reflectance was largely affected by snow cover from January to April and from November to December, we only used the MODIS monthly surface reflectance and vegetation indices from May through October. Seven sets of MODIS predictors were used in this study, including six sets of single-month MODIS composites from May to October and one set of multi-month MODIS predictors (containing all six sets of single-month MODIS composites). To avoid imputing to non-forest pixels, we separated forest and non-forest areas using the MODIS vegetation continuous fields (VCF) product (MOD44B; 250 m) (Schmitt et al. 2009) for year 2000 and removed areas with less than 10% tree cover.

2.2.3. Environmental variables

To reduce uncertainties in our predictions due to environmental heterogeneity, several environmental variables that were correlated to species-level biomass were selected as auxiliary explanatory data, including climate variables (mean annual precipitation and temperature from 1982 to 2009), topographic variables (elevation, slope, and cosine of aspect (COSASP)), soil variables (bulk density, pH, and the content of sand (%), clay (%), silt (%), gravel (%), and soil organic carbon (%)), and geospatial location (Table 2). Elevation, slope, and aspect were derived from the Shuttle Radar Topogra-

Fig. 2. (*a*) Ecoregion map, (*b*) forest inventory data, and (*c*) the nearest-neighbor distance between each pixel and its nearest reference polygon from forest inventory data with ecoregion map. The nearest-neighbor distance is calculated for each pixel from values for the spatial predictors based on a Random Forest proximity matrix of predictor variables and uncertainty increases with the nearest-neighbor distance. [Colour online.]

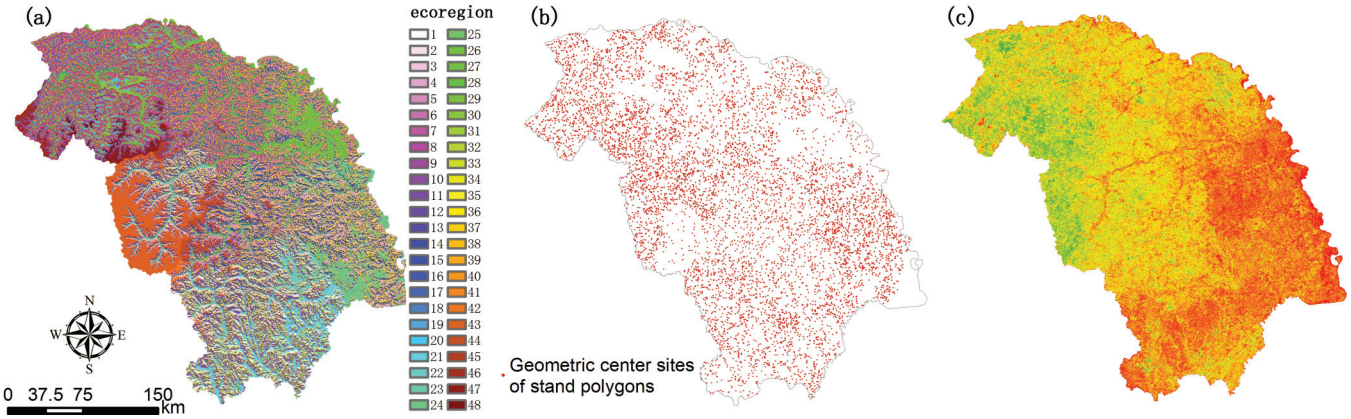


Table 1. Vegetation indices derived from the Moderate Resolution Imaging Spectroradiometer (MODIS) surface reflectance data.

Vegetation index	Formula	Full name	Reference
NDVI[5,6,9]	(b2 − b1)/(b2 + b1)	Normalized difference vegetation index	Rouse et al. (1973)
RVI[8,9]	b2/b1	Ratio vegetation index	Jordan (1969)
EVI	2.5(b2 − b1)/(b2 − 6b1 − 7.5b3 + 1)	Enhanced vegetation index	Huete et al. (2002)
MSAVI[9]	$2b2 + 1 - 0.5\sqrt{2b2 + 1^2 + 8(b2 - b1)}$	Modified soil adjusted vegetation index	Qi et al. (1994)
VARI[5,9]	(b4 − b1)/(b4 + b1 − b3)	Visible atmospherically resistant index	Gitelson et al. (2002)
NDWI[5,8]	(b2 − b5)/(b2 + b5)	Normalized difference water index	Gao (1996)
NDIIb6[6,9]	(b2 − b6)/(b2 + b6)	Normalized difference infrared index	Hunt and Rock (1989)
NDIIb7[5,6,9]	(b2 − b7)/(b2 + b7)	Normalized difference infrared index	Hunt and Rock (1989)
SAVI[5,6,7,8,10]	(b2 − b1)/(b2 + b1 + 0.5) × 1.5	Soil adjusted vegetation index	Huete (1988)
GEMI[8]	$n(1 - 0.25n) - (b2 - 0.125)/(1 - b2)$ $n = (2(b22 - b12) - 1.5b2 + 0.5b1)/(b2 + b1 + 0.5)$	Global environment monitoring index	Pinty and Verstraete (1992)
WDVI[5,7,8]	(0.2b2 − b1)/(0.2b2 + b1)	Wide dynamic range vegetation index	Gitelson (2004)
MSI[6,7]	b6/b5	Moisture stress index	Rock et al. (1986)

Note: b1–b7 indicate MODIS bands. Predictor variables in bold are used in the gradient nearest neighbor (GNN) modes based on multi-month MODIS variables, and superscript numbers represent the monthly information of the MODIS variables.

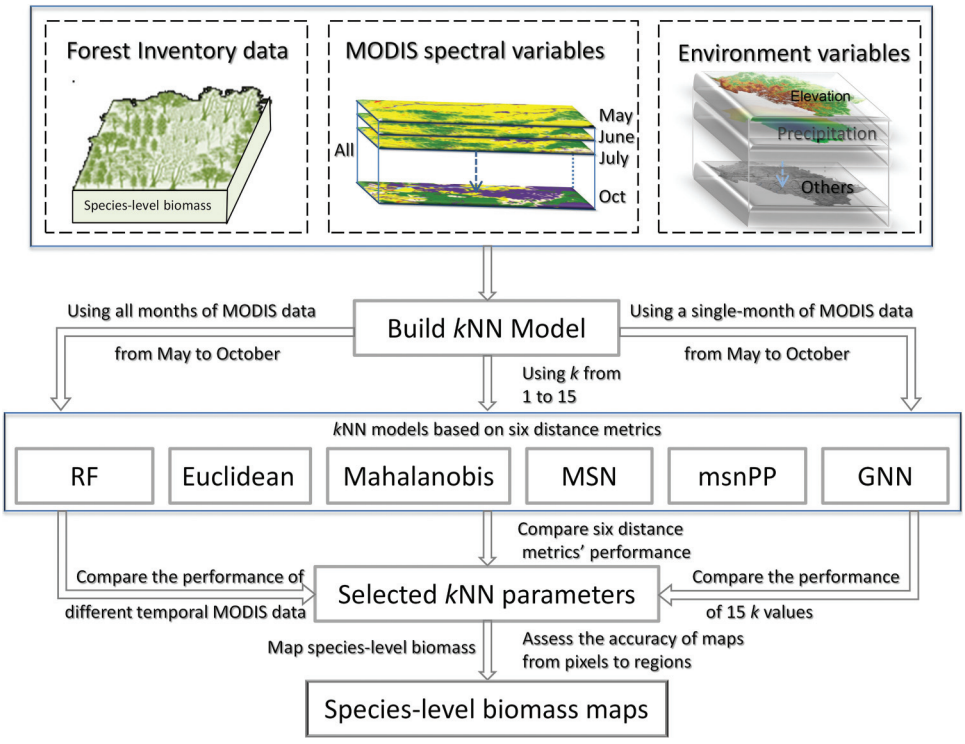
Table 2. Candidate predictor variables in this study.

Variable class and code	Definition
b1[8]	Band 1 (red, 620–670 nm)
b2[6,8,9]	Band 2 (short-wave near-infrared, 841–876 nm)
b3[6,9]	Band 3 (blue, 459–479 nm)
b4[5]	Band 4 (green, 545–565 nm)
b5[5,7,9]	Band 5 (long-wave near-infrared, 1230–1250 nm)
b6[5]	Band 6 (long-wave near-infrared, 1628–1652 nm)
b7[5]	Band 7 (long-wave near-infrared, 2105–2155 nm)
Vegetation indices	See Table 1
PRE, TEM	Mean annual precipitation and temperature from 1982 to 2009
ELEVATION, SLOPE, COSASP	Elevation (m); slope (°), aspect was derived from the SRTM DEM; COSASP is the cosine transformation value of the aspect
X	Coordinate *x* from each raster cell center (m)
Y	Coordinate *y* from each raster cell center (m)
SBULK	Bulk density of soil (g·cm^{-3})
SPH	pH of soil
SAND, CLAY, SILT, GRAVEL, SOC	Content (%) of sand, clay, silt, gravel, and soil organic carbon in the soil

Note: SRTM, Shuttle Radar Topography Mission; DEM, Digital Elevation Model. Variables in bold are used in the gradient nearest neighbor (GNN) modes based on multi-month MODIS variables, and superscript numbers indicate the monthly information of the MODIS variables.

phy Mission digital elevation model (90 m spatial resolution), provided by the International Scientific and Technical Data Mirror Site, Computer Network Information Center, Chinese Academy of Sciences (http://www.gscloud.cn/). Climate variables were collected from the National Meteorological Center of China and interpolated into a 1 km resolution map by Mao et al. (2012). Geospatial location data were in the form of *x* and *y* coordinates of each raster cell center. Soil data were extracted from Intergovernmental Panel on Climate Change (IPCC) default soil classes derived from the Harmonized World Soil Database (Batjes 2009). All of the environmental variables except the 90 m topographic variables were resampled to 250 m pixel resolution using a nearest-

Fig. 3. Flow diagram of species-level biomass imputation. kNN, *k* nearest neighbor; RF, Random Forest; MSN, most similar neighbor; msnPP, most similar neighbor with canonical correlation analysis; GNN, gradient nearest neighbor; MODIS, moderate resolution imaging spectroradiometer. [Colour online.]



neighbor algorithm to match the MODIS data resolution. The 90 m topographic variables were resampled to 250 m pixel resolution using bilinear interpolation.

2.3. The species-level biomass imputation approach

Forest stand polygons were used as the unit of observations for our imputation models. For both the MODIS composites and environmental predictors, we extracted the mean values of the raster cells with more than 50% of the pixel area covered by the stand polygon. The species-level biomass of each polygon was predicted by using these mean values in the *k*NN models. To reduce the effects of disturbance events occurring between MODIS acquisition dates and inventory dates, disturbance information derived from Landsat data from 1997 to 2001 with the vegetation change tracker (Huang et al. 2010) was used to identify disturbed polygons, and polygons that contained disturbed pixels (more than 50% pixel area was in the polygon) were excluded. This left the undisturbed polygons (98% or 7481 out of 7635 polygons) for further training and testing in this study. The stand polygons (Fig. 2*b*) were randomly split into training and testing data (training-to-testing ratio of 7:3). This process was replicated 20 times to reduce the effects of the sampling variability from a single split.

Formally, the equation of *k*NN is as follows (McRoberts 2012):

$$(1) \qquad \tilde{y}_i = \sum_{j=1}^{k} w_{ij} y_j^i$$

where y_j^i is the set of response variable observations for the k reference set elements that are nearest to the ith target in a feature space defined by some distance metrics, w_{ij} is the weight of the jth nearest neighbor reference, and $\sum_{j=1}^{k} w_{ij} = 1$. The inverse distance weight was selected to weight the k nearest neighbor reference elements (Crookston and Finley 2015), which is defined as follows:

$$(2) \qquad w_{ij} = \frac{1/(1 + d_{ij})}{\sum_{j=1}^{k} [1/(1 + d_{ij})]}$$

where d_{ij} is the distance of the jth nearest neighbor reference to the ith target. To determine whether different k values, distance metrics, and MODIS multi-temporal data improved the accuracy of the species-level biomass, we first built 630 *k*NN models based on six distance metrics, 15 k values, and seven sets of predictor variables (Fig. 3). The seven sets of predictor variables were seven sets of the monthly MODIS data combined with the environment variables (Table 2), separately. The distance metrics used in this study are listed in Table 3. Euclidean and Mahalanobis distances only depend on the predictor variables, and the remaining four distance metrics depend on the correlations between the response variables and predictor variables. We removed redundant predictor variables for the GNN distance metric by using forward stepwise canonical correspondence analysis (CCA) to keep significant variables ($p < 0.01$) using the vegan package in R (Oksanen et al. 2009) following methods used by Ohmann and Gregory (2002). For models using other distance metrics, we kept all predictor variables. We calculated the generalized root mean square distance (GRMSD) (Crookston and Finley 2015), the mean deviation (MD), the variance ratio (VR) (Powell et al. 2010), and the multivariate goodness of fit criterion T (McRoberts 2012) for all of the *k*NN models using 20 replicates of the testing data. T is defined as

$$(3) \qquad T = \sum_{p=1}^{m} w_p T_p^2$$

where p indexes response variables, w_p is the weight of the pth response variable (here all set equally to 1/m, where m is the num-

Table 3. Description of six distance metrics in this study defined in the yaImpute package of R (Crookston and Finley 2015).

Distance metrics	Description
RF	Distance in a Random Forest is calculated as one minus the proportion of classification trees where a target observation is in the same terminal node as a reference observation
Euclidean	Euclidean distance is computed in a multivariate predictor variable space normalized by subtracting the mean and dividing by the standard deviation, for each predictor variable
Mahalanobis	Mahalanobis distance is the dimensional components of Euclidean distance weighted by the inverse of the sample variance–covariance matrix (Mahalanobis 1936)
MSN	Most similar neighbor (Moeur and Stage 1995) distance is computed in a projected canonical space based on the canonical correlation analysis
msnPP	Like MSN, except that the canonical correlation is computed using projection pursuit from the ccaPP R package (Alfons 2013)
GNN	Gradient nearest neighbor (Ohmann and Gregory 2002) distance is computed using a projected ordination of predictors based on canonical correspondence analysis

ber of response variables), and T_p^2 is the variance explained for the pth response variable by the kNN prediction. The mean values of these measures for the 20 replicates were used as measures for comparing model performances with different distance metrics, k values, and spectral variable combinations. In this study, GRMSD was the root mean square distance between imputed and observed values in an orthogonal multivariate space defined by biomass values of the eight tree species. MD was calculated as the difference between the mean imputed and observed total AGB. VR was calculated as the standard deviation of imputed total AGB divided by the standard deviation of observed total AGB; VR values close to 1 indicate good model performance. In the multivariate goodness of fit criterion (T), p represented one of eight tree species, T_p^2 was the fractional amount of variance in response variable p explained by the kNN prediction, w_p was the weight of the pth species' biomass, which was the percent biomass of the pth species against the total AGB based on the observed value. The GRMSD function is

$$\text{GRMSD} = \frac{\sum_{i=1}^{n} D_i}{n} \quad (4)$$

where the D_i is the scaled root-mean-square distance of the ith response variable:

$$D_i = \sqrt{\frac{\sum_{j=1}^{nr} (O_{ij} - P_{ij})^2}{\text{nr}}} \quad (5)$$

where O_{ij} is the jth row and ith column element of the scaled observed response variables matrix **O**, and P_{ij} is the jth row and ith column element of the scaled predicted response variables matrix **P**, and nr is the row number of matrix **O**. The formulas of **P** and **O** are separately defined as follows:

$$\mathbf{O} = \mathbf{O}'\%*\%\mathbf{q} \quad (6)$$

$$\mathbf{P} = \mathbf{P}'\%*\%\mathbf{q} \quad (7)$$

where $\mathbf{O}'$ is the original observed responsible variables matrix, $\mathbf{P}'$ is the predicted variables matrix, %*% is the matrix product operator, and **q** is the weighted matrix defined as follows:

$$\mathbf{q} = \text{solve}(\text{chol}(\text{cov}(\mathbf{O}'))) \quad (8)$$

where cov() is the covariance matrix function, chol() is the function to compute the Choleski factorization, and solve() is the function to return the orthogonal matrix of a matrix.

After the best imputation method, MODIS predictor variables, and k values were determined, we computed the nearest-neighbor distance between all pixels and the corresponding nearest reference observations and imputed species-level biomass for all pixels. This imputation process was finished using the yaImpute package in R (Crookston and Finley 2015).

2.4. Accuracy assessment

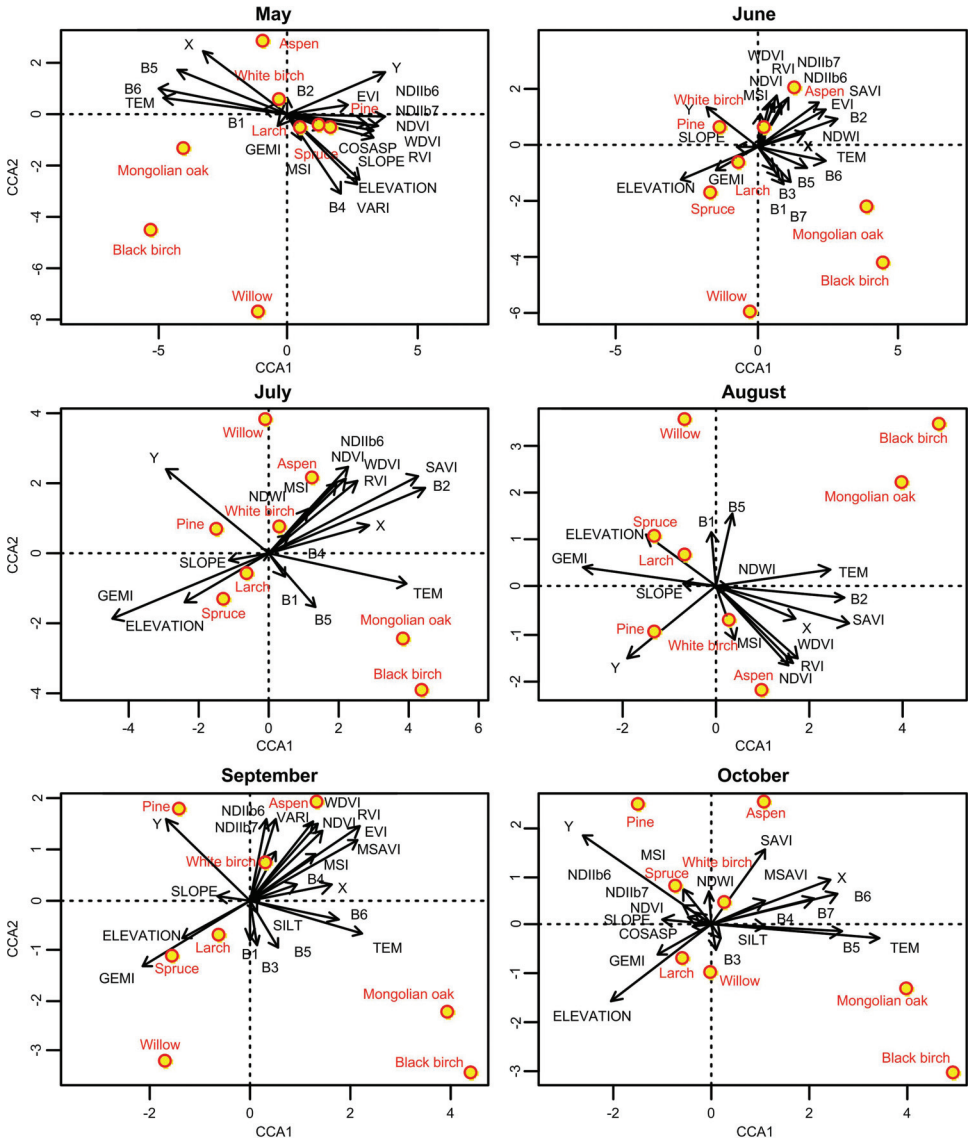
The accuracy of the species-level biomass maps in this study was assessed using a variety of methods designed to identify various measures of species biomass and composition at the pixel, ecoregion, and regional scales. All analyses were done using the raster (Hijmans 2015), vegan (Oksanen et al. 2017), yaImpute (Crookston and Finley 2008; Hudak et al. 2008), and base packages in R (R Core Team 2013). The training and testing data for the maps of species-level biomass were from one splitting of forest inventory data (training-to-testing ratio of 7:3).

At the pixel scale, the square of the Pearson correlation (R^2), mean deviation (MD), and root mean square deviation (RMSD) between observed and predicted species-level biomass were calculated using the testing data. At the same time, the difference between observed and imputed species-level biomass distributions of each species was quantified using empirical cumulative distribution functions (ECDFs) and the Kolmogorov–Smirnov (KS) statistic (Lopes et al. 2007; Riemann et al. 2010). The KS statistic makes no assumptions about the distribution of data, is independent of scale changes, and is defined as the maximum distance between two empirical distribution functions (Riemann et al. 2010). We also calculated the nearest-neighbor distance (Fig. 2*c*) between each 250 m pixel and its corresponding reference polygon based on the spatial predictors. Nearest-neighbor distance is an indicator of model uncertainty, with high values indicating high uncertainty in the predictions (Crookston and Finley 2008).

We averaged both the observed and imputed species-level biomass of the testing data at the ecoregion scale. In addition to the accuracy metrics used at the pixel scale, Bray–Curtis community dissimilarity (BC) (Bray and Curtis 1957) and Spearman rank order correlation between the field inventory data and the corresponding imputed pixels were calculated as a measure of imputation quality for each ecoregion. BC values range between 0 and 1, with 0 being the most similar and 1 being the most dissimilar (Bray and Curtis 1957). Spearman rank order correlation measures how well the order of species abundance was represented by the imputed map in each ecoregion. At the regional scale, we also calculated BC and Spearman rank order correlation. Additionally, the imputed and observed mean species-level biomass and its standard deviation were compared at the regional scale.

Metrics such as R^2 and RMSD provide an assessment of the overall accuracy of the total or single-species biomass maps; however, additional metrics are required to fully assess the accuracy of imputation results. For instance, the KS metric provides a robust

Fig. 4. Biplots for canonical correspondence analysis (CCA) axes 1 and 2 showing selected significant predictor variables (black arrows) for each month by stepwise CCA method and species centroids (circles). Arrow length and position of the arrowhead indicate the correlation between the explanatory variable and the CCA axes, and smaller angles between arrows indicate stronger correlations between variables. Species scores are linear combinations of plot scores. See Tables 1 and 2 for predictor variables. [Colour online.]



assessment of the differences in the distributions of the observed and imputed species-level biomass, and it is independent of changes in scale (Lopes et al. 2007). The BC dissimilarity and Spearman rank order correlation are effective measurements of forest composition and dominant species abundance between observed and imputed data (Duveneck et al. 2015; Ohmann et al. 2011).

3. Results

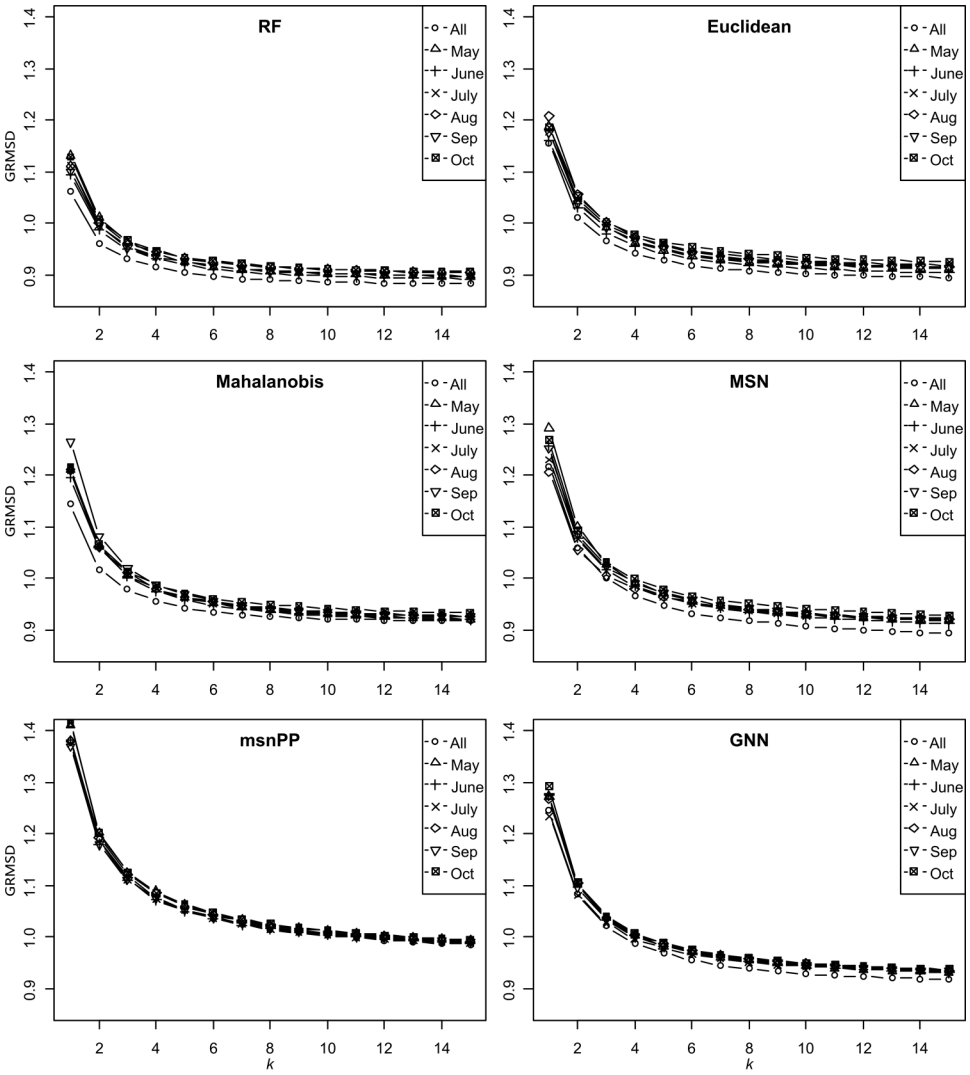
3.1. Importance of predictor variables

The most important variables selected by the stepwise CCA method were similar between models using MODIS composites from different months, although the importance scores of predictor variables varied across months (Fig. 4). The MODIS predictors tended to have high importance values, although many of the environmental variables had importance values as high as the MODIS predictors. For example, mean annual temperature had a strong positive correlation with the biomass of Mongolian oak and black birch and had the opposite correlation with pine. Elevation, slope, and the *x* coordinate also had strong effects on the biomass of larch, white birch, aspen, and spruce (Fig. 4); however, the soil-related variables provided almost no important contributions. Depending on the month used in the GNN model, different single-month MODIS predictor variables were selected in the *k*NN models due to collinearity. Most vegetation indices except GEMI from June to September showed high positive correlation with the biomass of white birch and aspen and negative correlation with the biomass of larch and spruce. The MODIS reflectance bands b2, b5, b6, and b7 also had high importance scores.

3.2. Performance of different *k*NN models and model parameter selection

The performance of the *k*NN models varied by distance metric, *k* value, and which MODIS spectral variables were included (Fig. 5; Appendix Figs. A1–A3). Models using RF had the best performance across the three measures of GRMSD, VR, and *T* (smallest GRMSD

Fig. 5. Multivariate generalized root mean square distance (GRMSD) curves vs. *k* for different *k* values and nearest-neighbor imputation models using different distance metrics based on combinations of environmental variables and seven sets of MODIS composite variables (all months or a single month of MODIS spectral variables from May to October in the legend). GRMSD is the scaled root mean square distance between the predicted and observed values for the eight response variables in an orthogonal multivariate space. Mean values are from 20 replicates; standard errors are very small and are not included for clarity. RF, Random Forest; MSN, most similar neighbor in canonical correlation space; msnPP, most similar neighbor computed using projection pursuit; GNN, gradient nearest neighbor.



and largest VR and *T* for the most predictor variables and *k* value combinations), but the worst performance (i.e., the greatest deviation from zero) for MD (Appendix Fig. A1). All values of MD ranged from –1.5 to 1.5 t·ha^{-1}, and the variation of the MD was small compared with the mean AGB of the forest inventory data (61.2 t·ha^{-1}) according to different distance metrics. Differences in GRMSD among the different distance metrics were also small (<10%) except for msnPP (Fig. 5).

Using multi-month instead of single-month MODIS composites (June) only slightly improved the accuracy of *k*NN models, with the mean of GRMSD reduced by 0.0003 to 0.024 (Fig. 5) and the mean of *T* increased by 0.011 to 0.025 (Appendix Fig. A3). Almost all of the *k*NN models showed the best performance when $k = 15$ for GRMSD and *T* but the worst performance for VR (the highest *T* and the lowest GRMSD and VR). The difference in GRMSD, MD, and *T* among the RF models using the single-month MODIS composite variables for June when $k > 6$ was essentially negligible (Fig. 5; Appendix Fig. A3), and VR only had 16.5% difference between $k = 1$ and $k = 6$. Therefore, we selected the RF distance metric, single-month MODIS composite variables for June, and $k = 6$ as the best model and used it for imputing species-level biomass.

3.3. Map accuracy assessment

3.3.1. *Accuracy of total AGB*

From the pixel scale to the ecoregion scale, accuracy improved substantially, with R^2 increasing from 0.60 to 0.91 and RMSD decreasing from 12.75 to 2.52 t·ha^{-1}, respectively (Figs. 6*a* and 6*c*, $p < 0.05$). Although the KS distance was slightly higher at the ecoregion scale (KS distance = 0.11) than at the pixel scale, a higher *p* value ($p = 0.95$) for KS distance at the ecoregion scale indicated that the imputed and observed AGB ECDFs had become more similar with increasing scale (Figs. 6*b* and 6*d*). For the regional scale, the MD between imputed and observed mean AGB was –0.43 t·ha^{-1} (Table 4). In addition, pixels with high distance values either had lower biomass or were located in non-forest areas (Fig. 2*c*). The imputed total AGB showed underestimation for the forests with high biomass and overestimation for the forests with low biomass (Fig. 6*a*).

Fig. 6. (*a* and *c*) Scatterplots and (*b* and *d*) cumulative distribution functions of imputed vs. observed total aboveground biomass (AGB; t·ha^{-1}) based on the testing data. The top two plots (*a* and *b*) are at the pixel scale and the bottom two plots (*c* and *d*) are at the ecoregion scale. The dotted line is the 1:1 line; the dashed line is the geometric mean functional regression line. RMSD (t·ha^{-1}), root-mean square deviation; KS, Kolmogorov–Smirnov statistic.

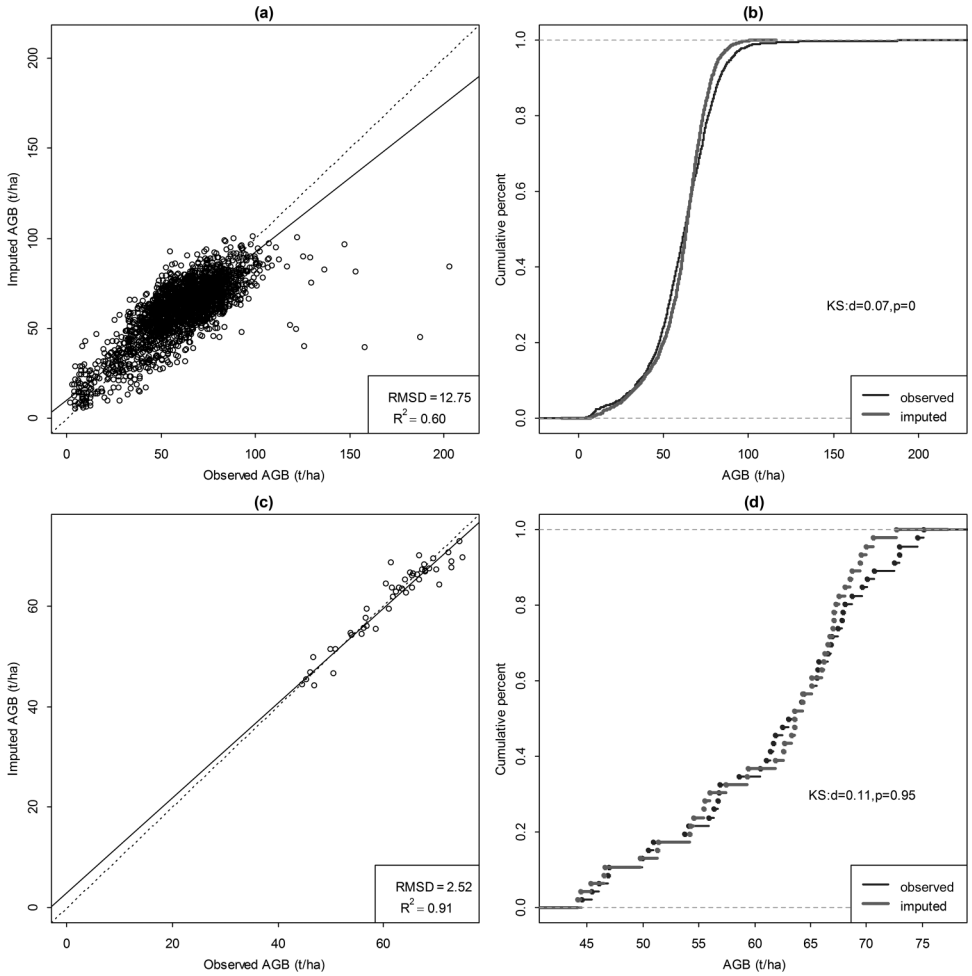


Table 4. Imputed and observed mean species-level biomass and their standard deviations, sample size, and mean deviation (MD) based on testing data at the regional scale.

	Mean biomass (t×ha^{-1})		Standard deviation (t×ha^{-1})		Sample size		
Species	Imputed	Observed	Imputed	Observed	Testing	Training	MD (t×ha^{-1})
Larch	30.34	29.44	17.36	21.86	1963	4568	0.90
White birch	23.73	23.67	10.84	16.55	2054	4842	0.06
Pine	1.37	1.83	3.24	7.40	253	602	−0.46
Aspen	3.18	3.66	5.21	8.44	635	1492	−0.48
Willow	0.02	0.10	0.31	1.84	10	22	−0.08
Spruce	0.11	0.20	0.80	1.72	44	97	−0.09
Mongolian oak	1.51	1.73	3.74	6.31	297	744	−0.22
Black birch	0.52	0.58	1.66	2.46	177	444	−0.06
Total AGB	60.78	61.21	16.44	20.01	2244	5237	−0.43

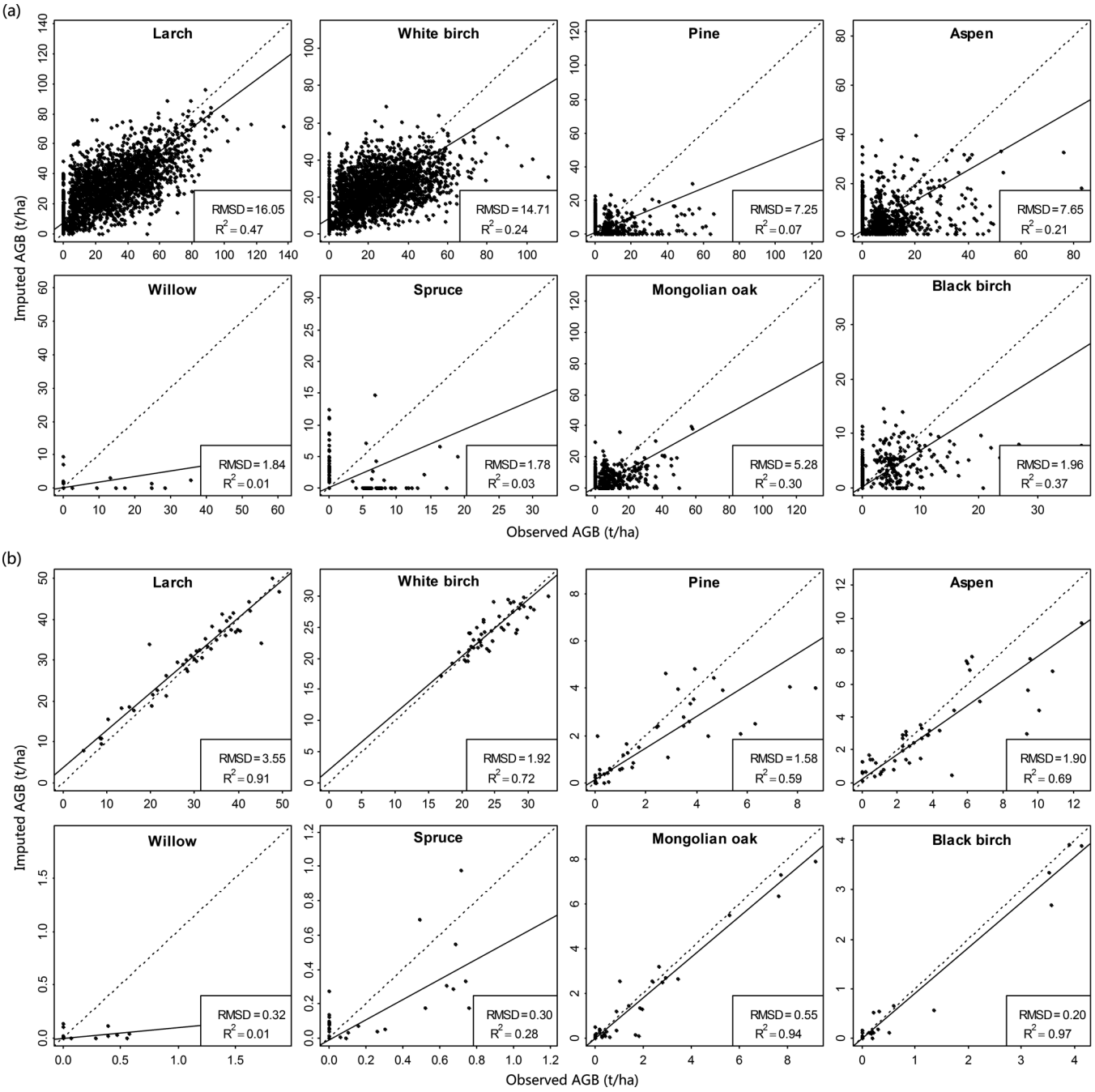
Note: AGB, aboveground biomass.

3.3.2. Accuracy of species-level biomass

Dominant tree species (larch and white birch) had relatively higher accuracy compared with other species at the pixel scale, and the results for most species for R^2, RMSD, and KS metrics showed improvement with increasing scales except willow (Figs. 7 and 8). The bias structure for dominant species (larch and white birch) indicated that the *k*NN imputation resulted in underestimation at high biomass levels and overestimation at low biomass levels (Figs. 7 and 8). The KS metrics showed that ECDFs between the observed and imputed species-level biomass at the pixel scale for almost all of the species except willow and spruce were significantly different ($p < 0.05$; Fig. 8*a*); however, at the ecoregion scale, the ECDFs were similar for all species and the KS metrics were not significant ($p > 0.05$; Fig. 8*b*).

The BC dissimilarity was close to 0 (mean value = 0.057, standard deviation = 0.032) and Spearman rank correlation was close to 1 (mean value = 0.930, standard deviation = 0.069) for most ecoregions, indicating that species composition was well repre-

Fig. 7. Scatterplots of imputed vs. forest inventory aboveground biomass (AGB) about eight species at (*a*) the pixel scale and (*b*) the ecoregion scale based on testing data. The dotted line is the 1:1 line; the dashed line is the geometric mean functional regression line. RMSD, root mean square deviation.



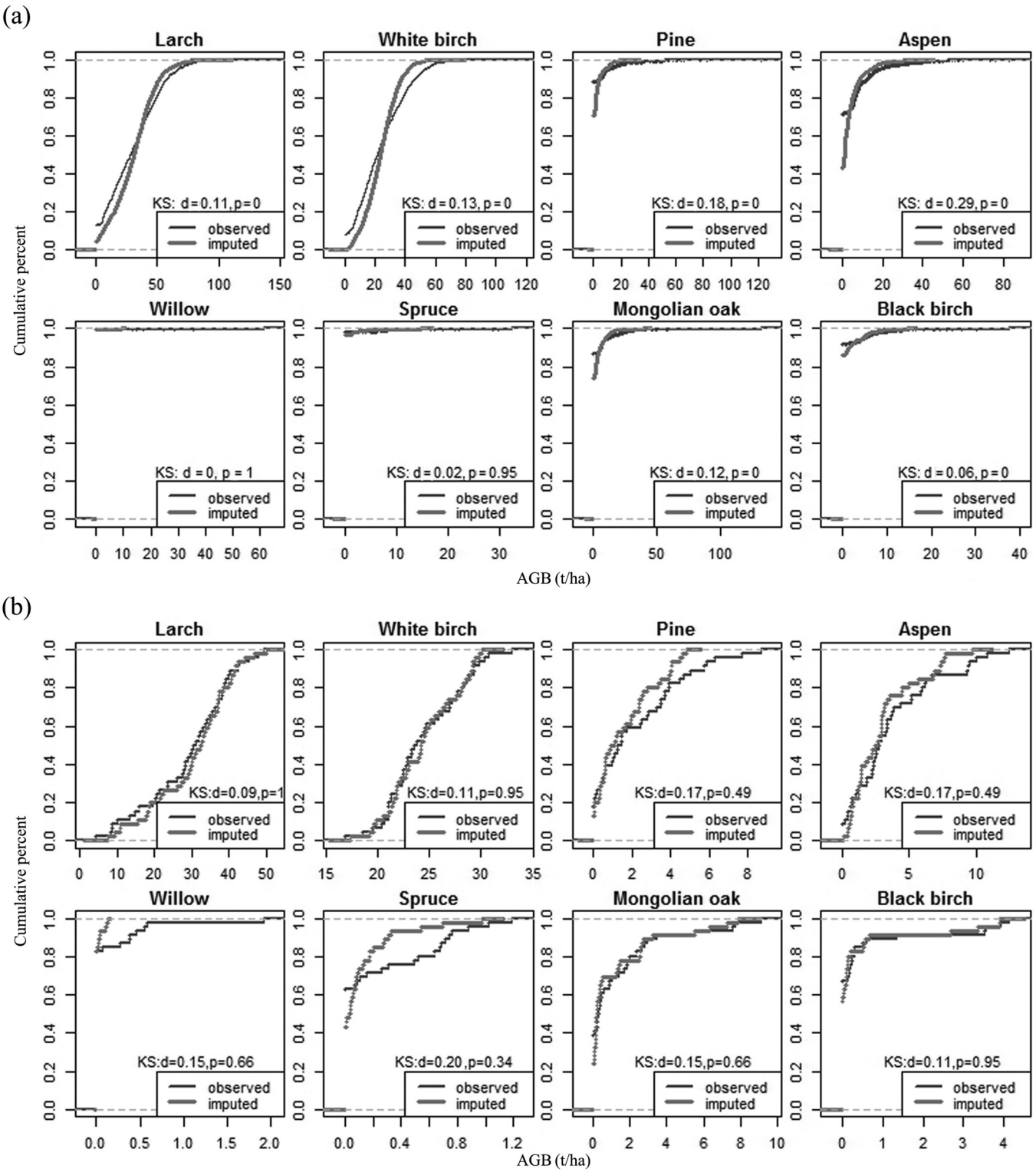
sented by the imputation map at the ecoregion scale (Fig. 9). The BC dissimilarity (0.019) and Spearman rank order correlation (0.976) also showed that our imputed results represent forest composition well at the regional scale. In addition, BC dissimilarity for each ecoregion was strongly related to the number of test polygons (Fig. 9).

3.4. Species-level biomass map

The total AGB across all species was mapped by summing the species-level biomass within each pixel (Fig. 10); the imputed mean AGB value across the study area was 57.21 t·ha^{-1}. Larch was the most dominant species with the largest biomass (28.4 t·ha^{-1}), followed by white birch (21.94 t·ha^{-1}). Aspen was also widely distributed over the study area, although it had lower biomass density (2.99 t·ha^{-1}) than larch and white birch. Black birch and Mongolian oak were imputed mainly in the southern part of the study area, whereas pine was mainly imputed in the northern part. Willow was mostly imputed along rivers, which is consistent with the inventory data. Spruce was sparse and scattered in the study area.

The locations of imputed pixels with low biomass were consistent with the areas influenced by fire and harvesting. For example, the area burned by the Black Dragon fire in 1987 in the northern part of our study area had lower total AGB overall. Other pixels with low biomass were mainly located in areas of relatively low elevation where they are most likely to be affected by human activities such as harvesting (Figs. 1 and 10); conversely areas with high biomass tended to be in high elevation areas (Fig. 10*b*).

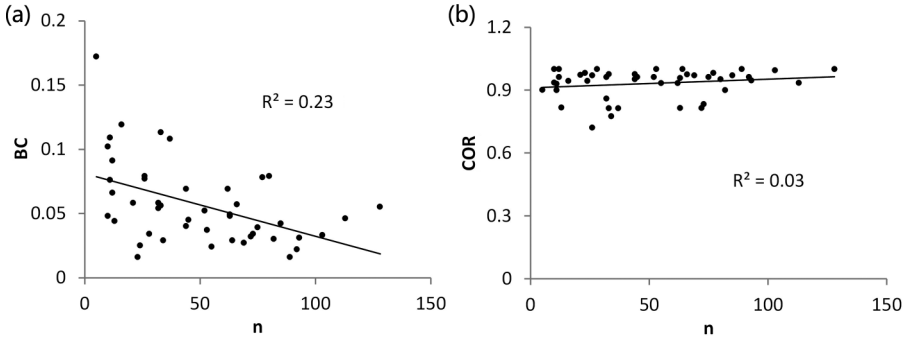
Fig. 8. Cumulative distribution functions of imputed vs. forest inventory biomass (t·ha^{-1}) for eight species separately at (*a*) the pixel scale and (*b*) the ecoregion scale based on testing data. The metrics of Kolmogorov–Smirnov (KS) statistic are also shown.



4. Discussion

We successfully mapped tree species level biomass in the Great Xing'an Mountains of northeastern China using MODIS spectral reflectance, vegetation indices, and additional environmental variables with a RF-based *k*NN imputation method. In our study area, several previous studies had mapped AGB, aboveground forest carbon stocks, and timber volume using either Landsat (Li 2010; Qi and Li 2015; Wang et al. 2014) or MODIS imagery (Cartus et al. 2011; Chi et al. 2015; Su et al. 2016; Zhang et al. 2014); however, this study is the first to produce a coherent set of species-level biomass across the Great Xing'an Mountains' well-inventoried forests. Although previous studies have imputed forest composition

Fig. 9. The number of testing polygons (*n*) vs. Bray–Curtis dissimilarity (BC) and Spearman rank order correlation coefficient (COR) in each ecoregion.



over large areas of the United States using Landsat data and GNN methods (Ohmann and Gregory 2002) or MODIS (Duveneck et al. 2015; Wilson et al. 2012), questions about distance metrics, multi-temporal spectral variables, and *k* values were explored in our study.

4.1. Selection of distance metrics and *k* value

Selecting the type of *k*NN distance metric and which predictor variables to use for imputation can be challenging because of the availability of a large number of distance metrics and a nearly endless number of potential predictors. Our study demonstrated that the RF distance metric had the highest accuracy when imputing species-level biomass using MODIS spectral variables. This finding is similar to the results that Hudak et al. (2008) reported using lidar metrics. The advantage of using the RF distance metric lies in its efficiency for processing high-dimensional data, flexibility to handle highly correlated predictors, and high levels of prediction accuracy (Prasad et al. 2006), which are largely because RF-based predictions are minimally affected by the inclusion of unimportant variables (Hastie et al. 2009). Although other methods such as stepwise CCA could be used to select the important variables from large sets of predictor variables, we found that using stepwise CCA as done in Ohmann and Gregory (2002) required fitting more models and thus required more time than the RF model when the number of predictor variables was large. This was mainly because RF variable selection is more automated than stepwise CCA; however, the potential benefit of using stepwise CCA to select the predictor variables might be that collinear predictor variables are removed and the relationships between the remaining predictor variables and the response variables can be more readily explained. After the RF distance metric, MSN-based imputation had the second highest accuracy and executed more efficiently than the RF-based *k*NN model. Given this, it might be worth considering use of the MSN distance metric for large areas instead of the RF distance metric to reduce computation times.

The similar trend with increasing *k* in our study for GRMSD, *T*, and MD was observed in contrast to the findings of previous studies (Beaudoin et al. 2014; McRoberts 2012; Muinonen et al. 2001; Wilson et al. 2012). Selecting a *k* value is a trade-off between the covariance structure and the imputation accuracy (Muinonen et al. 2001; McRoberts 2012). By selecting $k = 6$ in our study, covariance structure was kept as stable as possible on the basis of assuring the imputation accuracy.

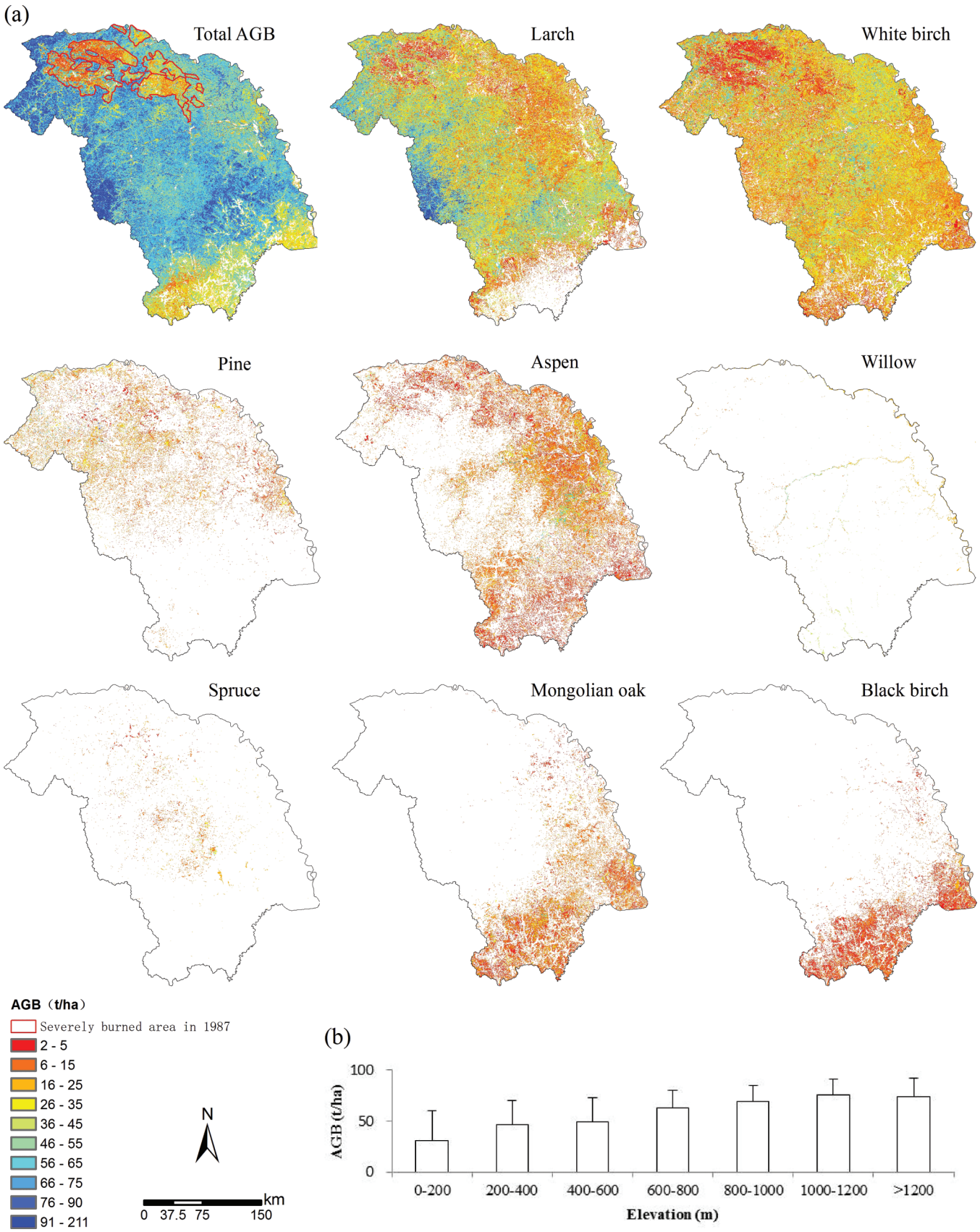
4.2. Performance of single- and multi-month MODIS composite predictors

Previous studies reported that multi-temporal spectral information improved the prediction accuracy of AGB by reducing limitations related to the saturation of spectral reflectance with forest biomass — at some point, spectral sensors are not responsive to increases in biomass beyond a certain threshold (Gómez et al. 2014; Powell et al. 2010; Zhu and Liu 2015). However, in our study, no appreciable improvement in performance for species-level biomass imputation was observed using multi-month MODIS composite variables compared with using single-month MODIS composite variables (Fig. 5). There could be several reasons for this. First, the biomass of species composition was considered in the species-level biomass accuracy assessment compared with the total AGB accuracy assessment. Multi-month MODIS composites could improve AGB estimation by providing additional information for the forests that have similar MODIS reflectance values at the peak of the growing season and different reflectance values at the start of the growing season. Such information also could divide one forest composition type into many others. In addition, MODIS composite variables for specific months may contain important information for distinguishing the tree species in our study area. For example, MODIS composite variables from June showed better performance than other monthly spectral variables for all of the distance metrics (Fig. 5). This may be because most species in the Great Xing'an Mountains leaf out in June (Yu and Zhuang 2006), and species-related differences in June spectral reflectance are good indicators of species-level differences in AGB. Finally, information derived from multi-month MODIS composite images could describe temporal dependence and have the ability to improve species-level biomass imputation. For example, Wilson et al. (2012) imputed tree species across the eastern half of the contiguous United States utilizing parameters from a harmonic regression fit to MODIS monthly composites.

Generally, lidar is currently considered to generate the most accurate data for estimation of forest structure (Pflugmacher et al. 2012). However, GRMSDs obtained in our study are comparable with the values obtained with lidar (Hudak et al. 2008). The following two reasons may reduce the accuracy gap between lidar and MODIS data in estimating forest species-level biomass. First, lidar metrics have an obvious advantage for estimating forest aboveground biomass compared with multispectral satellite images but are less effective for mapping tree species composition (Zald et al. 2014). Secondly, mid-seral, secondary forests are the main components of our study area, and such forests are more sensitive to the optical spectral reflectance because their spectral reflectance often cannot reach the saturation point (Lu 2006). This may indicate that it is more efficient to use MODIS or Landsat data for imputing species-level attributes of mid-seral, secondary forests over large areas rather than lidar data.

In Chinese boreal forests, the large area and constantly changing forest structure and biomass due to high-frequency disturbances presented a unique challenge for choosing predictor variables of *k*NN imputation, which need high temporal resolution and wide availability (Wu et al. 2013; Xu 1998). MODIS data, with higher temporal resolution and broader spatial coverage than Landsat data, are suitable for applying *k*NN to deriving forest attributes (Wilson et al. 2012). Our imputed results showed that using single-month MODIS data from June can produce accuracy comparable with using multi-month MODIS data for species-level

Fig. 10. (*a*) Maps of total aboveground biomass (AGB) and species-level biomass; (*b*) the mean values and standard errors of the imputed total AGB at different elevation gradients. [Colour online.]



biomass imputation. Our approach also captures forest disturbance and is an efficient way to impute forest species-level biomass over broad spatial extents.

4.3. Distribution of species-level biomass with environmental gradient

The environmental variables captured resource gradients that might not be well represented in the remote sensing data and were as important as the MODIS-derived predictor variables in our species-level biomass imputation (Fig. 4). The spatial pattern of species-level biomass in our imputed results was partially consistent with tree species' environmental niches. For example, black birch and Mongolian oak are limited by temperature and are abundant in the south, while pine is adapted to the cold, dry environments and is abundant in the north. The biomasses of white birch and aspen showed some negative correlation with the biomass of larch (Fig. 4). This may be because white birch and aspen are early successional species and often are the first to recolonize disturbed larch forests but are gradually replaced by larch due to succession.

The spatial pattern of species-level biomass in our imputed results also reflected the influence from forest harvest and disturbance. Overall, the total woody AGB of virgin larch forests should initially increase and then decrease along the elevation gradient according to the description by Xu (1998) because of permafrost due to the low soil temperatures and poor drainage in the areas with low elevation and strong wind and ultraviolet radiation on mountaintops; however, a slightly increasing trend in AGB along the elevation gradient was observed in our results, similar to the results reported in Wang et al. (2014). This may be because midseral, secondary forests have become the main components of the forest landscape except in high-elevation areas in the natural reserves due to long-term harvesting activities (Luo et al. 2014). In addition, Feng (1999) reported that the mean biomass of mature virgin larch forests in our study area should be nearly 161 $t \cdot ha^{-1}$, whereas the mean AGB of our imputed results was only 57.21 $t \cdot ha^{-1}$, similar to the results reported in Wang et al. (2014). This might indicate that forests in our study area have great potential for additional growth.

4.4. Imputation accuracy and limitations

Our imputation results have comparable accuracy with other mapping projects that use traditional field plots and satellite imagery. For example, our imputed AGB, estimated by merging species-level biomass in each pixel, was equal to or more accurate ($R^2 = 0.60$) than AGB predictions for northeastern China and Canada ($R^2 = 0.16–0.62$) that used forest inventory data and MODIS or Landsat imagery (Beaudoin et al. 2014; Chi et al. 2015; Wang et al. 2014; Zhang et al. 2014). Likewise, our estimates of species-level biomass and composition have accuracy comparable with recent forest composition imputation mapping in New England at the ecoregion scale (Duveneck et al. 2015). Although these comparisons with other studies are partially confounded by different statistical methods, inventory protocols, and forest types, they indicate that our imputation results have a relatively reliable accuracy.

At the pixel scale, the total AGB (Figs. 6*a* and 6*b*) and species-level biomass of larch and white birch (Fig. 7*a*) had relatively high accuracy compared with other species; however, the bias structure was obvious for the species-level biomass imputation, with underestimation in areas with high biomass and overestimation in areas with lower biomass potentially caused by MODIS spectral saturation in forests with high biomass. A large number of pixels also were imputed with unrealistic species biomass (Fig. 7*a*, a large number of zero values in the *X* and *Y* axes) at the pixel level. This might indicate a limitation of multispectral data for distinguishing tree species (Martin et al. 1998). Hyperspectral data may help to improve the imputation results, but these types of data are difficult to collect over large areas. Additionally, imputation accuracy could have been influenced by the coarse spatial resolution of our data — increasing the possibility for mismatch between the pixels and forest inventory data, especially along polygon boundaries.

At the ecoregion scale, the accuracy metrics all showed that our results had reasonable accuracy (Figs. 6–9). Our imputed results are relatively accurate and can be used for many analyses, e.g., assessing the effects of environment variables on forest composition (Liang et al. 2014), monitoring carbon stocks (He 2008; Scheller et al. 2007), and quantifying the impacts of forest management (Wu et al. 2013). In some ecoregions, relatively large dissimilarities between observed and imputed forest composition were found, especially where the density of inventory data was low (Fig. 9). It is possible that the actual species distribution in these ecoregions was not well represented by the inventory data. Our imputed results were more accurate at the regional scale than at the pixel and ecoregion scales. The increasing accuracy with increasing scale is consistent with results reported by others (e.g., Wilson et al. 2012). This may be because ecoregions in our study captured species-level biomass gradients associated with environmental conditions better than individual pixels, although the effects of spatial aggregation within each ecoregion also could be important.

The accuracy assessment of our imputed results at different scales can provide useful information for different applications. For example, forest landscape change models need species-level attributes at the pixel scale to initialize forest conditions (Duveneck et al. 2015), whereas ecosystem process research may require forest species-level attributes at the ecoregion or regional scale (Canadell and Raupach 2008; Führer 2000). The lower accuracy of some rare species (e.g., willow and spruce) at different scales might be caused by their very limited distributions. First, a limited number of samples included spruce and willow in our inventory data (Table 4), and it was difficult to calibrate the models using these samples. Second, it was difficult to capture uncommon species using the predictor variables with 250 m resolution, and predictor variables with finer spatial resolution may be needed to improve the imputed accuracy for such species.

5. Conclusions

In summary, the nonparametric RF-based *k*NN method integrating multispectral variables derived from June MODIS data and environmental variables with forest inventory data was used to map the species-level biomass across the forests of the Great Xing'an Mountains of northeastern China. The RF distance metric was slightly superior to others assessed in this study for imputing forest species-level biomass over large areas. Using multi-month MODIS predictor variables did not improve species-level biomass imputation compared with using single-month MODIS data for June. Using MODIS data to impute species-level attributes can produce accuracy comparable with using lidar data for young and middle-aged boreal forests. Imputed biomass for the dominant species (larch and white birch) had relatively good accuracy at the pixel, ecoregion, and regional scales, although other associated species had poorer accuracy at the pixel scale. The distribution of our imputed species-level biomass also captured the effects of disturbances such as fire and harvest, and the method could be used for broad-scale monitoring of changes in biomass over time.

Acknowledgements

The study was funded by the National Key Research and Development Project (2017YFA0604402 and 2016YFA0602301) and the Natural Science Foundation of China (31570461 and 41371199), and the Major State Basic Research Development Program of China (2016YFA0600804). Contributions from authors from the U.S. Geological Survey (USGS) were supported by the National Biologic

Carbon Sequestration Assessment Program under the USGS Climate and Land Use Mission Area. Andrew Hudak, Eugene Schweig, Randall Schumann, Janet Slate, and three anonymous reviewers provided many comments and suggestions during the preparation of this manuscript. We appreciate their input as it helped us strengthen this manuscript considerably.

References

Alfons, A. 2013. ccaPP: (Robust) canonical correlation analysis via projection pursuit [online]. R package version 0.3.0. Available from http://CRAN.R-project.org/package=ccaPP.

Batjes, N.H. 2009. IPCC default soil classes derived from the Harmonized World Soil Data Base (Ver. 1.1) [online]. Report 2009/02b, Carbon Benefits Project (CBP) and ISRIS – World Soil Information, Wageningen (with dataset). Available from http://heihedata.org.

Beaudoin, A., Bernier, P.Y., Guindon, L., Villemaire, P., Guo, X.J., Stinson, G., Bergeron, T., Magnussen, S., and Hall, R.J. 2014. Mapping attributes of Canada's forests at moderate resolution through *k*NN and MODIS imagery. Can. J. For. Res. **44**(5): 521–532. doi:10.1139/cjfr-2013-0401.

Bitterlich, W. 1948. Die Winkelzahlprobe. Allg. Forst- Holzwirtsch. Ztg. **59**: 4–5.

Blackard, J.A., Finco, M.V., Helmer, E.H., Holden, G.R., Hoppus, M.L., Jacobs, D.M., Lister, A.J., Moisen, G.G., Nelson, M.D., Riemann, R., et al. 2008. Mapping U.S. forest biomass using nationwide forest inventory data and moderate resolution information. Remote Sens. Environ. **112**(4): 1658–1677. doi:10.1016/j.rse.2007.08.021.

Bray, J.R., and Curtis, J.T. 1957. An ordination of the upland forest communities of southern Wisconsin. Ecol. Monogr. **27**(4): 325–349. doi:10.2307/1942268.

Breiman, L. 2001. Random forests. Mach. Learn. **45**(1): 5–32. doi:10.1023/A:1010933404324.

Canadell, J.G., and Raupach, M.R. 2008. Managing forests for climate change mitigation. Science, **320**(5882): 1456–1457. doi:10.1126/science.1155458. PMID:18556550.

Carlson, M., Wells, J., and Roberts, D. 2009. The carbon the world forgot: conserving the capacity of Canada's Boreal Forest region to mitigate and adapt to climate change. Boreal Songbird Initiative and Canadian Boreal Initiative, Seattle, Wash., and Ottawa, Ont.

Cartus, O., Santoro, M., Schmullius, C., and Li, Z. 2011. Large area forest stem volume mapping in the boreal zone using synergy of ERS-1/2 tandem coherence and MODIS vegetation continuous fields. Remote Sens. Environ. **115**(3): 931–943. doi:10.1016/j.rse.2010.12.003.

Chi, H., Sun, G., Huang, J., Guo, Z., Ni, W., and Fu, A. 2015. National forest aboveground biomass mapping from ICESat/GLAS data and MODIS imagery in China. Remote Sens. **7**(5): 5534–5564. doi:10.3390/rs70505534.

Crookston, N.L., and Finley, A.O. 2008. yaImpute: an R package for kNN imputation. J. Stat. Softw. **23**(10): 1–16. doi:10.18637/jss.v023.i10.

Crookston, N.L., and Finley, A.O. 2015. yaImpute: nearest neighbor observation imputation and evaluation tools [online]. R package version 1.0-26. Available from http://CRAN.R-project.org/package=yaImpute.

Duveneck, M.J., Thompson, J.R., and Wilson, B.T. 2015. An imputed forest composition map for New England screened by species range boundaries. For. Ecol. Manage. **347**: 107–115. doi:10.1016/j.foreco.2015.03.016.

Fang, J.-y., Wang, G.G., Liu, G.-h., and Xu, S.-l. 1998. Forest biomass of China: an estimate based on the biomass–volume relationship. Ecol. Appl. **8**(4): 1084–1091. doi:10.1890/1051-0761(1998)008[1084:FBOCAE]2.0.CO;2.

Feng, Z. 1999. The biomass and productivity of forest ecosystem of cool temperate forest. *In* The biomass and productivity of forest ecosystem in China. Science Press, Beijing. pp. 51–60.

Führer, E. 2000. Forest functions, ecosystem stability and management. For. Ecol. Manage. **132**(1): 29–38. doi:10.1016/S0378-1127(00)00377-7.

Gao, B.-c. 1996. NDWI — a normalized difference water index for remote sensing of vegetation liquid water from space. Remote Sens. Environ. **58**(3): 257–266. doi:10.1016/S0034-4257(96)00067-3.

Gitelson, A.A. 2004. Wide Dynamic Range Vegetation Index for remote quantification of biophysical characteristics of vegetation. J. Plant Physiol. **161**(2): 165–173. doi:10.1078/0176-1617-01176. PMID:15022830.

Gitelson, A.A., Kaufman, Y.J., Stark, R., and Rundquist, D. 2002. Novel algorithms for remote estimation of vegetation fraction. Remote Sens. Environ. **80**(1): 76–87. doi:10.1016/S0034-4257(01)00289-9.

Gómez, C., White, J.C., Wulder, M.A., and Alejandro, P. 2014. Historical forest biomass dynamics modelled with Landsat spectral trajectories. ISPRS J. Photogramm. Remote Sens. **93**: 14–28. doi:10.1016/j.isprsjprs.2014.03.008.

Hastie, T., Tibshirani, R., and Friedman, J. 2009. The elements of statistical learning: data mining, inference, and prediction. 2nd ed. Springer-Verlag, New York. doi:10.1007/978-0-387-84858-7.

He, H.S. 2008. Forest landscape models: definitions, characterization, and classification. For. Ecol. Manage. **254**(3): 484–498. doi:10.1016/j.foreco.2007.08.022.

He, H.S., Mladenoff, D.J., Radeloff, V.C., and Crow, T.R. 1998. Integration of GIS data and classified satellite imagery for regional forest assessment. Ecol. Appl. **8**: 1072–1083. doi:10.1890/1051-0761(1998)008[1072:IOGDAC]2.0.CO;2.

Hijmans, R.J. 2015. raster: geographic data analysis and modeling. R package version 2.4-20. R Foundation for Statistical Computing, Vienna, Austria.

Huang, C., Goward, S.N., Masek, J.G., Thomas, N., Zhu, Z., and Vogelmann, J.E. 2010. An automated approach for reconstructing recent forest disturbance history using dense Landsat time series stacks. Remote Sens. Environ. **114**(1): 183–198. doi:10.1016/j.rse.2009.08.017.

Huang, S., Ramirez, C., Conway, S., Kennedy, K., Kohler, T., and Liu, J. 2017. Mapping site index and volume increment from forest inventory, Landsat, and ecological variables in Tahoe National Forest, California, USA. Can. J. For. Res. **47**(1): 113–124. doi:10.1139/cjfr-2016-0209.

Hudak, A.T., Crookston, N.L., Evans, J.S., Hall, D.E., and Falkowski, M.J. 2008. Nearest neighbor imputation of species-level, plot-scale forest structure attributes from lidar data. Remote Sens. Environ. **112**(5): 2232–2245. doi:10.1016/j.rse.2007.10.009.

Huete, A.R. 1988. A Soil-Adjusted Vegetation Index (Savi). Remote Sens. Environ. **25**(3): 295–309. doi:10.1016/0034-4257(88)90106-X.

Huete, A., Didan, K., Miura, T., Rodriguez, E.P., Gao, X., and Ferreira, L.G. 2002. Overview of the radiometric and biophysical performance of the MODIS vegetation indices. Remote Sens. Environ. **83**(1–2): 195–213. doi:10.1016/S0034-4257(02)00096-2.

Hunt, E.R., Jr., and Rock, B.N. 1989. Detection of changes in leaf water content using near- and middle-infrared reflectances. Remote Sens. Environ. **30**(1): 43–54. doi:10.1016/0034-4257(89)90046-1.

Ji, L., Wylie, B.K., Brown, D.R.N., Peterson, B., Alexander, H.D., Mack, M.C., Rover, J., Waldrop, M.P., McFarland, J.W., Chen, X., and Pastick, N.J. 2015. Spatially explicit estimation of aboveground boreal forest biomass in the Yukon River Basin, Alaska. Int. J. Remote Sens. **36**(4): 939–953. doi:10.1080/01431161.2015.1004764.

Jordan, C.F. 1969. Derivation of leaf-area index from quality of light on the forest floor. Ecology, **50**(4): 663–666. doi:10.2307/1936256.

Li, M. 2010. Estimation and analysis of forest biomass in northeast forest region using remote sensing technology. School of Forestry, Northeast Forestry University, Harbin, China.

Liang, Y., He, H.S., Wu, Z., and Yang, J. 2014. Effects of environmental heterogeneity on predictions of tree species' abundance in response to climate warming. Environ. Modell. Softw. **59**: 222–231. doi:10.1016/j.envsoft.2014.05.025.

Lopes, R.H.C., Reid, I., and Hobson, P.R. 2007. The two-dimensional Kolmogorov–Smirnov test. *In* Proceedings of the XI International Workshop on Advanced Computing and Analysis Techniques in Physics Research, Amsterdam, 23–27 April 2007. Proceedings of Science. pp. 196–206.

Lu, D. 2006. The potential and challenge of remote sensing-based biomass estimation. International Journal of Remote Sensing, **27**(7): 1297–1328. doi:10.1080/01431160500486732.

Luo, X., He, H.S., Liang, Y., Wang, W.J., Wu, Z., and Fraser, J.S. 2014. Spatial simulation of the effect of fire and harvest on aboveground tree biomass in boreal forests of Northeast China. Landscape Ecol. **29**(7): 1187–1200. doi:10.1007/s10980-014-0051-x.

Mahalanobis, P.C. 1936. On the generalised distance in statistics. Proc. Natl. Inst. Sci. India, **2**(1): 49–55.

Mao, D., Wang, Z., Luo, L., and Ren, C. 2012. Integrating AVHRR and MODIS data to monitor NDVI changes and their relationships with climatic parameters in Northeast China. Int. J. Appl. Earth Obs. Geoinform. **18**: 528–536. doi:10.1016/j.jag.2011.10.007.

Martin, M.E., Newman, S.D., Aber, J.D., and Congalton, R.G. 1998. Determining forest species composition using high spectral resolution remote sensing data. Remote Sens. Environ. **65**(3): 249–254. doi:10.1016/S0034-4257(98)00035-2.

McRoberts, R.E. 2009. A two-step nearest neighbors algorithm using satellite imagery for predicting forest structure within species composition classes. Remote Sens. Environ. **113**(3): 532–545. doi:10.1016/j.rse.2008.10.001.

McRoberts, R.E. 2012. Estimating forest attribute parameters for small areas using nearest neighbors techniques. For. Ecol. Manage. **272**: 3–12. doi:10.1016/j.foreco.2011.06.039.

Moeur, M., and Stage, A.R. 1995. Most similar neighbor: an improved sampling inference procedure for natural resource planning. For. Sci. **41**(2): 337–359.

Muinonen, E., Maltamo, M., Hyppänen, H., and Vainikainen, V. 2001. Forest stand characteristics estimation using a most similar neighbor approach and image spatial structure information. Remote Sens. Environ. **78**(3): 223–228. doi:10.1016/S0034-4257(01)00220-6.

Ohmann, J.L., and Gregory, M.J. 2002. Predictive mapping of forest composition and structure with direct gradient analysis and nearest-neighbor imputation in coastal Oregon, USA. Can. J. For. Res. **32**(4): 725–741. doi:10.1139/x02-011.

Oksanen, J., Blanchet, F.G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., Minchin, P.R., O'Hara, R.B., Simpson, G.L., Solymos, P., Stevens, M.H.H., Szoecs, E., and Wagner, H. 2017. vegan: Community Ecology Package. https://CRAN.R-project.org/package=vegan.

Ohmann, J.L., Gregory, M.J., Henderson, E.B., and Roberts, H.M. 2011. Mapping gradients of community composition with nearest-neighbour imputation: extending plot data for landscape analysis. J. Veg. Sci. **22**(4): 660–676. doi:10.1111/j.1654-1103.2010.01244.x.

Pflugmacher, D., Cohen, W.B., and Kennedy, R.E. 2012. Using Landsat-derived disturbance history (1972–2010) to predict current forest structure. Remote Sens. Environ. **122**: 146–165. doi:10.1016/j.rse.2011.09.025.

Pinty, B., and Verstraete, M.M. 1992. GEMI: a non-linear index to monitor global vegetation from satellites. Vegetatio, **101**(1): 15–20. doi:10.1007/BF00031911.

Powell, S.L., Cohen, W.B., Healey, S.P., Kennedy, R.E., Moisen, G.G., Pierce, K.B., and Ohmann, J.L. 2010. Quantification of live aboveground forest biomass dynamics with Landsat time-series and field inventory data: a comparison of empirical modeling approaches. Remote Sens. Environ. **114**(5): 1053–1068. doi:10.1016/j.rse.2009.12.018.

Prasad, A.M., Iverson, L.R., and Liaw, A. 2006. Newer classification and regression tree techniques: bagging and random forests for ecological prediction. Ecosystems, **9**(2): 181–199. doi:10.1007/s10021-005-0054-1.

Pu, R., and Landry, S. 2012. A comparative analysis of high spatial resolution IKONOS and WorldView-2 imagery for mapping urban tree species. Remote Sens. Environ. **124**: 516–533. doi:10.1016/j.rse.2012.06.011.

Qi, J., Chehbouni, A., Huete, A.R., Kerr, Y.H., and Sorooshian, S. 1994. A modified soil adjusted vegetation index. Remote Sens. Environ. **48**(2): 119–126. doi:10.1016/0034-4257(94)90134-1.

Qi, Y., and Li, F. 2015. Remote sensing estimation of aboveground forest carbon storage in Daxing'an Mountains based on kNN method. Sci. Silvae Sin. **51**(5): 10.

R Core Team. 2013. R: a language and environment for statistical computing [online]. R Foundation for Statistical Computing, Vienna, Austria. Available from http://www.R-project.org.

Riemann, R., Wilson, B.T., Lister, A., and Parks, S. 2010. An effective assessment protocol for continuous geospatial datasets of forest characteristics using USFS Forest Inventory and Analysis (FIA) data. Remote Sens. Environ. **114**(10): 2337–2352. doi:10.1016/j.rse.2010.05.010.

Rock, B.N., Vogelmann, J.E., Williams, D.L., Vogelmann, A.F., and Hoshizaki, T. 1986. Remote detection of forest damage. BioScience, **36**(7): 439–445. doi:10.2307/1310339.

Rouse, J.W., Jr, Haas, R.H., Schell, J.A., and Deering, D.W. 1973. Monitoring vegetation systems in the Great Plains with ERTS. *In* NASA. Goddard Space Flight Center 3d ERTS-1 Symp., Vol. 1, Sect. A, Paper A20. pp. 309–317.

Scheller, R.M., Domingo, J.B., Sturtevant, B.R., Williams, J.S., Rudy, A., Gustafson, E.J., and Mladenoff, D.J. 2007. Design, development, and application of LANDIS-II, a spatial landscape simulation model with flexible temporal and spatial resolution. Ecol. Modell. **201**(3–4): 409–419. doi:10.1016/j.ecolmodel.2006.10.009.

Schmitt, C.B., Burgess, N.D., Coad, L., Belokurov, A., Besançon, C., Boisrobert, L., Campbell, A., Fish, L., Gliddon, D., Humphries, K., Kapos, V., Loucks, C., Lysenko, I., Miles, L., Mills, C., Minnemeyer, S., Pistorius, T., Ravilious, C., Steininger, M., and Winkel, G. 2009. Global analysis of the protection status of the world's forests. Biol. Conserv. **142**(10): 2122–2130. doi:10.1016/j.biocon.2009.04.012.

Shataee, S., Kalbi, S., Fallah, A., and Pelz, D. 2012. Forest attribute imputation using machine-learning methods and ASTER data: comparison of k-NN, SVR and random forest regression algorithms. Int. J. Remote Sens. **33**(19): 6254–6280. doi:10.1080/01431161.2012.682661.

Su, Y., Guo, Q., Xue, B., Hu, T., Alvarez, O., Tao, S., and Fang, J. 2016. Spatial distribution of forest aboveground biomass in China: estimation through combination of spaceborne lidar, optical imagery, and forest inventory data. Remote Sens. Environ. **173**: 187–199. doi:10.1016/j.rse.2015.12.002.

Temesgen, B.H., LeMay, V.M., Froese, K.L., and Marshall, P.L. 2003. Imputing tree-lists from aerial attributes for complex stands of south-eastern British Columbia. For. Ecol. Manage. **177**(1–3): 277–285. doi:10.1016/S0378-1127(02)00321-3.

Tomppo, E., and Katila, M. 1991. Satellite image-based national forest inventory of Finland for publication in the IGARSS'91 digest. *In* Geoscience and Remote Sensing Symposium, IGARSS'91. Remote Sensing: Global Monitoring for Earth Management, International, Espoo, Finland, 3–6 June 1991. IEEE. pp. 1141–1144.

Tomppo, E., Olsson, H., Ståhl, G., Nilsson, M., Hagner, O., and Katila, M. 2008. Combining national forest inventory field plots and remote sensing data for forest databases. Remote Sens. Environ. **112**(5): 1982–1999. doi:10.1016/j.rse.2007.03.032.

van Ewijk, K.Y., Randin, C.F., Treitz, P.M., and Scott, N.A. 2014. Predicting fine-scale tree species abundance patterns using biotic variables derived from lidar and high spatial resolution imagery. Remote Sens. Environ. **150**: 120–131. doi:10.1016/j.rse.2014.04.026.

Wang, X., Yu, C., Hongwei, C., Yuanman, H., Linlin, J., Yuting, F., Wen, W., and Haifeng, W. 2014. Spatial pattern of forest biomass and its influencing factors in the Great Xing'an Mountains, Heilongjiang Province, China. Chin. J. Appl. Ecol. **4**: 974–982. doi:10.13287/j.1001-9332.2014.0082.

Wilson, B.T., Lister, A.J., and Riemann, R.I. 2012. A nearest-neighbor imputation approach to mapping tree species over large areas using forest inventory plots and moderate resolution raster data. For. Ecol. Manage. **271**: 182–198. doi:10.1016/j.foreco.2012.02.002.

Wilson, B.T., Woodall, C.W., and Griffith, D.M. 2013. Imputing forest carbon stock estimates from inventory plots to a nationally continuous coverage. Carbon Balance Manage. **8**(1): 1. doi:10.1186/1750-0680-8-1.

Wu, Z., He, H.S., Liu, Z., and Liang, Y. 2013. Comparing fuel reduction treatments for reducing wildfire size and intensity in a boreal forest landscape of northeastern China. Sci. Total Environ. **454–455**: 30–39. doi:10.1016/j.scitotenv.2013.02.058. PMID:23542479.

Xu, H. 1998. Forest in Great Xing'an Mountains of China. Science Press, Beijing.

Yu, X., and Zhuang, D. 2006. Monitoring forest phenophases of Northeast China based on MODIS NDVI Data. Resources Science, **28**(4): 111–117.

Zald, H.S.J., Ohmann, J.L., Roberts, H.M., Gregory, M.J., Henderson, E.B., McGaughey, R.J., and Braaten, J. 2014. Influence of lidar, Landsat imagery, disturbance history, plot location accuracy, and plot size on accuracy of imputation maps of forest composition and structure. Remote Sens. Environ. **143**: 26–38. doi:10.1016/j.rse.2013.12.013.

Zald, H.S.J., Wulder, M.A., White, J.C., Hilker, T., Hermosilla, T., Hobart, G.W., and Coops, N.C. 2016. Integrating Landsat pixel composites and change metrics with lidar plots to predictively map forest structure and aboveground biomass in Saskatchewan, Canada. Remote Sens. Environ. **176**: 188–201. doi:10.1016/j.rse.2016.01.015.

Zellweger, F., Braunisch, V., Baltensweiler, A., and Bollmann, K. 2013. Remotely sensed forest structural complexity predicts multi-species occurrence at the landscape scale. For. Ecol. Manage. **307**: 303–312. doi:10.1016/j.foreco.2013.07.023.

Zhang, Y., He, H.S., Dijak, W.D., Yang, J., Shifley, S.R., and Palik, B.J. 2009. Integration of satellite imagery and forest inventory in mapping dominant and associated species at a regional scale. Environ. Manage. **44**(2): 312–323. doi:10.1007/s00267-009-9307-7. PMID:19488811.

Zhang, Y., Liang, S., and Sun, G. 2014. Forest biomass mapping of Northeastern China using GLAS and MODIS data. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. **7**(1): 140–152. doi:10.1109/JSTARS.2013.2256883.

Zhu, X., and Liu, D. 2015. Improving forest aboveground biomass estimation using seasonal Landsat NDVI time-series. ISPRS J. Photogramm. Remote Sens. **102**: 222–231. doi:10.1016/j.isprsjprs.2014.08.014.

Appendix A

Appendix Figures A1–A3 appear on the following pages.

Fig. A1. Mean deviation (MD) vs. *k* for different *k* values and nearest-neighbor imputation models using different distance metrics based on combinations of environmental variables and seven sets of MODIS summary variables (all months or a single month of MODIS spectral variables from May to October in the legend). MD was calculated as the difference between the mean imputed and mean observed total aboveground biomass. Mean values are from 20 replicates; standard errors are very small and are not included for clarity. RF, Random Forest; MSN, most similar neighbor in canonical correlation space; msnPP, most similar neighbor computed using projection pursuit; GNN, gradient nearest neighbor.

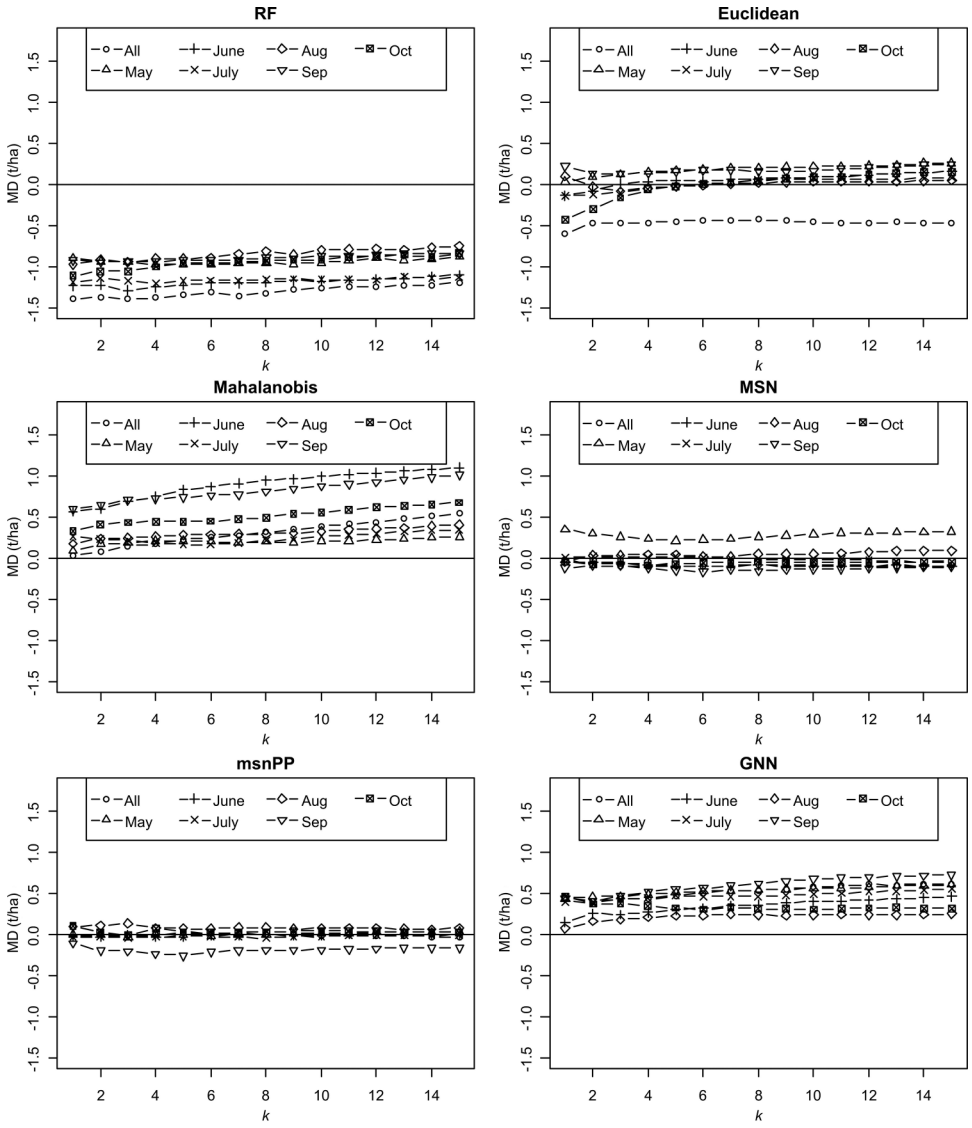


Fig. A2. Variance ratio (VR) vs. *k* for different *k* values and nearest-neighbor imputation models using different distance metrics based on combinations of environmental variables and seven sets of MODIS summary variables (all months or a single month of MODIS spectral variables from May to October in the legend). Mean values are from 20 replicates; standard errors are very small and are not included for clarity. RF, Random Forest; MSN, most similar neighbor in canonical correlation space; msnPP, most similar neighbor computed using projection pursuit; GNN, gradient nearest neighbor.

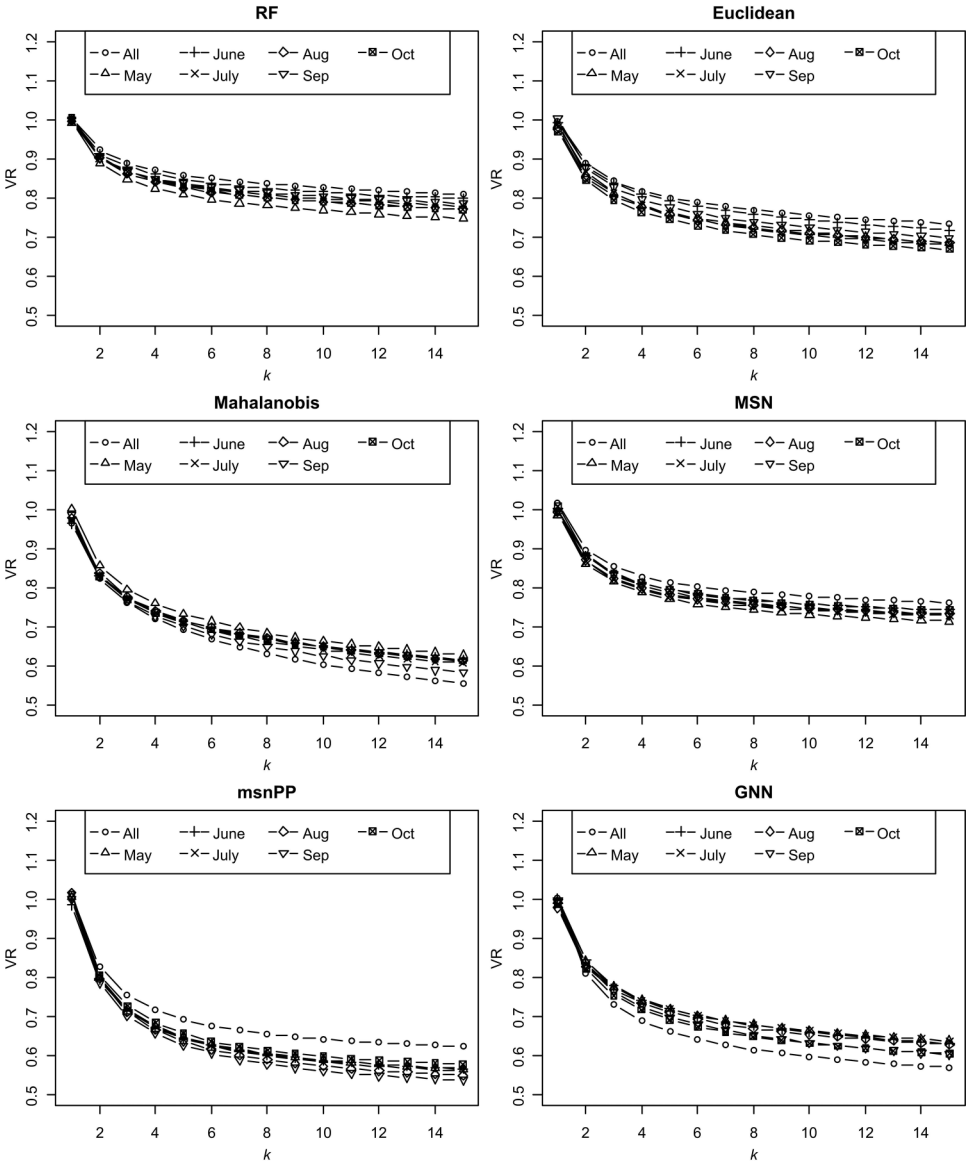


Fig. A3. Multivariate goodness of fit criterion (*T*) curves vs. *k* for different *k* values and nearest-neighbor imputation models using different distance metrics based on combinations of environmental variables and seven sets of MODIS summary variables (all months or a single month of MODIS spectral variables from May to October in the legend). Mean values are from 20 replicates; standard errors are very small and are not included for clarity. RF, Random Forest; MSN, most similar neighbor in canonical correlation space; msnPP, most similar neighbor computed using projection pursuit; GNN, gradient nearest neighbor.