Supplementary Material

S1

1. Harvested trees

1.1 Plot setup

The purpose of plot selection was to establish fixed and permanent plots representing regional *Eucalyptus* growing conditions and to provide harvested tree data on the single-tree scale with adequate consideration of spatial heterogeneity. Patches were selected first and met the following six conditions: (1) patch records were available from FMPI data for 2009; (2) forest stands were classified as timber or commercial forest; (3) forest patches were disturbance-free during the previous seven years, including but not limited to logging, fire, and pests; (4) forest patches were not replanted; (5) patches contained closed canopy forests; and (6) patches were monocultures, not mixed stands. Based on these six conditions, 2,980 *Eucalyptus* patches were selected from the FMPI data and fixed and permanent plots were established. The 2,980 selected patches were divided into ten groups based on forest age. Each stand group had been planted at the same time. We calculated the mean basal area for each group and used this as the basis for fixed plot selection, which was obtained from specified plot design and sampling procedures. In parallel, we considered site conditions, forest use, and forest origin (natural vs. man-made), and subsequently established 30 permanent square plots (20 m × 20 m). We recorded fixed-plot conditions by assigning a code to each fixed plot and recorded environmental conditions, including the following direct and indirect attributes: age, community structure, canopy density, and understory shrub conditions. Finally, a full tree survey was conducted in each fixed plot to obtain the following: DBH for every tree ≥ 8 cm in diameter, tree height, and other tree attributes.

1.2 Selection and cutting of standard wood

Standard wood was selected following a full tree survey. The following selection criteria were used: (1) Wood was located within the plot; stems were representative of the plot, with no disturbances (e.g., pests, fire, or anthropogenic activities); and the wood was healthy. (2) Based on the full tree survey data, a tree sampling method was used to calculate average basal area and three trees closest to the average values were selected (i.e., standard trees). These standard trees were cut down and the average biomass was calculated and multiplied by the stems per unit area to obtain the total *Eucalyptus* biomass per unit area.

1.3 Harvested tree measurements

Aboveground biomass was divided into three tissue types: stems, branches, and foliage. Four to six branches were systematically sampled from each tree at regular intervals over the entire crown length. Foliage was collected from each of the sampled branches. Stems were sectioned into meter-long pieces using a chainsaw.

The fresh weight of three tissue types was obtained in the field and 500 g of each tissue type (i.e., stems, branches, and foliage) were placed in plastic bags. The samples were stored under refrigeration during transportation to the laboratory. Fresh samples were oven dried at 85 °C to determine the constant dry weight.

2. Introduction to machine learning

2.1 Support vector machines for regression

A support vector machine (SVM) is a type of categorized algorithm that improves generalized machine learning ability by minimizing structural risks in order to minimize empirical risk and confidence intervals. In this way, it achieves adequate statistical trends from a limited number of samples. Compared with traditional machine learning methods, SVM adopts the principle of minimizing structural risks. Along with minimizing sample point errors, SVM simultaneously narrows the upper bound of generalized error in the model to improve the generalization ability of the model and to solve the problems of excessive model learning, nonlinearity, and dimensionality (Ukil, 2002).

The SVM classification model was trained using a C-classification method, with longitude, DBH, tree height, and forest age as the selection characteristics and the biomass data from the 30 plots as model training samples. The Gaussian inner product function served as the kernel function.

2.2 Radial basis function artificial neural networks

The basic components of radial basis function artificial neural networks (RBF-ANNs) include an input layer, a hidden layer, and an output layer, which are able to provide the best approximation for nonlinear functions and optimal global performance (Elanayar and Shin, 1994). The change from the input layer space to the hidden layer space is nonlinear, whereas the spatial transformation from the hidden layer to the output layer space is linear. The RBF-ANN has good generalizability, requires fewer calculations, and has a faster learning speed than other machine learning algorithms. Therefore, the RBF-ANN avoids lengthy iterative calculations, such as those found in the learning algorithms of back propagation neural networks, and the possibility of falling into a local extremum. RBF-ANN is widely used in many fields, including meteorology (Nath et al., 2016), soil (Zakian, 2017), vegetation (Hilbert and Ostendorf, 2001), and engineering control (Sarimveis et al., 2004).

2.3 Random forest

The random forest (RF) algorithm model is a relatively new machine learning technique and data mining method developed by Breiman in 2001. It is a modern classification and regression technology that combines self-learning technologies (Breiman, 2001). In order to achieve a better performance than individual classifiers, combinatorial learning approaches integrate several individual classifiers to determine the final classification of a case. If a single classifier is considered as a decision maker, the method of combinatorial learning is equivalent to a decision-making process involving multiple decision makers.

3. Introduction to P-BSHADE

P-BSHADE is an extension of the BSHADE method, which stands for the best linear unbiased estimation (BLUE) model for biased-spatial-location data (Hu et al., 2013). With the BSHADE model, the spatial correlation and heterogeneity of the target data are added into the model using prior knowledge (such as forest AGB). In addition, through rectification of sample points, the BLUE model can estimate the target subjects. The strategy of the algorithm is to transform the problem into one of solving for the extremum of a multivariate function with constraint conditions, followed by using the Lagrange multiplier method and the overall estimate to acquire the corresponding parameters (Wang et al., 2011) (i.e., each sample in this method is given a certain weight, so that the variance between each sample and the true value is minimized to achieve rectification).

Based on the BSHADE method, P-BSHADE is a BLUE-based interpolation method that considers both temporal and spatial heterogeneity. It can use biased samples to deduce the corresponding attributes of regions with missing samples. Therefore, the P-BSHADE model includes the following characteristics and assumptions: (1) the spatial distribution of the target data (such as forest AGB) is heterogeneous and (2) the correlations and differences among the target data in different forests (or sites) is included in the operation of the model (Xu et al., 2013). The performance of the P-BSHADE method has been tested using average annual temperature data in China from 1950 to 2000 (Xu, 2013).

4 Forest Management and Planning Inventory (FMPI) data

The FMPI data for the entire study area were provided by the Forestry Department of Fujian Province, China. This forest inventory used large-scale sampling methods to collect detailed information about the characteristics and conditions of each forest type. The FMPI data consisted of irregular polygons that were drawn based on the structured characteristics of the forest. Each polygon was homogeneously structured. In this study, we selected FMPI data for *Eucalyptus* plantation forests (2,980 patches).

In every patch, all trees with a diameter at breast height (DBH) greater than 8 cm were measured. The data contained patch area, tree age (which was the same for all trees in a given patch because they were planted at the same time), plantation density, mean DBH, mean tree height, and total volume of each patch. All variables were measured within each forest patch and the average values were used as the factor value for each patch. The accuracy of forest patch variables was tested using systematic sampling. A 95% sampling precision was required. Table B.1 lists the statistical description of the forest patch data.

5 Robustness of combined models

We established 22 independent sample plots (Figure C.2) and conducted non-destructive measurements of each tree in July 2019. We then repeated the plot-level model construction workflow for these data and evaluated the models. The independent sample plots were widely distributed throughout the eastern section of the study area.

6 Model application and upscaling of AGB mapping

We applied the chosen optimal model to each *Eucalyptus* forest patch (2,980 patches) and estimated the total AGB for all patches in the study area. We regarded the irregular polygon forest patches from the FMPI as a homogenous sample plot and applied the optimal plot-level model to upscale forest AGB. We compared this upscaled forest AGB with the AGB map obtained by an allometric model and calculated the relative error (RE) (see Equation A.1) of AGB between the two methods.

$$\mathrm{RE} = |y_i - y_j|/y_i \times 100\% \qquad \text{(A.1)}$$

where y_j represents the predictive AGB value of each irregular polygon forest patch by the optimal model and y_i is the predicted AGB value of each irregular polygon forest patch by the allometric model.

The allometric model was expressed as follows:

$$\mathrm{AGB} = \mathrm{a}((DBH)^2 H)^b \qquad \text{(A.2)}$$

where, DBH is the diameter at breast height (m), H is the tree height (m), and a and b are constants. This model is acknowledged as a fast, simple, and basic method to calculate regional AGB. In our study, we used the AGB, mean H, and mean DBH of the 30 sample plots to create the plot-level allometric model.

Figure C.3 shows the spatial distribution of the AGBs predicted by the RF & P-BSHADE model. The range of AGBs was 7.54-89.93 $\mathrm{Mg \cdot ha^{-1}}$, with an average AGB of 41.21 $\mathrm{Mg \cdot ha^{-1}}$, a median AGB of 43.53 $\mathrm{Mg \cdot ha^{-1}}$, a standard deviation of 18.83 $\mathrm{Mg \cdot ha^{-1}}$, and a coefficient of variation of 45.69%.

The total AGB of the Nanjing area (2,980 forest patches) estimated by RF & P-BSHADE was 122,812.1 $\mathrm{Mg \cdot ha^{-1}}$ and that estimated by the allometric model was 123,021.5 $\mathrm{Mg \cdot ha^{-1}}$. The relative percent difference in total AGB between the two methods was 0.17%. Meanwhile, the MRE of AGB between the two methods ranged from 0.04% to 99.8%, with an average MRE of 19.93%.

S2

Table B.1 Statistical description of forest patch data.

	Number of patches	Minimum	Maximum	Mean	Standard deviation
Age (years)	2,980	1	51	5.05	2.42
Stand density (stems/ha)	2,980	135	3450	1377.63	241.10
DBH (cm)	2,980	5.0	60.0	12.30	3.55
Tree height (m)	2,980	1.5	48.50	13.40	3.99

Note: Of the 2,980 forest patches, for which the maximum age was 51 years, only 24 forest patches were older than 10 years, all of which were identified as mature forest.

Table B.2 Tree structures for calculating the biomass of the 90 harvested trees.

Age	DBH	Height	Individual biomass (kg)	Age	DBH	Height	Individual biomass (kg)
(yr)	(cm)	(m)	Aboveground	(yr)	(cm)	(m)	Aboveground
1	3.3	4.3	1.9376	6	15.0	20.8	82.2273
	3.0	4.0	2.2500		15.3	20.8	99.3969
	3.2	4.3	1.8514		15.0	21.1	102.5718
	2.1	3.3	1.1061		15.3	19.9	97.7377
	2.1	3.4	1.0697		15.0	21.2	93.3897
	2.4	3.3	1.3143		14.5	20.8	89.4676
	3.4	4.6	2.2976		14.6	19.4	81.7034
	3.3	4.7	2.3782		15.0	19.4	81.8693
	3.3	4.5	2.0494		14.6	20.1	87.1974
2	7.6	10.1	14.4861	7	18.0	20.4	119.9316
	8.0	8.5	14.7833		17.8	20.8	106.3167
	8.1	9.9	14.3030		18.0	20.4	143.0096
	7.2	10.5	12.1682		16.7	20.0	113.6738

	7.0	10.4	11.7154		16.6	20.9	99.6045
	7.0	10.8	11.1324		16.4	21.4	98.7499
	7.2	9.2	12.3033		16.9	19.8	102.7874
	7.2	9.5	11.0665		16.9	20.2	97.2996
	7.0	8.1	10.2483		15.6	20.3	89.5590
3	6.1	6.3	5.5350	8	14.3	21.1	89.6489
	7.0	6.9	8.8532		14.5	19.8	72.6971
	6.4	6.8	7.5987		14.0	19.2	90.9861
	6.2	7.6	6.3156		16.4	19.7	99.4468
	7.2	7.9	9.5706		16.4	20.1	97.8657
	7.2	7.7	9.7457		17.2	21.2	112.4650
	6.1	6.9	6.4039		14.0	17.7	63.5059
	6.2	9.4	9.2803		15.0	20.3	81.3824
	5.4	6.6	5.7853		14.9	19.3	84.1050
4	11.1	18.6	36.7169	9	16.9	25.5	110.3010
	12.1	17.3	50.7412		17.2	25.1	146.4738
	11.8	17.3	44.8078		17.5	24.5	130.5710
	8.9	11.7	16.5647		16.1	23.5	117.4427
	9.2	17.4	27.9658		15.8	22.9	106.7083
	8.8	15.2	24.5316		15.9	23.3	112.0993
	13.2	17.9	56.0009		18.4	26.6	168.4229
	13.1	18.2	58.7273		18.4	24.7	144.5210
	12.4	17.8	51.5655		18.3	26.0	167.0830
5	13.2	19.7	62.9911	10	18.2	27.0	136.6728
	13.9	16.5	68.7846		18.5	25.0	163.4031
	12.9	16.1	58.5322		18.2	26.2	150.9330
	13.4	19.3	81.9325		14.0	18.5	69.9841
	13.4	19.4	84.0987		13.9	22.1	76.9977
	13.1	18.9	73.2317		13.9	24.0	91.4171
	13.4	19.0	70.4283		17.6	23.8	118.4468
	12.9	17.1	70.5207		17.6	22.2	149.1616
	13.8	18.6	96.5537		17.6	25.6	138.2509

Table B.3 Construction of the optimal model.

Leave-one-out		Model 1	Model 2	…	Model 7
Validation data (Plot AGB)	Training data (Plot AGB and predictor variables)	Simulated data1 (Simulated AGB 1)	Simulated data2 (Simulated AGB2)	Simulated data (Simulated AGB)	Simulated data7 (Simulated AGB7)
Plot ID	Plot ID	Plot ID	Plot ID	Plot ID	Plot ID
1	2-30	1 S1	1 S2	…	1 S7
2	1,3-30	2 S1	2 S2	…	2 S7
3	1-2,4-30	3 S1	3 S2	…	3 S7
…	…	…	…	…	…
29	1-28,30	29 S1	29 S2	…	29 S7
30	1-29	30 S1	30 S2	…	30 S7
AGB (group M)		AGB (group1)	AGB (group2)	…	AGB (group7)
		MAE1, MRE1 and RMSE1	MAE2, MRE2 and RMSE2	…	MAE7, MRE7 and RMSE7

Table B.4 Statistical description of AGB and selected variables for sample plots.

Variables	Mean	Median	Standard deviation	Coefficient of variation	Minimum	Maximum
Aboveground biomass, AGB	47.34	46.64	34.46	0.73	1.02	135.79

(t/ha)						
Longitude	117.48	117.47	0.02	$0.13*10^{-5}$	117.446	117.503
Diameter at breast height, DBH (cm)	12.29	13.19	4.48	0.36	2.19	17.99
Tree height, h (m)	12.98	14.42	4.72	0.36	2.83	18.23
Age (years)	5.5	5.5	2.92	0.53	1	10

Table B.5 List of model accuracy indexes and their definitions.

Model accuracy index	**Description**	**Interpretation**
Mean Absolute Error (MAE)	Mean absolute error is the mean of the absolute deviations of all individual measurements from arithmetical mean values.	This represents the mean of absolute deviations of the true biomass of the 30 sample plots from the average biomass of the 30 sample plots obtained by a given prediction method. Because the deviations are expressed in absolute values, the mean absolute error is not cancelled out by positive and negative numbers. Therefore, the mean absolute error can better reflect the actual prediction error.
Mean Relative Error (MRE)	Mean relative error is the average value of the relative error, which is usually expressed as the absolute value (i.e., the absolute value of mean relative error). The relative error is the ratio of the absolute error to the measured value or the average of multiple measurements.	This represents the average value of the ratio of the absolute error (the absolute value of the difference between the true value and the simulated value) for the biomass of each of the 30 sample plots to the predicted values. It is used to analyze the accuracy and precision of the results.
Root Mean Square Error (RMSE)	Square root of the ratio between the square of the deviation of the observed value from the true value and n, the number of observations. In actual measurement, the number of observations, n, is always limited and the true value can only be substituted by the most reliable (best) value.	This represents the average of the square root of the following value: for real and simulated values of the biomass of each of the 30 sample plots, the square of their difference is divided by 30. Because the results are very sensitive to extremely large or small errors in a set of measurements, it

		can better reflect the precision of the measurement.
The Normalized Root Mean Square Error (nRMSE)	The normalized root mean square error is the RMSE divided by the average of the observed values of a variable being predicted.	When comparing the modelling accuracies of different studies presenting different forest types, nRMSE is more meaningful because the intrinsic AGB variability is very different between drastically different forest types (e.g., a tropical rainforest (large) and a Eucalyptus plantation (small)).

Table B.6 Leave-one-out cross-validation for machine learning (support vector machine, artificial neural network, and random forest), spatial statistical analysis (P-BSHDE), and results from paired combinations of the two types.

Method	MAE	MRE	RMSE	nRMSE
SVM	11.168	0.2479	10.388	0.2182
ANN	12.149	0.267	10.388	0.2182
RF	10.155	0.259	9.429	0.1980
P-BSHADE	18.371	0.391	14.077	0.2957
SVM-＆P-BSHADE	6.883	0.125	6.304	0.1324
ANN-＆P-BSHADE	10.136	0.205	9.633	0.2023
RF-＆P-BSHADE	5.679	0.130	5.299	0.1113

S3

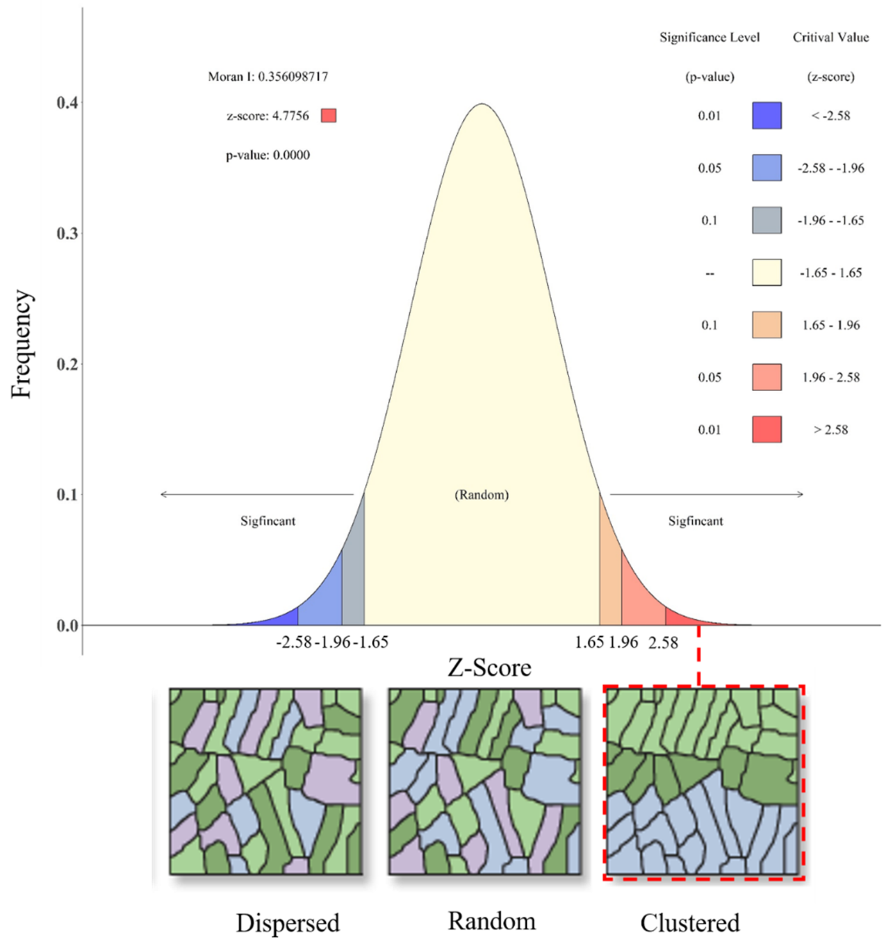

Figure C.1 Spatial autocorrelation report.

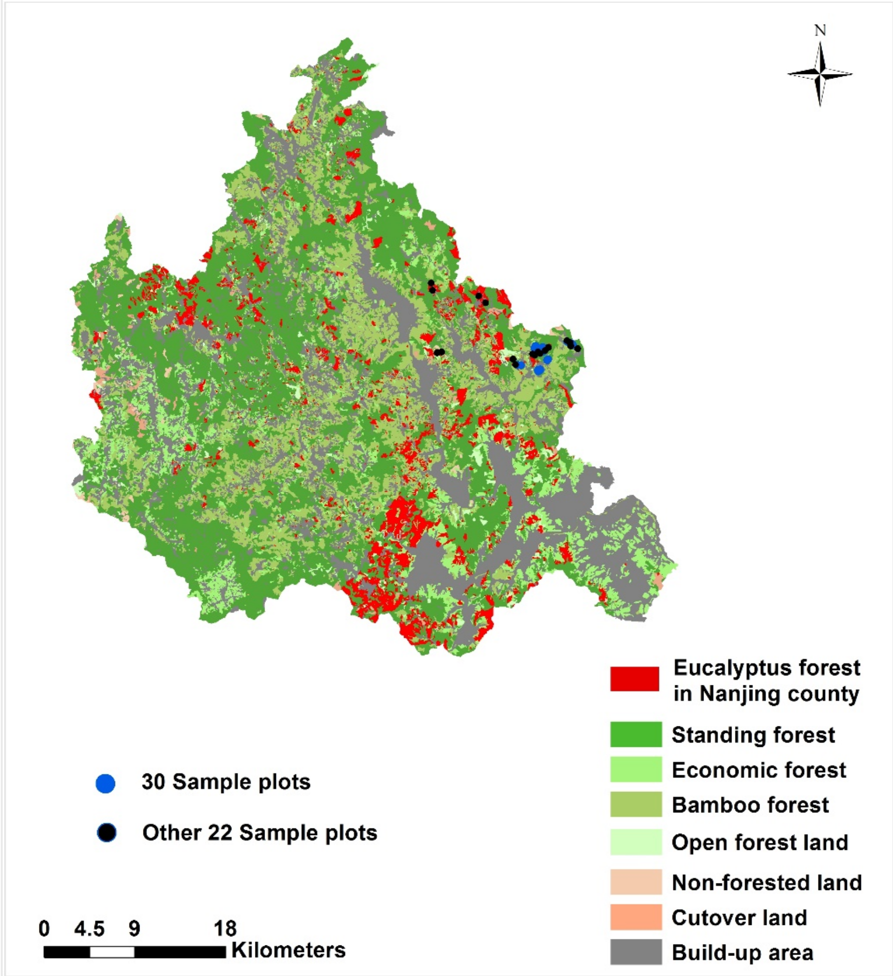

Figure C.2 The location of experimental sample plots (blue dots) and independent sample plots (black dots).

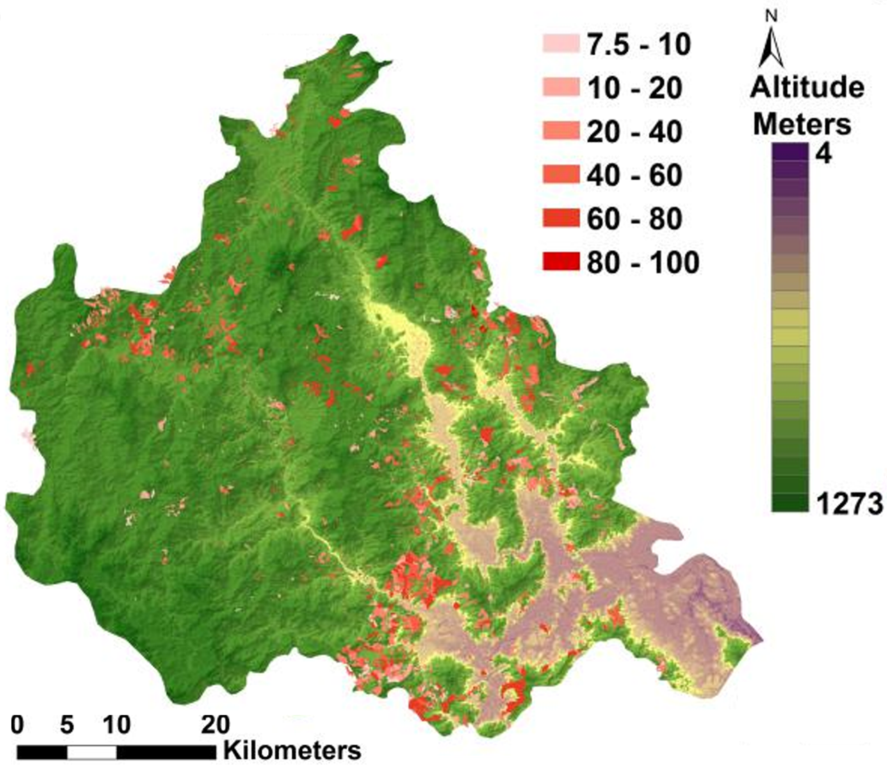

Figure C.3 Upscaling map of AGB using RF & P-BSHADE.

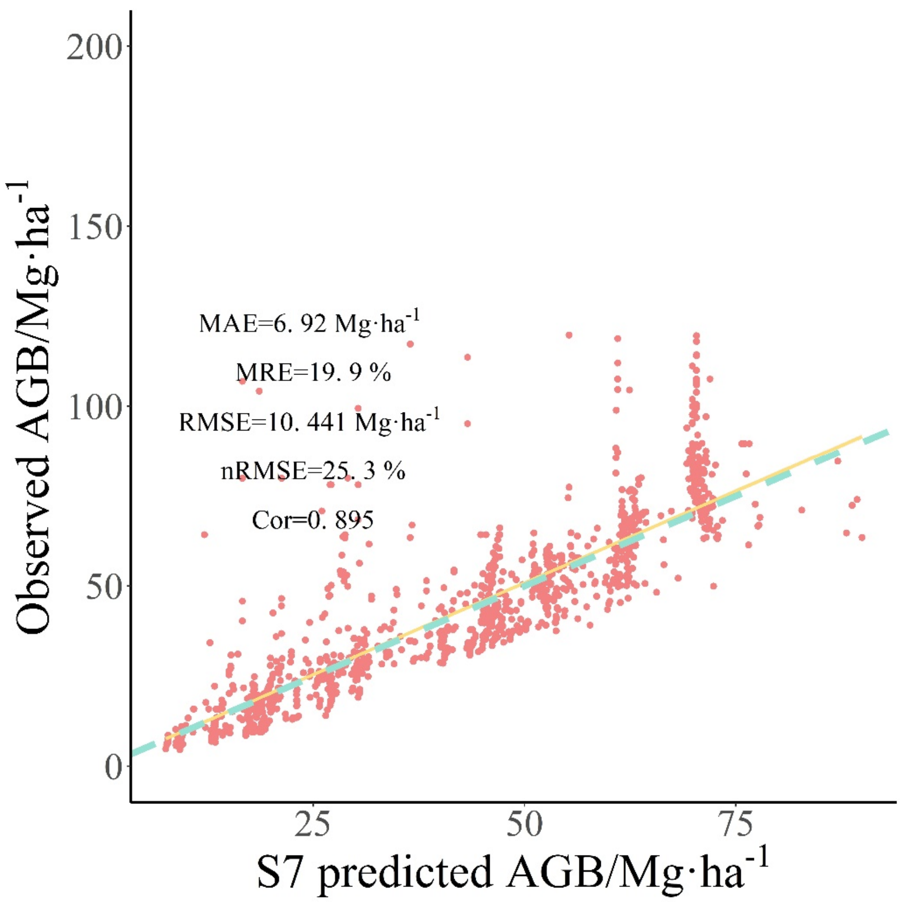

Figure C.4 Comparison of upscaling by RF & P-BSHADE with upscaling by the allometric model. The green dashed line corresponds to a 1:1 relationship; each dot represents an individual forest patch; the solid yellow line indicates the trend line for the dots.

References

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

Elanayar, V.T.S., Shin, Y.C., 1994. Radial basis function neural network for approximation and estimation of nonlinear stochastic dynamic systems. IEEE Transactions on Neural Networks 5(4) 594-603.

Hilbert, D.W., Ostendorf, B., 2001. The utility of artificial neural networks for modelling the distribution of vegetation in past, present and future climates. Ecological Modelling 146(1-3) 311-327.

Hu, M.G., Wang, J.F., Zhao, Y., Jia, L., 2013. A B-SHADE based best linear unbiased estimation tool for biased samples. Environmental Modelling & Software 48(48) 93-97.

Nath, S., Kotal, S.D., Kundu, P.K., 2016. Seasonal prediction of tropical cyclone activity over the north Indian Ocean using three artificial neural networks. Meteorology and Atmospheric Physics 128(6) 751-762.

Sarimveis, H., Alexandridis, A., Mazarakis, S., Bafas, G., 2004. A new algorithm for developing dynamic radial basis function neural network models based on genetic algorithms ☆. Computers & Chemical Engineering 28(1–2) 209-217.

Ukil, A., 2002. Support vector machine. Computer Science 1(4) 1-28.

Wang, J.F., Reis, B.Y., Hu, M.G., George, C., Yang, W.Z., Qiao, S., Li, Z.J., Li, X.Z., Lai, S.J., Chen, H.Y., 2011. Area Disease Estimation Based on Sentinel Hospital Records. Plos One 6(8) e23428.

Xu, C.D., 2013. Modeling of uncertainty of temperature observation and anomaly stratification. University of Chinese Academy of Sciences.

Xu, C.D., Wang, J.F., Hu, M.G., Li, Q.X., 2013. Interpolation of missing temperature data at meteorological stations using P-BSHADE*. Journal of Climate 26(19) 7452-7463.

Zakian, P., 2017. An efficient stochastic dynamic analysis of soil media using radial basis function artificial neural network. Frontiers of Structural and Civil Engineering 11(4) 470-479.