Journal of Applied Remote Sensing

RemoteSensing.SPIEDigitalLibrary.org

Modeling wetland aboveground biomass in the Poyang Lake National Nature Reserve using machine learning algorithms and Landsat-8 imagery

Rongrong Wan
Peng Wang
Xiaolong Wang
Xin Yao
Xue Dai

SPIE.

Rongrong Wan, Peng Wang, Xiaolong Wang, Xin Yao, Xue Dai, "Modeling wetland aboveground biomass in the Poyang Lake National Nature Reserve using machine learning algorithms and Landsat-8 imagery," *J. Appl. Remote Sens.* **12**(4), 046029 (2018), doi: 10.1117/1.JRS.12.046029.

Modeling wetland aboveground biomass in the Poyang Lake National Nature Reserve using machine learning algorithms and Landsat-8 imagery

Rongrong Wan,[a,b,*] Peng Wang,[a,b] Xiaolong Wang,[a,b,*] Xin Yao,[c] and Xue Dai[a,b]

[a]Chinese Academy of Sciences, Nanjing Institute of Geography and Limnology, Key Laboratory of Watershed Geographic Sciences, Nanjing, China

[b]University of Chinese Academy of Sciences, College of Resources and Environment, Beijing, China

[c]Nanjing University of Information Science and Technology, School of Geographic Sciences, Nanjing, China

Abstract. Quantitative estimation of wetland aboveground biomass (AGB) is an essential aspect in evaluating the health and conservation of this valuable ecosystem. We combine AGB field measurements and remote sensing data to establish a suitable model for estimating wetland AGB in the Poyang Lake National Nature Reserve (PLNNR), which is included in the Ramsar Convention's List of Wetlands of International Importance. All field sampling points cover four dominant vegetation communities (*Carex cinerascen*, *Phalaris arundinacea*, *Artemisia selengensis*, and *Miscanthus sacchariflorus*) in the PLNNR. Wetland AGB is retrieved from the Landsat-8 OLI image. To improve the accuracy of wetland AGB estimation, we compare the performances of three machine learning algorithms, namely, random forest (RF), back-propagation neural network (BPNN), and support vector regression (SVR), with linear regression (LR) in estimating the AGB in the PLNNR. Results are as follows: (1) the RF model with a root-mean-square error of 0.25 $kg\,m^{-2}$ performs better than BPNN (0.29 $kg\,m^{-2}$), SVR (0.27 $kg\,m^{-2}$), and LR (0.31 $kg\,m^{-2}$) in our testing dataset, and AGB density in the PLNNR is between 0 and 1.973 $kg\,m^{-2}$. (2) The four most important features for AGB modeling are near-infrared, short-wave infrared 1 band, enhanced vegetation index, and red band. Our study presents an effective and operational RF model that estimates wetland AGB from Landsat data, providing a scientific basis for floodplain wetland carbon accounting and possible future studies, such as the linkage between wetland AGB and the great water level fluctuations. [DOI: 10.1117/1.JRS.12.046029]

Keywords: aboveground biomass; machine learning algorithms; random forests; Landsat-8 OLI image; Poyang Lake National Nature Reserve.

Paper 180701 received Aug. 27, 2018; accepted for publication Nov. 29, 2018; published online Dec. 27, 2018.

1 Introduction

In recent years, quantitative evaluation of the wetland vegetation biomass has attracted increasing attention worldwide, considering that this method is an important index for evaluating the health of wetland ecosystems.[1,2] Quadrat survey, one of the main traditional methods,[3,4] has considerable disadvantages when used in complex ecosystems, such as heavy workload, huge costs, and large-scale information insufficiency when measured over a short time period. In comparison with traditional methods, remote sensing technology can rapidly, accurately, and nondestructively estimate the vegetation biomass of wetlands.

Studies of wetland biomass have focused mainly on aboveground biomass (AGB). Optical remote sensing, synthetic aperture radar (SAR), and light detection and ranging (LiDAR) are the three main methods for mapping wetland AGB. The differences among structure, crown

*Address all correspondence to Rongrong Wan, E-mail: rrwan@niglas.ac.cn; Xiaolong Wang, E-mail: wangxl@niglas.ac.cn

width, and plant diameter allow SAR to utilize the backscattering ratio for predicting vegetation biomass. Microwave technology not only interacts with the canopy but also can penetrate into vegetation stalks.[5] SAR can obtain the surface and body scattering information of vegetation due to its capability to penetrate clouds and vegetation. Thus, SAR is suitable for inverting vegetation parameters with evident structural characteristics, such as tall trees in forests. In addition, the saturation problem of backscattering coefficients limits SAR application in wetlands. Therefore, SAR is rarely used for the biomass inversion of wetland vegetation. Recently, the advantages of acquiring structural information about objects on the ground have rendered LiDAR further attractive to vegetation biomass inversion researchers. The earliest application of LiDAR was in measuring forest biomass. Then, LiDAR was utilized successfully for wetland biomass, especially mangrove forests in costal zones.[6–9] However, LiDAR applications are restricted by weak penetrability, low saturation in plants found in high-density canopies, and deficiencies in spectral information, particularly with tall and lush trees, mangrove forests, and low-lying herbs in freshwater wetlands.

Optical remote sensing technology is a common and well-tested method in terms of data availability, processing simplicity, and extensive applications over a large region. However, high-resolution remotely sensed images and LiDAR or SAR data are often restricted by their limited spatial and temporal coverage. Accordingly, numerous researchers prefer medium-resolution satellite images for measuring AGB over long time periods and at large areas.[10] Landsat is a trade-off of spatial, temporal, and spectral resolution; thus, it is a good option for large-scale AGB modeling.[11–14] Generally, the optical remote sensing method utilizes the spectral characteristics of plants, particularly the huge difference in reflectance in the red and near-infrared (NIR) bands, to construct a vegetation index (VI) for analyzing relationships. Normalized difference vegetation index (NDVI) is the most commonly used VI;[15–18] however, it is greatly affected by the atmosphere, soil composition, and heavy saturation in dense vegetation. Thus, researchers have proposed modified VIs, including soil-adjusted vegetation index (SAVI),[19] modified SAVI,[20] and enhanced vegetation index (EVI).[21] These indices are widely used and frequently combined to overcome the effects of background noise and improve the accuracy of biomass evaluation.[22–25]

Remote sensing methods for modeling AGB can be divided into two groups: statistical and physical models. Physical models, such as SAIL, Kuusk, and PROSEPCT, have been used to establish the link between the vegetation spectral reflectance (leaf or canopy) and biomass by analyzing the entire radiation transmission process of light inside and outside the vegetation. These models are useful under certain circumstances; however, their complexity, the over-abundance of parameters, and the uncertainty of measurements limit their application in large-scale regions.[26] For statistical models, a single or multiple VIs are traditionally adopted as the predictors for establishing a linear, exponential, logarithmic, or power model.[27,28] The development of machine learning and artificial intelligence has allowed for improved predictive accuracy. Such techniques can produce complex nonlinear mappings due to advanced learning strategies by utilizing the information contained in a set of reference samples. Another advantage is that no assumptions have to be formulated about data distribution. Thus, nonlinear machine learning methods are often regarded as distribution-free. Given this property, the retrieval process can integrate data from different sources with poorly defined (or unknown) probability density functions that relate well to the target variable. Regardless of the approach, either empirical or physical models, the high complexity and nonlinearity of retrieval problems require the development and usage of further advanced methods. The artificial neural network (ANN)[29] is a commonly used technique in the field of geo-/biophysical variable retrieval.[30,31] The ANN, due to its effectiveness and relatively higher accuracy, is more effective for estimating wetland AGB than the traditional linear model.[32,33] Numerous studies[34–36] have shown that the ANN model exhibits better accuracy, stability, and computational speed than the other investigated strategies.

Support vector regression (SVR)[37] has also become popular in the last few years and is particularly effective in the field of wetland AGB retrieval.[30,38] Study results reveal the promising features of this method, such as its good intrinsic generalization capability and its capacity for overcoming noise interference when reference samples are limited. Ensemble methods, such as random forest (RF),[39] have successfully been used to enhance predictive accuracy in the ecology

field.[40,41] The RF algorithm is a nonparametric statistical technique that can synthesize regression or classification functions on the basis of discrete or continuous datasets. The RF can also handle the complex relationships between predictors due to noise when using large amounts of data and weighing the importance of each input variable. In the remote sensing field, the RF has been widely applied in various domains as a classification algorithm.[42–44] Mutanga et al.[45] investigated the capability of RF to model AGB in iSimangaliso Wetland Park on the basis of WorldView-2 images. Recently, Byrd et al.[46] generated a remote sensing model based on RF to model tidal marsh AGB and carbon stocks in the United States. Studies that compare the effectiveness of several machine learning algorithms in modeling and mapping AGB based on Landsat images are limited, particularly at a vegetation landscape scale for seasonal lake wetland in floodplain areas. Therefore, such a study must be conducted.

In this study, we evaluated the effectiveness of linear regression (LR), back-propagation neural network (BPNN), SVR, and RF models in estimating wetland AGB in the Poyang Lake National Nature Reserve (PLNNR). The objectives of this study are as follows: (1) to explore which machine learning algorithms and spectral features can yield the most accurate AGB, (2) to estimate the AGB and their distribution and various characteristics in the PLNNR quantitatively, and (3) to evaluate the importance of each input band variable derived from Landsat images for predicting AGB.

2 Materials and Methods

2.1 *Study Area*

Poyang Lake, the largest freshwater lake in China, is located at 115°47 to 116°45'E and 28°22' to 29°45′N on the southern bank of the Yangtze River (Fig. 1). The lake is fed primarily by five tributaries (Ganjiang, Fuhe, Xinjiang, Raohe, and Xiushui Rivers) and is connected to the Yangtze River at Hukou. Poyang Lake has a subtropical monsoon climate with an average annual temperature of 17.6°C and mean annual precipitation level from 1450 to 1550 mm, with the rainy season generally occurring in summer. Interactions among the hydrology, soil, and plants of Poyang Lake have formed a unique wetland ecosystem, which provides essential functions, such as water supply, floodwater storage, and biodiversity maintenance. Poyang Lake wetlands are home to 102 species of aquatic plants from 38 families and to 122 species of fish from 23 families. More than 280 bird species are also available, representing 12 genera and 51 families, including 50 rare species.

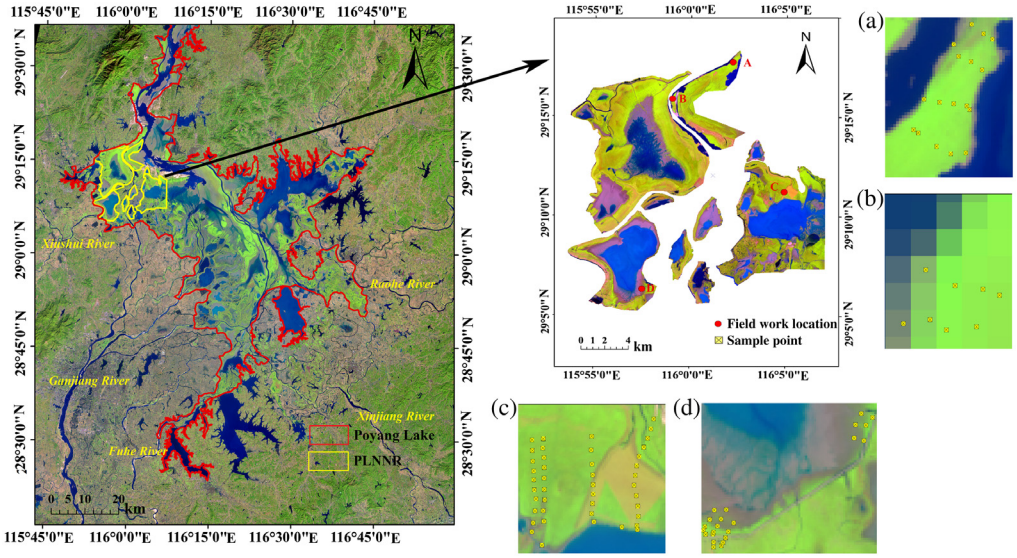

Fig. 1 Location of the PLNNR and the distribution of field sampling points in 2016. (a) Ganjiang River delta, (b) Sidu Island, (c) Dachahu sublake, and (d) Dahuchi sublake.

The PLNNR is located northwest of Poyang Lake, at the intersection of Ganjiang and Xiushui Rivers (Fig. 1). The PLNNR, with an area of 224 km^2, was established in 1988 to preserve wintering birds.[47] Twenty-three threatened species in the International Union for Conservation of Nature and Natural Resources red list[48] were found in the PLNNR, and approximately 95% of the entire population of critically endangered Siberian cranes (*Grus leucogeranus* Pallas), nearly 80% of endangered Oriental storks (*Ciconia boyciana*), and over 70% of vulnerable white-naped cranes (*Grus vipio*) wintered in the PLNNR.[49,50] For these reasons, Poyang Lake was one of the first wetlands to be included in the Ramsar Convention's List of Wetlands of International Importance.[51]

Complex inflow, outflow, and backflow patterns lead to large seasonal water level fluctuations.[52] The plant distribution of the PLNNR wetland, accompanied by the fluctuating water level, is characterized by a typical concentric pattern along the elevation gradient from the lake to the shoreland.[53] Four main types of plants are abundant in the PLNNR wetlands, namely, *Carex cinerascen*, *Phalaris arundinacea*, *Miscanthus sacchariflorus*, and *Artemisia selengensis*. They form three belts, namely, bulrushes (*Miscanthus sacchariflorus* or *Phragmites australis* communities), sedges (*Carex cinerascen* or *Artemisia selengensis* communities), and sparse emergent vegetation (*Phalaris arundinacea* communities), which occur naturally along a moisture gradient from the higher lands to the lake shoreline. Wetland vegetation in the PLNNR is distributed in different types of bottomlands, which are often inundated during flood season. These bottomlands include the littoral land of the main lake, inflow river delta, sublakes detached from the main lake during autumn and winter, and small islands seasonally submerged during summer. From October to December, water levels are low, thereby exposing the areas of these vegetation communities in Poyang Lake. At this time of the year, emergent vegetation (e.g., *Miscanthus sacchariflorus*) experiences a heading stage and subsequently withers and dies; sedges (e.g., *Carex cinerascen*), which have a long growing period, continue blooming; and sparse emergent vegetation (e.g., *Phalaris arundinacea* and *Artemisia selengensis*) begin to wither and die. During this period, the spectral characteristics and AGB of these vegetation communities do not change considerably. Thus, this type of phenology phenomenon necessitates the implementation of a field work that covers all four dominant vegetation communities in the PLNNR.

2.2 *Field Surveying and Data Collection*

The field campaign was conducted on November 23 to 30, 2016. We selected four typical bottomlands, which are representative of PLNNR wetland for sampling, and a total of 94 sampling points, which covered the four main vegetation communities in Poyang Lake wetland (Fig. 1 and Table 1). Then, in view of the concentric pattern of vegetation communities along the elevation gradient, a predetermined fixed number of 1 m × 1 m sample plots at each bottomland were created from the shoreline to the relatively higher land, where flood cannot overflow. The interval between plots is 50 to 120 m (in accordance with the distribution of slop and vegetation belts in the sites) to cover all the main vegetation communities at different elevations in various

Table 1 Number of sample points in every sample field in 2016.

Sample field	C.C	Pha.A	A.S.	Mis.S	Total
(a) Ganjiang River delta	5	4	6	2	17
(b) Sidu Island	2	2	3	1	8
(c) Dachahu sublake	28	7	0	7	42
(d) Dahuchi sublake	11	6	3	7	27
Total	46	19	12	17	94

Note: The locations of sample fields are shown in Fig. 1. C.C, Pha.A, A.S, and Mis.S refer to *Carex cinerascen*, *Phalaris arundinace*, *Aremisia selengensis*, and *Miscanthus saccharifloruss* communities, respectively.

types of bottomlands (Fig. 1). Once the sample plot was located, we recorded its geographic coordinates and elevation through GPS (Trimble) with an accuracy of 0 to 0.20 m for position and 0.10 m for elevation. Then, all plant types in the plot were identified, recorded, and excavated. All dead materials were removed from clipped plants and fresh biomass was measured immediately using a digital scale. Then, the average fresh AGB per plot was calculated from these measurements ($n = 3$). The Landsat 8 image, which was acquired from USGS,[54] with 11 bands (bands 1 to 7 and 9 to 11 with a spatial resolution of 30 m and band 8 with a spatial resolution of 15 m) from December 16, 2016, covering the PLNNR, was used to complete the study.

2.3 *Data Preprocessing and Preparation*

Image preprocessing, executed via ENVI 5.2, included geometric, radiometric, atmospheric corrections, and spatial subsets. On the basis of the georeferenced images of Poyang Lake, the root-mean-square errors (RMSEs) in the image registration were ensured at <0.3 pixel for the seven images. The FLAASH atmospheric correction module, a feature of ENVI 5.2, was used to finish the atmospheric correction. NDVI, SAVI, EVI, and the second band from the Kauth–Thomas transformation[55] were added to the Landsat 8 OLI image with 7 multispectral bands by layer stacking to create an 11-band layer-stacked Landsat image. Then, the layer-stacked image from December 16, 2016, was used to extract 94 spectral sampling points, based on geographical coordinates recorded by GPS in preparation for image classification. The proposed methods are briefly explained in the flowchart (Fig. 2).

2.4 *AGB Model Methods*

We selected eight variables, namely, NDVI, SAVI, EVI, B3 (red band), B4 (NIR band), greenness (the second band from the Kauth–Thomas transformation), B6 (SWIR1, short-wave infrared 1 band), and B7 (SWIR2, short-wave infrared 2 band), as inputs. The descriptions and computational formulas of four VIs in this study are shown in Table 2. The variable values of 94 sampling points were extracted in accordance with their geographic coordinates. The effectiveness of LR, BPNN, SVR, and RF models in estimating AGB in the PLNNR was evaluated. Then, we utilized the trained models with the highest testing accuracy to map AGB in the PLNNR. We used spectral features as predictors to improve the accuracy of the models.

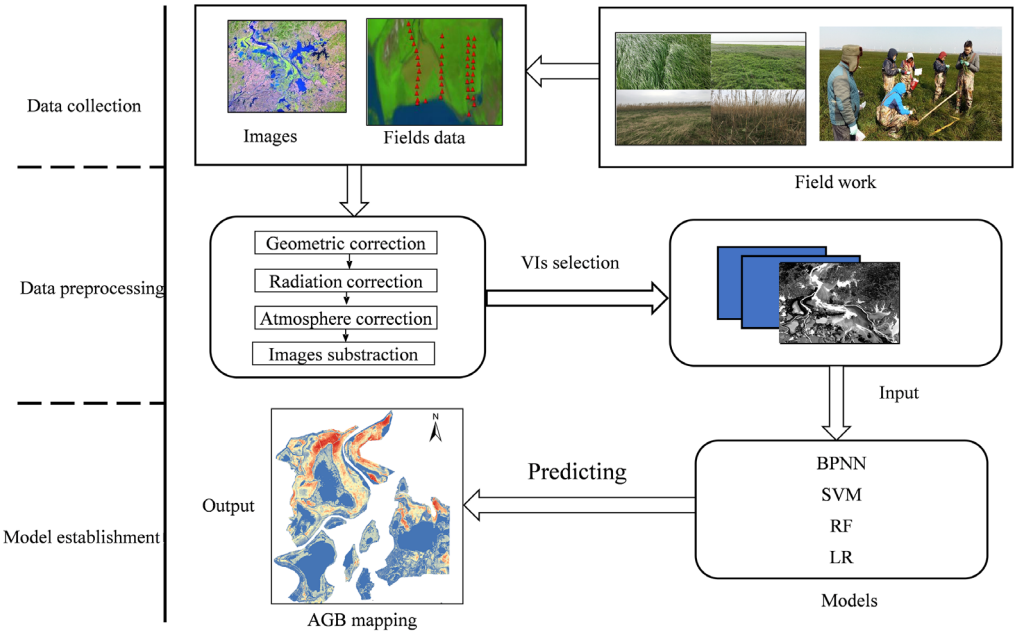

Fig. 2 Flowchart used to map AGB in Poyang Lake using Landsat images.

2.4.1 *RF model*

The RF model is an ensemble learning technique developed by Breiman to improve the classification and regression tree method by combining a large set of decision trees.[39] In RF regression, each tree is constructed by selecting a random set of variables and a random sample from the training dataset via a deterministic algorithm. Three parameters must be optimized in this model: (1) n_{tree}, the number of regression trees grown based on a bootstrap sample of the observations, with a default value of 500 trees; (2) m_{try}, the number of predictors tested at each node, with a default value 1/3 of the total number of variables; and (3) node size, the minimum size of the terminal nodes of the trees. To determine the n_{tree} and m_{try} values that can most accurately predict the wetland biomass, the two parameters were optimized on the basis of the RMSE. In addition, as the importance of each predictor is measured by an increase of mean squared errors and node purity, we excluded these predictors individually from the RF models. In our study, we tried 500 parameter sets, including n_{tree}, m_{try}, and node size, for the RF model and selected the one with the highest accuracy.

2.4.2 *SVR model*

Support vector machine (SVM) is a supervised nonparametric statistical learning technique, with no requirement for data distribution. The SVM can solve regression problems, which are generally regarded as the SVR. The two advantages of this technique include unique and globally optimal architectures and its easily accepted results. Nonlinear SVR maps input data X to a high-dimensional feature space using a kernel function. For our study, we utilized the commonly used RBF kernel, because it is associated with fewer numerical difficulties than any other kernel. Given the training data $[(x_1, y_1), (x_2, y_2) \ldots (x_n, y_n)]$, where x_i and y_i are the input and output data, respectively, we used ε-SVR to determine the function $f(x)$ with the most ε deviation from the input data and that is as flat as possible. The RBF kernel formula is as follows:

$$f(x, w) = \sum_{j=1}^{n} w_j \exp(-\gamma \|x - x_j\|^2), \tag{1}$$

where γ is a parameter and vector x_j denotes the training data input. The unknown vector of w is determined to minimize the function:

$$\min_{w \in R} \cdots \frac{1}{2}\|w\|^2 + C * \sum_{1=1}^{N} \max[|y_i - f(x_i, w) - \varepsilon, 0], \tag{2}$$

where cost $(C) > 0$ controls the trade-off in the flatness of $f(x)$, and deviations greater than ε are tolerated. Further details are provided by Awad and Khanna.[56] We adopted the most commonly used method, in which γ, C, and ε are calibrated to a certain range by a grid search. Similarly, 500 pairs of parameters were tried, and the set with the best performance is selected.

2.4.3 *BPNN model*

BPNN has a good generalization capability,[57,58] and it consists of input, hidden, and output layers, including their nodes and activation functions. The main mathematical expression of BPNN is as follows:

$$y_j = f\left(\sum_{i}^{n} w_{ji} x_i + b_j\right), \tag{3}$$

x_i is the i'th node value of the previous layer, y_j denotes the j'th node value of the present layer, w_{ji} represents the weighted value connecting x_i and y_j, n refers to the number of nodes in the previous layer, and f indicates the activation function. The BPNN model is discussed in further detail by Buscema.[59] Levenberg–Marquardt algorithm was used to determine the weighting and bias matrices for each iteration. We selected a bagging method (n_estimators: 400, max_samples:0.2) to ensure the stability and robustness of the trained model.

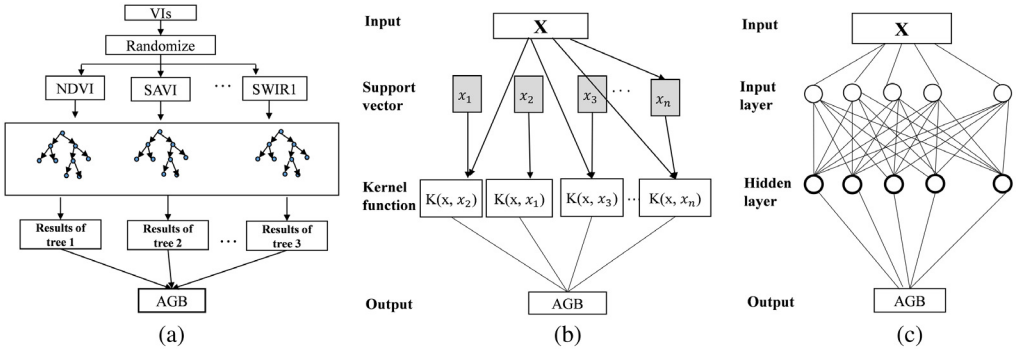

Fig. 3 Structures of: (a) RF, (b) SVR, and (c) BPNN models for estimating the AGB of the PLNNR.

Table 2 Descriptions and computational formulas of four VIs in this study.

VI	Description	Calculation
NDVI	A remote sensing index that reflects the state of land cover vegetation. It is defined as the quotient of the difference and sum between the reflectance of the NIR and visible light channels	$\mathrm{NDVI} = \frac{\rho_{\mathrm{NIR}}-\rho_{\mathrm{RED}}}{\rho_{\mathrm{NIR}}+\rho_{\mathrm{RED}}}$
SAVI	An SAVI based on NDVI and a large number of observations to reduce soil background effects	$\mathrm{SAVI} = \frac{(\rho_{\mathrm{NIR}}-\rho_{\mathrm{RED}})*(1+L)}{\rho_{\mathrm{NIR}}+\rho_{\mathrm{RED}}+L}$
EVI	An optimized VI that increases sensitivity to high-biomass areas and improves vegetation monitoring by decoupling canopy background signals and reducing atmospheric effects	$\mathrm{EVI} = \frac{2.5*(\rho_{\mathrm{NIR}}-\rho_{\mathrm{RED}})}{\rho_{\mathrm{NIR}}+6*\rho_{\mathrm{RED}}-7.5*\rho_{\mathrm{BLUE}}}$
Greenness	The second component from the Kauth–Thomas transformation	

Figure 3 shows the structures of the three models for estimating the AGB of Poyang Lake.

2.4.4 *Accuracy assessment*

In this study, we implemented all four modeling methods through packages in Python: scikit-learn.[60,61] Considering that we do not have enough sample points for every year, we divided the 94 sample points from the field survey in 2016 into two parts: training (80%) and testing (20%) datasets. We used three criteria, namely, RMSE, coefficient of determination (R^2), and mean absolute error (MAE), to evaluate the performance of these models in predicting AGB. RMSE [Eq. (4)] is a standard metric for measuring the discrepancies between the simulated and actual AGB values; however, it is easily influenced by outliers.[62] Therefore, MAE [Eq. (5)] is suggested to be used with RMSE for determining the variations of errors in the model.[63] R^2 [Eq. (6)] is utilized to determine the collinearity between the predicted and observed AGB values. RMSE and MAE values close to 0 and an R^2 value close to 1 indicate that the model is an accurate predictor:

$$\mathrm{RMSE} = \sqrt{\sum_{i=1}^{n} \frac{(y_{is} - y_{it})^2}{n}}, \tag{4}$$

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n} |y_{is} - y_{it}|}{n}, \tag{5}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_{is} - y_{it})^2}{\sum_{1=1}^{n} (y_{is} - \bar{y}_{is})^2}. \tag{6}$$

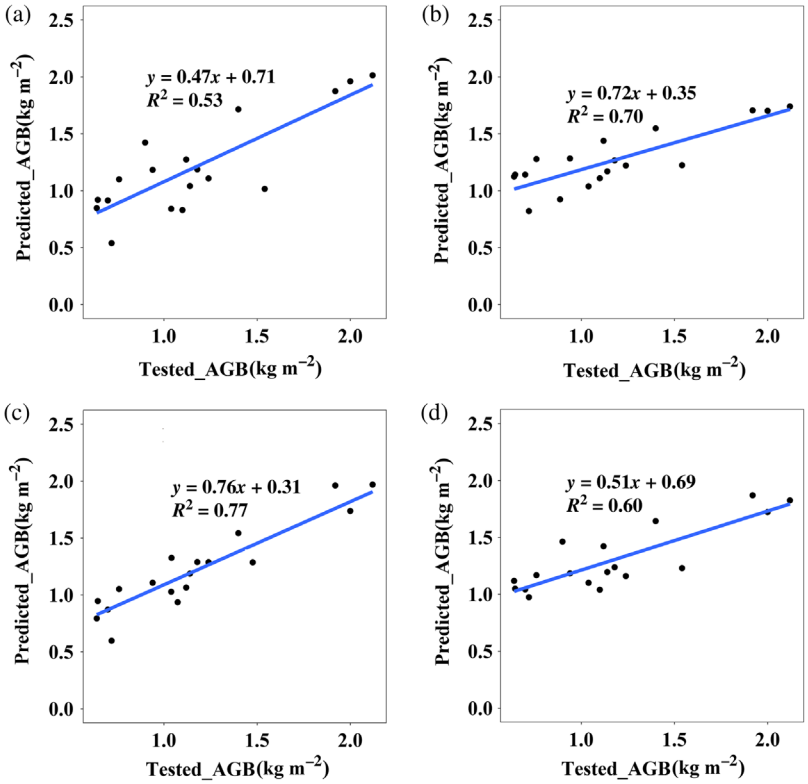

Fig. 4 Performance of the four models in estimating AGB in the testing dataset: (a) LR, (b) SVR, (c) RF, and (d) BPNN.

y_{is} is the i'th simulated AGB value, y_{it} denotes the i'th real AGB value among the tested sample points, $\bar{y}_{is}$ represents the mean simulated AGB for all tested sample points, and n indicates the size of tested samples. The RMSE, R^2, and MAE in Tables 3 and 5 have an average value after fivefold cross validation.

3 Results

3.1 *Comparison of the AGB Simulation Accuracy of Various Models*

Table 3 presents the specific value for the three criteria of four machine learning algorithms. Among the 76 sample points from the training dataset, the SVR had the lowest RMSE (0.25 kg m^{-2}) and the highest R^2 (0.84), and it also performed best in MAE (0.20 kg m^{-2}), which was considerably lower than that of the other three models. RF had the second best performance in training dataset: RMSE (0.30 kg m^{-2}), R^2 (0.71), and MAE (0.31 kg m^{-2}). The BPNN and LR were similar in magnitudes of RMSE, R^2, and MAE. For the 18 sample

Table 3 RMSE (kg m^{-2}), R^2, and MAE (kg m^{-2}) values of the four models for estimating AGB in the training and testing datasets.

	Training dataset			Testing dataset		
Model	RMSE	R^2	MAE	RMSE	R^2	MAE
LR	0.49	0.39	0.36	0.31	0.53	0.26
SVR	0.25	0.84	0.2	0.27	0.64	0.22
RF	0.3	0.71	0.31	0.25	0.70	0.21
BPNN	0.47	0.39	0.34	0.29	0.59	0.23

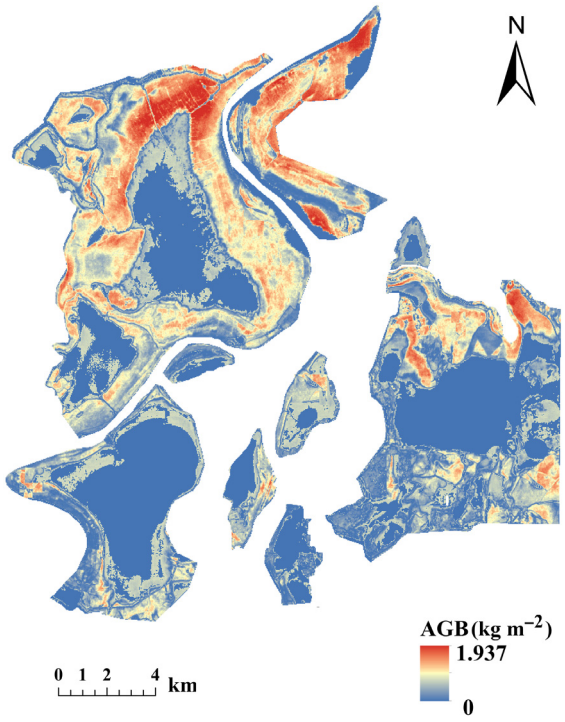

Fig. 5 AGB density distribution in the PLNNR.

points from the testing dataset, the RF model showed a better generalization capability than the SVR (RMSE, 0.25 versus 0.27 kg m^{-2}; R^2, 0.70 versus 0.64; MAE, 0.21 versus 0.22 kg m^{-2}). Finding an overfitting problem that occurs in the SVR is easy. Although the predicting capability of BPNN and LR in the training dataset almost had no difference, the R^2 (0.59) of BPNN is considerably higher than that of LR (0.53) in the testing dataset, indicating that BPNN had a relatively better generalization capability in this study. Figure 4 shows that the deviation between the simulated and actual values of RF had a relatively even distribution compared with the other models. From the scatter plot, the prediction values of the BPNN and LR had a relatively more dispersed distribution around the fitting line than those of the RF and SVR, indicating that these two models had worse stability of predictions. The least accurate model was the LR, with the highest RMSE of 0.31 kg m^{-2} and the lowest R^2 of 0.53 in the testing dataset. No discernible difference among the generalization capabilities of the three models except LR existed. Thus images with high spatial and spectral resolution might be essential to improve modeling accuracy. We concluded that RF was a slightly better model for predicting wetland AGB in the PLNNR that the other models.

3.2 *Predicting AGB in the PLNNR*

We utilized the most accurate model, namely, RF, for exploring the AGB distribution in the PLNNR (Fig. 5). The maps show that the AGB density is between 0 and 1.973 kg m^{-2}. On the whole, a higher than average AGB value occurred in the north part, including the Ganjiang River delta and Banghu sublake, whereas the south part, including bottomlands at the Dahuchi sublake experienced relatively low AGB values.

4 Discussion

4.1 *Accuracy and Uncertainty of the Study*

This study presents for the first time that a landscape-scale remote sensing model of the AGB for seasonal lake wetland in floodplain areas has been developed on the basis of machine learning

Table 4 Comparison of the accuracy of various models for simulating AGB in wetlands.

Methods	RMSE ($kg\,m^{-2}$)	References
LR	0.5 to 0.7	64 and 65
BPNN	0.3 to 0.4	66 and 67
SVR	0.35	67
RF	0.44	45

algorithms and Landsat images. We used RMSE to compare the predictive performance of the RF model to that of other models (Table 4). During our remote sensing analysis of wetland biomass, we concluded that simple LR had the worst performance regarding simulation effects, whereas the further advanced machine learning algorithm performed better with regard to RMSE. In our study, the RF model had a 0.21 $kg\,m^{-2}$ RMSE value in the testing dataset, which is considerably lower than the mean level of ~0.3 to 0.5 $kg\,m^{-2}$. Researchers might select different input variables; thus, the final simulated results would be affected by the randomness of sample points and the species types in wetlands. However, we cannot ignore that the RF model is useful and effective for predicting AGB in wetlands.

The time inconsistency of remote sensing imaging and sampling may cause the error on AGB inversion. For example, the December 16 scene is the only appropriate Landsat image closest to our field survey time in 2016. The atmospheric conditions of imaging time can affect the gray value of each pixel, thereby resulting in a heterogeneous gray value of the same vegetation type on the ground in different times. Thus, the training model generalization capability will be reduced.

4.2 *Implications of the Input Variables for Modeling AGB*

The rank of a feature used as a decision node in a tree can be utilized to assess the relative importance of that feature with respect to the predictability of the target variable. Features at the top of the tree greatly influence the final prediction decision of a large fraction of the input samples. Thus, the expected percentage of the samples that they contribute to can be used to estimate the relative importance of the features. By averaging the expected activity rates over several randomized trees, one can reduce the variance of such an estimate and use it for feature selection. Fig. 6 shows the results of applying RF with least squares loss and 500 base learners to the AGB in Poyang Lake wetland. Plot (a) shows the training and

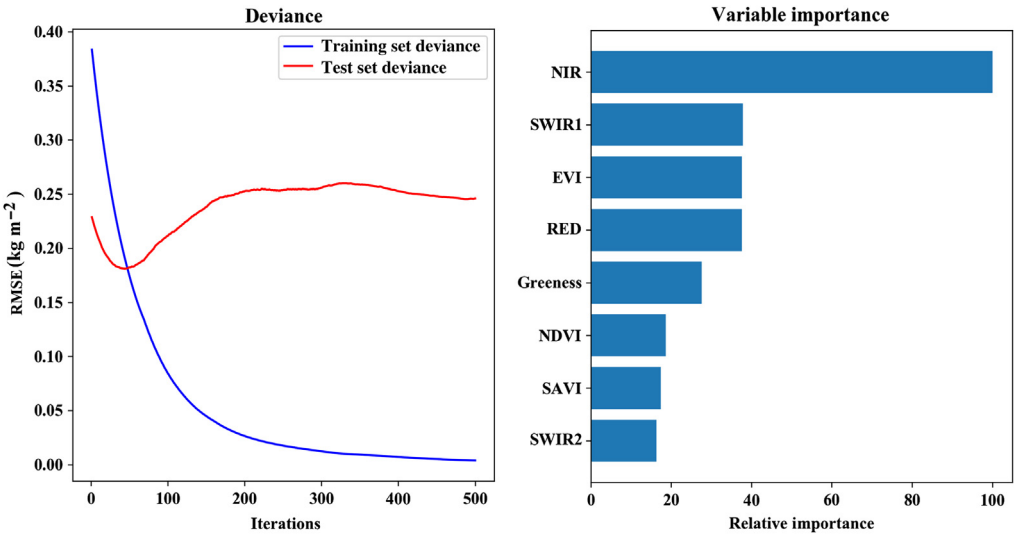

Fig. 6 Relative importance of all input variables identified using the RF model in this study.

Table 5 RMSE ($kg\,m^{-2}$), R^2, and MAE values ($kg\,m^{-2}$) of the four models for predicting AGB in training and testing datasets using NIR as input variable.

	Training dataset			Testing dataset		
Model	RMSE	R^2	MAE	RMSE	R^2	MAE
LR	0.52	0.31	0.38	0.30	0.55	0.25
SVR	0.37	0.66	0.28	0.34	0.45	0.27
RF	0.50	0.36	0.36	0.30	0.48	0.23
BPNN	0.54	0.28	0.39	0.33	0.46	0.26

testing errors at each iteration. Plot (b) shows the feature importance, which can be obtained via the feature importance property. Details on how the RF determined the importance of variables are discussed by Genuer et al.[68] When the number of trees is close to 450 or more, the test dataset error is ~0.2, which is the minimum value. Thus, we select 450 as the best N of the RF model in this study. Furthermore, the top four most important features were NIR, SWIR1, EVI, and red band. NIR performed best among all the variables, followed by SWIR1. NDVI was not the best predictive factor for estimating AGB in the PLNNR, possibly because it is considerably more affected by Poyang Lake wetland's complex environmental background conditions than the other modified VIs.

There is no doubt that EVI, SAVI, NDVI, NIR, and red band are highly correlated. However, in general, the models would be further effective for estimation if their input variables were independent of each other. Thus, we conducted an experiment to determine whether using the NIR alone as the input variable of these four models would produce further accurate results for estimating AGB in the PLNNR. Table 5 shows the specific performance. The RMSE, R^2, and MAE values changed remarkably in the training and testing datasets. The RMSE values for the training dataset increased by 0.11 $kg\,m^{-2}$ on average (RF, 0.15; SVR, 0.12; BPNN, 0.07; and LR, 0.03), whereas the R^2 values decreased. In the testing dataset, a slight improvement was observed in the LR. The RMSE and MAE decreased by 0.01 $kg\,m^{-2}$ and R^2 increased by 0.02, indicating that LR could not cope with the collinearity of variables and could not efficiently extract other variables' useful information. The RMSE and R^2 values of the other three models were increased by more than 0.05 $kg\,m^{-2}$ and reduced by 0.18 on average (RF, 0.22; SVR, 0.19; and BPNN, 0.13). Therefore, we insisted that the RF and SVR had better capability for processing further complicated information than the other methods. We concluded that placing these VIs into models together, which may decrease the influence of environmental background to some extent, improves the capability of the models for measuring AGB in the PLNNR. This improvement surpassed the influence of the collinearity of variables to some extent.

4.3 *Limitations of Predicting AGB Using Optical Remote Sensing Data such as Landsat*

The complexity of species composition and the density of vegetation in wetland areas present a huge challenge for remote sensing.[1] In fact, VIs calculated from broadband sensors will rapidly approach a saturation level when the AGB estimation is limited by the asymmetrical nature of the relationship between the AGB and VIs calculated from medium-spatial-resolution (10 to 100 m) multispectral sensors using NIR and red bands. Therefore, the RF model is likely to overestimate biomass at low observed values and underestimate biomass at high observed values, which may explain why errors are associated with high biomass values. Despite these limitations, our findings showed that the Landsat NIR band was sensitive to the AGB in the PLNNR. Recent efforts have been geared toward using narrow band VIs from hyperspectral data or WorldView-2 (eight bands including red edge band and 2-m spatial resolution) to estimate high canopy density biomass.[15,23,69,70] Results from these studies have shown that modified

VIs calculated from the red edge and NIR shoulder domains can more accurately estimate biomass at a full canopy cover than the standard red/NIR indices.[2,23] A reasonable explanation for this finding is that the indices calculated from the red edge are more sensitive to vegetation properties, such as canopy biomass and chlorophyll content, than that from other regions of the electromagnetic spectrum. A slight change in vegetation properties could result in a shift in the red edge curve, and NIR can minimize the influence of the atmospheric and water absorption as well as the soil background. However, the use of fine spatial and spectral resolution sensors (<5 m and >100 bands) for estimating AGB is limited by the cost, availability, and complexity of processing high-dimensional data.[70,71] This technology may be widely used in the future when costs decrease.

5 Conclusions

The quantitative estimation of wetland AGB is crucial for evaluating the health and conservation of this vital ecosystem. Traditional methods do not meet the requirements for rapid, accurate, and effective observation demands for a seasonal and changeable wetland such as Poyang Lake wetland. Therefore, numerous researchers are compelled to conduct an overall estimation of wetland AGB without considering the AGB information of different wetland vegetation communities. Landsat is a good remote sensing method alternative due to its relatively high spatial, spectral, and time resolution. In this study, we compared the performances of three machine learning algorithms, namely, RF, SVR, and BPNN, in estimating the AGB in the PLNNR. The RF model with 0.25 $\mathrm{kg\,m^{-2}}$ RMSE performed better than BPNN (0.29 $\mathrm{kg\,m^{-2}}$), SVR (0.27 $\mathrm{kg\,m^{-2}}$), and LR (0.31 $\mathrm{kg\,m^{-2}}$) models in the testing dataset. Furthermore, the AGB density in the PLNNR was found to be between 0 and 1.973 $\mathrm{kg\,m^{-2}}$ using the trained RF model to map the ABG distribution.

Our results indicated that RF had a relatively better generalization capability than LR, BPNN, and SVR in predicting AGB in the PLNNR. By considering the variable importance selection of the RF model, we regarded NIR, SWIR1, EVI, and red band as the most critical variables for estimating AGB in the PLNNR. Moreover, we found that the introduction of modified VIs can greatly improve the estimation accuracy, as opposed to only using NIR as the input variable. Furthermore, images with high spatial and spectral resolution are essential for improving AGB modeling precision and overcoming the saturation problem. This study presents an effective and operational RF model that estimates seasonal lake wetland AGB from Landsat-8 data, thereby providing a scientific basis for floodplain wetland carbon accounting.

Acknowledgments

This study was supported by the National Natural Science Foundation of China (No. 41571107), the Key Research Program of Chinese Academy of Sciences (No. KFZD-SW-318), and the Key Project of Water Resources Department of Jiangxi Province (No. KT201503).

References

1. E. Adam, O. Mutanga, and D. Rugege, "Multispectral and hyperspectral remote sensing for identification and mapping of wetland vegetation: a review," *Wetlands Ecol. Manage.* **18**(3), 281–296 (2010).
2. J. Chen et al., "Estimating aboveground biomass of grassland having a high canopy cover: an exploratory analysis of in situ hyperspectral data," *Int. J. Remote Sens.* **30**(24), 6497–6517 (2009).
3. R. A. Fournier et al., "Mapping aboveground tree biomass at the stand level from inventory inf," *Can. J. For. Res.* **33**(10), 1846–1863 (2003).
4. B. R. Parresol, "Assessing tree and stand biomass: a review with examples and critical comparisons," *For. Sci.* **45**(4), 573–593 (1999).
5. E. S. Kasischke et al., "Effects of seasonal hydrologic patterns in south Florida wetlands on radar backscatter measured from ERS-2 SAR imagery," *Remote Sens. Environ.* **88**(4), 423–441 (2003).

6. Z. J. Bortolot and R. H. Wynne, "Estimating forest biomass using small footprint LiDAR data: an individual tree-based approach that incorporates training data," *ISPRS J. Photogramm. Remote Sens.* **59**(6), 342–360 (2005).
7. S. A. Hall et al., "Estimating stand structure using discrete-return lidar: an example from low density, fire prone ponderosa pine forests," *For. Ecol. Manage.* **208**(1), 189–209 (2005).
8. M. A. Lefsky et al., "Lidar remote sensing of above-ground biomass in three biomes," *Global Ecol. Biogeogr.* **11**(5), 393–399 (2002).
9. Q. Wang and J. J. Liao, "Estimation of wetland vegetation biomass in the Poyang Lake Area using Landsat TM and Envisat ASAR Data," *Proc. SPIE* **7841**, 78411D (2010).
10. S. L. Powell et al., "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 (2010).
11. M. E. J. Cutler et al., "Estimating tropical forest biomass with a combination of SAR image texture and Landsat TM data: an assessment of predictions between regions," *ISPRS J. Photogramm. Remote Sens.* **70**(3), 66–77 (2012).
12. N. I. Gasparri et al., "Assessing multi-temporal landsat 7 ETM+ images for estimating above-ground biomass in subtropical dry forests of Argentina," *J. Arid Environ.* **74**(10), 1262–1270 (2010).
13. M. Mainknorn et al., "Evaluating the remote sensing and inventory-based estimation of biomass in the Western Carpathians," *Remote Sens.* **3**(7), 1427–1446 (2011).
14. X. Zhu and D. Liu, "Improving forest aboveground biomass estimation using seasonal Landsat NDVI time-series," *ISPRS J. Photogramm. Remote Sens.* **102**, 222–231 (2015).
15. E. M. I. Adam and O. Mutanga, "Estimation of high density wetland biomass: combining regression model with vegetation index developed from Worldview-2 imagery," *Proc. SPIE* **8531**, 85310V (2012).
16. P. J. Curran, J. L. Dungan, and H. L. Gholz, "Seasonal LAI in slash pine estimated with Landsat TM," *Remote Sens. Environ.* **39**(1), 3–13 (1992).
17. J. R. Jensen et al., "The measurement of mangrove characteristics in southwest Florida using SPOT multispectral data," *Geocarto Int.* **6**(2), 13–21 (1991).
18. X. Li et al., "Regression and analytical models for estimating mangrove wetland biomass in South China using Radarsat images," *Int. J. Remote Sens.* **28**(24), 5567–5582 (2007).
19. A. R. Huete, "A soil-adjusted vegetation index (SAVI)," *Remote Sens. Environ.* **25**(3), 295–309 (1988).
20. J. Qi et al., "A modified soil adjusted vegetation index," *Remote Sens. Environ.* **48**(2), 119–126 (1994).
21. H. Q. Liu and A. Huete, "A feedback based modification of the NDVI to minimize canopy background and atmospheric noise," *IEEE Trans. Geosci. Remote Sens.* **33**(2), 457–465 (1995).
22. G. M. Foody et al., "Mapping the biomass of bornean tropical rain forest from remotely sensed data," *Global Ecol. Biogeogr.* **10**(4), 379–387 (2001).
23. O. Mutanga and A. K. Skidmore, "Narrow band vegetation indices overcome the saturation problem in biomass estimation," *Int. J. Remote Sens.* **25**(19), 3999–4014 (2004).
24. S. Nandy et al., "Neural network-based modelling for forest biomass assessment," *Carbon Manage.* **8**(4), 305–317 (2017).
25. S. W. Todd, R. M. Hoffer, and D. G. Milchunas, "Biomass estimation on grazed and ungrazed rangelands using spectral indices," *Int. J. Remote Sens.* **19**(3), 427–438 (1998).
26. S. Jacquemoud et al., "PROSPECT+SAIL models: a review of use for vegetation characterization," *Remote Sens. Environ.* **113**(2009), S56–S66 (2009).
27. Y. Tian et al., "Remote sensing estimation of the aboveground biomass of reed wetland in the Western Songnen Plain, China, based on MODIS data," *Acta Ecol. Sin.* **36**(24), 8071–8080 (2016).
28. P. Xie, B. He, and M. Xing, "Estimation above-ground biomass of wetland bulrush in Qaidam Basin, China, combining regression model with vegetation index," in *Int. Conf. Geoinformatics*, pp. 1–4 (2011).
29. H. White, "Learning in artificial neural networks: a statistical perspective," *Neural Comput.* **1**(4), 425–464 (1989).

30. Y. Guo et al., "Optimal support vector machines for forest above-ground biomass estimation from multisource remote sensing data," in *IEEE Int. Geoscience and Remote Sensing Symp.*, pp. 6388–6391 (2012).
31. C. Notarnicola et al., "Neural network adaptive algorithm applied to high resolution C-band SAR images for soil moisture retrieval in bare and vegetated areas," *Proc. SPIE* **7829**, 78290F (2010).
32. J. M. Kovacs and F. Flores, "Estimating leaf area index of a degraded mangrove forest using high spatial resolution satellite data," *Aquat. Bot.* **80**(1), 13–22 (2004).
33. C. Proisy, P. Couteron, and F. Fromard, "Predicting and mapping mangrove biomass from canopy grain analysis using Fourier-based textural ordination of IKONOS images," *Remote Sens. Environ.* **109**(3), 379–392 (2007).
34. C. A. D. Castro et al., "High-performance prediction of macauba fruit biomass for agricultural and industrial purposes using artificial neural networks," *Ind. Crops Prod.* **108**, 806–813 (2017).
35. Q. D. Yang, H. B. Gao, and W. J. Zhang, "Biomass concentration prediction via an input-weighed model based on artificial neural network and peer-learning cuckoo search," *Chemom. Intell. Lab. Syst.* **171**, 170–181 (2017).
36. S. X. Yang et al., "Modeling grassland above-ground biomass based on artificial neural network and remote sensing in the Three-River Headwaters Region," *Remote Sens. Environ.* **204**, 448–455 (2018).
37. H. Drucker et al., "Support vector regression machines," in *Proc. 9th Int. Conf. Neural Information Processing Systems*, vol. **9**, pp. 155–161 (1997).
38. N. R. A. Jachowski et al., "Mangrove biomass estimation in Southwest Thailand using machine learning," *Appl. Geogr.* **45**(45), 311–321 (2013).
39. L. Breiman, "Random forests," *Mach. Learn.* **45**(1), 5–32 (2001).
40. V. Avitabile et al., "Capabilities and limitations of Landsat and land cover data for above-ground woody biomass estimation of Uganda," *Remote Sens. Environ.* **117**(1), 366–380 (2012).
41. M. Main-Knorn et al., "Monitoring coniferous forest biomass change using a Landsat trajectory-based approach," *Remote Sens. Environ.* **139**(4), 277–290 (2013).
42. Q. Feng, J. Liu, and J. Gong, "Urban flood mapping based on unmanned aerial vehicle remote sensing and random forest classifier—a case of Yuyao, China," *Water* **7**(4), 1437–1455 (2015).
43. P. O. Gislason, J. A. Benediktsson, and J. R. Sveinsson, "Random forest classification of multisource remote sensing and geographic data," in *Proc. IEEE Int. Geoscience and Remote Sensing Symp. (IGARSS '04)*, vol. **1042**, pp. 1049–1052 (2004).
44. M. Pal, "Random forest classifier for remote sensing classification," *Int. J. Remote Sens.* **26**(1), 217–222 (2005).
45. O. Mutanga, E. Adam, and M. A. Cho, "High density biomass estimation for wetland vegetation using WorldView-2 imagery and random forest regression algorithm," *Int. J. Appl. Earth Obs. Geoinf.* **18**(1), 399–406 (2012).
46. K. B. Byrd et al., "A remote sensing-based model of tidal marsh aboveground carbon stocks for the conterminous United States," *ISPRS J. Photogramm. Remote Sens.* **139**, 255–271 (2018).
47. W. Ji, "Biological resources survey in the international important wetland—The Poyang Lake National Nature Reserve," in *The Wetlands of Jiangxi Province*, X. Liu, Ed., China forestry Publishing House, Beijing (2000).
48. International Union for Conservation of Nature and Natural Resources (IUCN) 2018, "The IUCN Red List of Threatened Species," Version 2018-2, http://www.iucnredlist.org/ (10 May 2018).
49. G. F. Wu et al., "Will the three Gorges Dam affect the underwater light climate of Vallisneria spiralis L. and food habitat of Siberian crane in Poyang Lake?" *Hydrobiologia* **623**(1), 213–222 (2009).
50. Z. L. Huang et al., "Analysis of the correlations between environmental factors and rare cranes in the Poyang Lake region of China," *J. Great Lakes Res.* **44**(1), 140–148 (2018).

51. Ramsar Convention Secretariat, *Designating Ramsar Sites: Strategic Framework and guidelines for the future development of the List of Wetlands of International Importance*, Ramsar handbooks for the wise use of wetlands, 4th ed., vol. **17**, Ramsar Convention Secretariat, Gland, Switzerland (2010).
52. X. Dai, R. R. Wan, and G. S. Yang, "Non-stationary water-level fluctuation in China's Poyang Lake and its interactions with Yangtze River," *J. Geogr. Sci.* **25**(3), 274–288 (2015).
53. X. Dai et al., "Responses of wetland vegetation in Poyang Lake, China to water-level fluctuations," *Hydrobiologia* **773**(1), 35–47 (2016).
54. U.S. Department of the Interior, "The EarthExplorer (EE) tool, USGS science for a changing world," https://earthexplorer.usgs.gov/ (10 January 2018).
55. R. J. Kauth and G. S. Thomas, "The tasseled cap—a graphic description of the spectral-temporal development of agricultural crops as seen by Landsat," in *Proc. Symp. Machine Processing of Remotely Sensed Data*, pp. 4b41–44b51 (1976).
56. M. Awad and R. Khanna, "Support vector regression," *Neural Inf. Process. Lett. Rev.* **11**(10), 203–224 (2007).
57. W. B. Chen, W. C. Liu, and M. H. Hsu, "Predicting typhoon-induced storm surge tide with a two-dimensional hydrodynamic model and artificial neural network model," *Nat. Hazards Earth Syst. Sci.* **12**(12), 3799–3809 (2012).
58. D. E. Rumelhart, G. E. Hinton, and R. J. Williams, *Learning Representations by Back-Propagating Errors*, MIT Press, Cambridge (1988).
59. M. Buscema, "Back propagation neural networks," *Subst. Use Misuse* **33**(2), 233–270 (1998).
60. Scikit-learn developers, "Scikit-learn: machine learning in python," scikit-learn 0.19.0, http://scikit-learn.org/stable/ (15 March 2018).
61. F. Pedregosa et al., "Scikit-learn: machine learning in python," *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
62. T. Chai and R. R. Draxler, "Root mean square error (RMSE) or mean absolute error (MAE)?" *Geosci. Model Dev.* **7**(1), 1247–1250 (2014).
63. W. G. Pollett et al., "Spatial prediction models for shallow landslide hazards: a comparative assessment of the efficacy of support vector machines, artificial neural networks, kernel logistic regression, and logistic model tree," *Landslides* **13**(2), 361–378 (2016).
64. R. Li and J. Liu, "An estimation of wetland vegetation biomass in the Poyang Lake using Landsat ETM Data," *Acta Geogr. Sin.* **5**(56), 532–540 (2001).(in Chinese with English abstract)
65. C. Lin, "Study on wetland information extraction and above ground biomass estimation supported by Worldvies-2 images," PhD Dissertation, Chinese Academy of Forestry, Beijing (2013).(in Chinese)
66. L. Dong, J. Liao, and G. Shen, "Neural network-based biomass estimation in the Poyang Lake wetland using Envisat ASAR Data," *Remote Sens. Technol. Appl.* **24**(3), 325–330 (2009).
67. T. D. Pham, K. Yoshino, and D. T. Bui, "Biomass estimation of Sonneratia caseolaris (l.) Engler at a coastal area of Hai Phong city (Vietnam) using ALOS-2 PALSAR imagery and GIS-based multi-layer perceptron neural networks," *Mapp. Sci. Remote Sens.* **54**(3), 329–353 (2017).
68. R. Genuer, J.-M. Poggi, and C. Tuleau-Malot, "Variable selection using random forests," *Pattern Recognit. Lett.* **31**(14), 2225–2236 (2010).
69. S. J. Hurcom and A. R. Harrison, "The NDVI and spectral decomposition for semi-arid vegetation abundance estimation," *Int. J. Remote Sens.* **19**(16), 3109–3125 (1998).
70. P. S. Thenkabail, R. B. Smith, and E. D. Pauw, "Hyperspectral vegetation indices and their relationships with agricultural crop characteristics," *Remote Sens. Environ.* **71**(2), 158–182 (2000).
71. D. Lu, "The potential and challenge of remote sensing-based biomass estimation," *Int. J. Remote Sens.* **27**(7), 1297–1328 (2006).

Rongrong Wan received her PhD in physical geography from the Chinese Academy of Sciences (CAS) in 2005. She is currently an associate professor at Nanjing Institute of

Geography and Limnology, CAS. Her research interests include land-use and land-cover changes (LULC) and their hydrological and ecological effects, assessment of watershed ecosystem services, and spatially distributed watershed modeling. She has obtained three grants from the National Science Foundation of China as a principal investigator.

Xiaolong Wang is an associate professor at the Nanjing Institute of Geography and Limnology, Chinese Academy of Sciences (CAS). He received his BS and MS degrees in ecology from Anhui Agricultural University in 2000 and Nanjing Agricultural University in 2003, respectively, and his PhD in ecology from the Research Center for Eco-Environmental Sciences, CAS. He is the author of more than 60 journal papers. His current research interests include wetland ecology and eco-hydrological processes.

Biographies of the other authors are not available.