forests

MDPI

Article

Estimating Forest Gross Primary Production Using Machine Learning, Light Use Efficiency Model, and Global Eddy Covariance Data

Zhenkun Tian [1], Yingying Fu [2,*], Tao Zhou [3,4], Chuixiang Yi [5,6], Eric Kutter [7], Qin Zhang [8] and Nir Y. Krakauer [6,9]

[1] School of Computer Science, China University of Labor Relations, Beijing 100048, China; tzhenkun@hotmail.com
[2] School of Mathematics and Statistics, Beijing Technology and Business University, Beijing 100048, China
[3] State Key Laboratory of Earth Surface Processes and Resource Ecology, Beijing Normal University, Beijing 100875, China; tzhou@bnu.edu.cn
[4] Key Laboratory of Environmental Change and Natural Disaster of Ministry of Education, Faculty of Geographical Science, Beijing Normal University, Beijing 100875, China
[5] School of Earth and Environmental Sciences, Queens College, City University of New York, New York, NY 11367, USA; chuixiang.yi@qc.cuny.edu
[6] Earth and Environmental Sciences Department, Graduate Center, City University of New York, New York, NY 10016, USA
[7] Barry Commoner Center for Health and the Environment, Queens College, City University of New York, New York, NY 11367, USA; eric.kutter@qc.cuny.edu
[8] Institution of Water and Environment Research, Dalian University of Technology, Dalian 116024, China; zhangqinhan@mail.dlut.edu.cn (Q.Z.); nkrakauer@ccny.cuny.edu (N.Y.K.)
[9] Department of Civil Engineering and NOAA-CREST, The City College of New York, New York, NY 10031, USA
* Correspondence: yingyingfu2015@hotmail.com

check for updates

Citation: Tian, Z.; Fu, Y.; Zhou, T.; Yi, C.; Kutter, E.; Zhang, Q.; Krakauer, N.Y. Estimating Forest Gross Primary Production Using Machine Learning, Light Use Efficiency Model, and Global Eddy Covariance Data. *Forests* **2024**, *15*, 1615. https://doi.org/10.3390/f15091615

Academic Editor: Bahman Jabbarian Amiri

Received: 7 August 2024
Revised: 7 September 2024
Accepted: 10 September 2024
Published: 13 September 2024

Abstract: Forests play a vital role in atmospheric CO_2 sequestration among terrestrial ecosystems, mitigating the greenhouse effect induced by human activity in a changing climate. The LUE (light use efficiency) model is a popular algorithm for calculating terrestrial GPP (gross primary production) based on physiological mechanisms and is easy to implement. Different versions have been applied for many years to simulate the GPP of different ecosystem types at regional or global scales. For estimating forest GPP using different approaches, we implemented five LUE models (EC-LUE, VPM, GOL-PEM, CASA, and C-Fix) in forests of type DBF, EBF, ENF, and MF, using the FLUXNET2015 dataset, remote sensing observations, and Köppen–Geiger climate zones. We then fused these models to additionally improve the ability of the GPP estimation using an RF (random forest) and an SVM (support vector machine). Our results indicated that under a unified parameterization scheme, EC-LUE and VPM yielded the best performance in simulating GPP variations, followed by GLO-PEM, CASA, and C-fix, while MODIS also demonstrated reliable GPP estimation ability. The results of the model fusion across different forest types and flux net sites indicated that the RF could capture more GPP variation magnitudes with higher R^2 and lower *RMSE* than the SVM. Both RF and SVM were validated using cross-validation for all forest types and flux net sites, showing that the accuracy of the GPP simulation could be improved by the RF and SVM by 28% and 27%.

Keywords: climate change; forest ecology; modeling; machine learning

1. Introduction

The carbon sequestration of terrestrial ecosystems, which have provided a net carbon sink for about 30% of CO_2 emitted by human activities, plays a significant role in mitigating climate warming caused by rising global average temperature [1]. Since the majority of estimated carbon sequestration occurs in forests, forests are considered to be the main nonoceanic force slowing the rate of CO_2 accumulation; thus, accurately estimating forest GPP has been a research hotspot of carbon cycle for decades [2,3]. During the past two decades, gross carbon sequestration by global forests were about 15.6 ± 49 Gt CO_2 equivalent per year [4]. It is hard to achieve the climate change mitigation goal without the contribution from forests in atmospheric greenhouse gas (GHG) removal. The models for estimating terrestrial gross primary production (GPP) developed in past decades include process-based methods and empirical methods [5–10]. Process-based models such as Biome-BGC require meteorological, eco-physiological, and soil input data because these algorithms involve physical and biological processes which control the exchanges of mass and energy, which makes it hard to drive these process models at large scales given data availability [11]. Empirical methods could reach higher accuracy with limited input data but are limited by the number and representativeness of training examples with no rigorous foundation of ecological theory, while suffering from poor generalizability [7].

The LUE model is a kind of process model which has relatively few parameters, enables easy data acquisition, and has low computational requirements and is thus commonly used to simulate GPP and seasonal variations at global or regional scales [8,12–18]. LUE is defined as the efficiency of the vegetation photosynthesis system to use absorbed solar energy to assimilate CO_2. In this type of model, GPP is treated as a dependent variable of the APAR (Absorbed Photosynthetically Active Radiation) and the efficiency of carbon uptake [6,9]. LUE models can be written as:

$$GPP = \varepsilon_g \times APAR \tag{1}$$

$$APAR = fPAR \times PAR \tag{2}$$

where $fPAR$ is the fraction of incident PAR (Photosynthetically Active Radiation, $\mathrm{MJm^{-2}}$) absorbed by the vegetation canopy at daily or monthly scales. $fPAR$ could be calculated by linear regression of remote sensing indices. The actual light use efficiency ε_g is down-regulated from its theoretical maximum LUE ε_0 $\left(\mathrm{gCm^{-2}day^{-1}MJ^{-1}}\right)$ by environmental conditions including water and temperature stress. It can be calculated as:

$$\varepsilon_g = \varepsilon_0 \times f(T, W, \ldots\ldots) \tag{3}$$

where $f(T, W, \ldots\ldots)$ is a scalar in the range (from zero to one) that describes the reduction in theoretical LUE for the reason of environmental conditions. The key point lies in how to define the environmental stress functions and integrate them to compute GPP, which is also the difference between different LUE models. Representative LUE models such as EC-LUE (Eddy Covariance–Light Use Efficiency) [19], GLO-PEM (GLObal Production Efficiency Model) [12], and VPM (Vegetation Photosynthesis Model) [20] differ in the structure of the regulation function, Equation (3). For example, VPM and GLO-PEM use the multiplicative method to integrate water and temperature stress, while EC-LUE takes only the most serious stress factor according to Liebig's law.

As a practical technology in scientific research and commercial use, machine learning has experienced rapid development in the past twenty years [21], and a series of machine learning methods has been developed and applied in many fields such as natural language processing, medical diagnosis, computer vision, remote sensing, and other interdisciplinary fields [22–25]. In the research field of geoscience, machine learning methods have been widely applied in the inversion of land surface parameters, especially in forest classification, forest aboveground biomass estimation, and other forest parameters estimation [26–28]. Previous research has shown that even a simple averaging or linear combination could improve the accuracy of estimation compared to single models, and machine learning methods such as RFs and SVMs can outperform linear combinations [29–33]. Although many LUE models have been developed and applied in terrestrial GPP estimation at ground-site, regional, or global scales [5,10,13,16,19,20,34], there is still a big gap between the actual and desired GPP estimation accuracy, and further improvements through machine learning are necessary for deeply understanding the contribution of forests in reducing greenhouse gas concentrations in the atmosphere. Thus, for the purpose of estimating accurate forest GPP, we ran five LUE models (CASA, GLO-PEM, EC-LUE, VPM, and C-Fix) and ensembled them to improve the capacity of forest GPP estimation across biomes by employing two widely used machine learning methods (SVMs and RFs) based on satellite, eddy covariance data, and the Köppen–Geiger climate zones. Our objectives were to (1) evaluate the five LUE models in estimating forest GPP with site observations; (2) fuse the five LUE models based on climate zones using an SVM and an RF; and (3) compare the accuracy of different machine learning methods with individual models to clarify the improvements raised by machine learning.

2. Materials and Methods

2.1. Data

2.1.1. FLUXNET2015 Data

The FLUXNET2015 dataset includes sites with records over twenty years (from 1991 to 2014) and shares data on water, CO_2, and energy exchange between the atmosphere and biosphere, and synchronous meteorological and biological observations at the ecosystem scale from 212 global sites [35]. Considering the data availability and the observation duration, 45 forest sites (see Figure 1 and Table S1) were selected in which observations covered at least 5 years and could provide input parameters for all five LUE models. These sites included ENF (evergreen needleleaf forest), EBF (evergreen broadleaf forest), DBF (deciduous broadleaf forest), and MF (mixed forest) biomes. All in situ data were available at the FLUXNET2015 website, and data quality was controlled under a standardized data processing [35]. The daily VPD (Vapor Pressure saturation Deficit), sensible heat, air temperature, latent heat, CO_2 mole fraction, shortwave radiation, and remote sensing data were utilized for driving single LUE models. The missing data were eliminated using the quality control flags. The shortwave radiation (SR) was used to compute photosynthetically active radiation (PAR) as SR × 0.45 [36]. Ground GPP products were used to validate single models and machine learning methods. After the data quality check, the average of nighttime and daytime ground GPP products was used to carry out validation at site and forest-type scales [34].

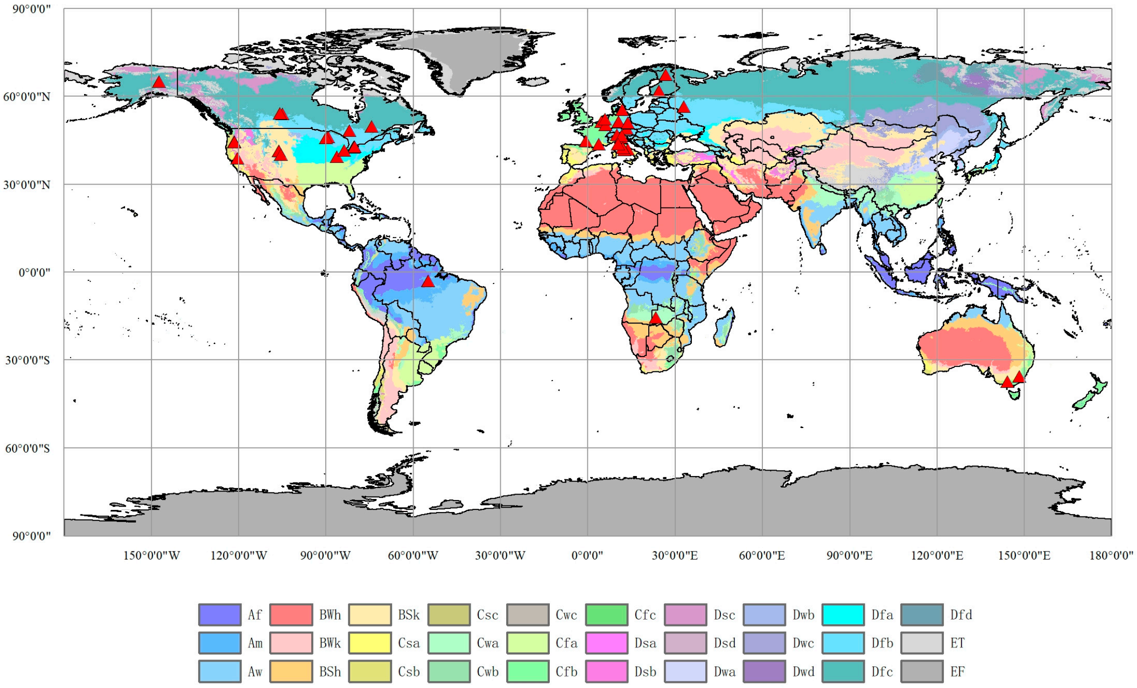

Figure 1. Köppen–Geiger climate zones and 45 FLUXNET2015 forest sites (red triangles) distribution. Köppen–Geiger climate symbols are listed in Table S2 in the Supporting Information File.

2.1.2. MODIS Data

MODIS has multispectral bands and provides visible light and thermal infrared data, with a higher resolution from 250 to 1000 m. MODIS data are widely used in fields such as monitoring global climate change, vegetation ecology, air pollution, land-use cover change, urban development, and environmental monitoring. In this study, MOD09A1 was used to provide land surface reflectance to calculate the vegetation/water index, and MOD15A2H was used as $fPAR$ in 8-day and 500 m spatiotemporal resolution during the period of ground-site observations. MOD17A2H, as a popular GPP product for research worldwide, was used as a reference value to compare with the results of both individual LUE models and the models fused by machine learning. ORNL DAAC (Oak Ridge National Laboratory's Distributed Active Archive Center, Oak Ridge, TN, USA) provides the download link and the Fixed Sites Subsets Tool for global users to match ground sites and corresponding MODIS pixel data. All missing or poor-quality data in MODIS products were removed based on the MODIS quality control flags. The MODIS products were interpolated temporally to daily time scale from composite products with a linear interpolation to match with ground-site observations.

2.1.3. Köppen–Geiger Climate Classification

The Köppen–Geiger climate classification is a partition that divides complex climate areas into climate zones with ecological meaning [37]. It has been widely used for many applications, for example, climate change impact assessments or ecological modeling. Beck et al., (2018) presented new and improved global Köppen–Geiger climate zone maps [37]. The high-precision current map (as shown in Figure 1) was derived from a set of high-resolution and terrain-corrected climate maps. The map of Köppen–Geiger climate zones was used as auxiliary data in fusing individual LUE models by machine learning in this study. The Köppen–Geiger system divides climate into 5 main classes and corresponding 30 sub-classes as shown in Figure 1 and Table S2. Further details of this classification can be found in Beck et al., (2018) [37].

2.2. LUE Models' Overview

Table 1 shows the functions of environmental constraints and structure of the 5 LUE models considered: EC-LUE, VPM, CASA, GLO-PEM, and C-Fix.

Table 1. A summary of the LUE models.

Number	Model Name	Equation	Environmental Scalars	Reference
1	CASA	$GPP = PAR \times fPAR \times \varepsilon_0 \times f(T_1) \times f(T_2) \times f(W)$	$f(T_1) = 0.8 + 0.02 \times T_{opt} - 0.0005 \times {T_{opt}}^2$ $f(T_2) = \frac{1.1814 \times \left(1 + e^{0.3 \times (-T_{opt} - 10 + T)}\right)}{1 + e^{0.2 \times (T_{opt} - 10 - T)}}$ $f(W) = \frac{1}{2} + \frac{EET}{2 \times PET}$	[8]
2	GLO-PEM	$GPP = PAR \times fPAR \times \varepsilon_0 \times f(T) \times f(\delta_q) \times f(\delta_\theta)$	$f(T) = \frac{(T - T_{min}) \times (T - T_{max})}{[(T - T_{min}) \times (T - T_{max})] - (T - T_{opt})^2}$ $f(\delta_q) = \begin{cases} 1 - 0.05 \times \delta_q & \left(0 < \delta_q \leq 15\right) \\ 0.25 & (\delta_q > 15) \end{cases}$ $\delta_q = Qw(T) - q$ $f(\delta_\theta) = 1 - exp(0.081 \times (\delta_\theta - 83.03))$	[12]

Table 1. *Cont.*

Number	Model Name	Equation	Environmental Scalars	Reference
3	C-Fix	$GPP = PAR \times fPAR \times \varepsilon_0 \times f(T) \times f(CO_2)$	$f(T) = \frac{e^{(C_1 - \frac{\Delta H_{a,P}}{R_g T})}}{1 + e^{(\frac{\Delta ST - \Delta H_{d,P}}{R_g T})}}$ $f(CO_2) = \frac{[CO_2] - \frac{[O_2]}{2\tau}}{[CO_2]^{ref} - \frac{[O_2]}{2\tau}} \times \frac{K_m\left(1 + \frac{[O_2]}{K_0}\right) + [CO_2]^{ref}}{K_m\left(1 + \frac{[O_2]}{K_0}\right) + [CO_2]}$	[15]
4	VPM	$GPP = PAR \times fPAR \times \varepsilon_0 \times f(T) \times f(W) \times f(P)$	$f(T) = \frac{(T - T_{min}) \times (T - T_{max})}{[(T - T_{min}) \times (T - T_{max})] - (T - T_{opt})^2}$ $f(W) = \frac{LSWI + 1}{LSWI_{max} + 1}$ $f(P) = \begin{cases} \frac{LSWI + 1}{2} (during\ leaf\ growth) \\ 1 \quad (after\ leaf\ growth) \end{cases}$ $LSWI = \frac{\rho_{nir} - \rho_{swir}}{\rho_{nir} + \rho_{swir}}$	[20]
5	EC-LUE	$GPP = PAR \times fPAR \times \varepsilon_0 \times min(f(T), f(W))$	$f(T) = \frac{(T - T_{min}) \times (T - T_{max})}{[(T - T_{min}) \times (T - T_{max})] - (T - T_{opt})^2}$ $f(W) = EF = \frac{LE}{LE + H}$	[19]

2.2.1. CASA

The CASA model was proposed by Potter et al., (1993) and can be applied to simulate seasonal patterns of NPP or GPP in monthly steps as a biosphere model with meteorological data and soil properties [8]. $f(T_1)$ and $f(T_2)$ account for high and low temperature stress, and T_{opt} and T represent the optimum and actual temperature, respectively. In the water stress factor $f(W)$, PET (potential evapotranspiration) can be calculated by the Priestley–Taylor model [38], and EET is the estimated evapotranspiration.

2.2.2. GLO-PEM

The GLO-PEM model was proposed by Prince & Goward (1995) [12]. It has been utilized to estimate global terrestrial GPP or NPP [12,39,40]. In Table 1, $f(T)$ defines the temperature regulation on the actual LUE. T_{opt}, T_{min}, and T_{max} denote the optimal, minimum, and maximum temperature for photosynthetic activities of vegetation, and T represents the actual temperature. Water limitation is determined by $f(\delta_q)$ and $f(\delta_\theta)$. In GLO-PEM, δ_q represents the specific humidity deficit; $Qw(T)$ denotes the saturated specific humidity at the air temperature; q represents the specific humidity of the air; and δ_θ is the soil water deficit in the top soil layer (1.0 m). Because of the lack of soil moisture deficit data in site observations, here, we adopted the water limitation factor of CASA for GLO-PEM.

2.2.3. C-Fix

The C-Fix model uses CO_2 fertilization effect and temperature constraint to simulate the CO_2 flux [15,41] and calculate GPP as shown in Table 2. In the temperature dependency factor $f(T)$, C_1 is a fixed number; T denotes the air temperature, $\Delta H_{a,P}$ is the activation energy, $\Delta H_{d,P}$ is the deactivation energy, R_g denotes the gas constant, and ΔS represents the denaturation equilibrium entropy of CO_2. In $f(CO_2)$, τ is the specificity ratio as CO_2/O_2; $[CO_2]$ is the actual CO_2 concentration; $[CO_2]^{ref}$ is the referenced CO_2 concentration; $[O_2]$ is the concentration of O_2; K_m represents the affinity constant of Rubisco for CO_2; and K_0 is the inhibition constant for O_2. These validated parameters can be obtained from the published paper [15].

Table 2. Look-up table of maximum LUE, optimal, minimum, and maximum temperature parameters for each forest type.

PFT	ε_0 (gC/m^2/d/MJ)	Tmin (°C)	Tmax (°C)	Topt (°C)
DBF	2.14 for EC-LUE 1.93 for other LUE Models	−1	40	20
EBF		−2	48	28
ENF		−1	40	20
MF		−1	48	19

2.2.4. VPM

In the VPM model proposed by Xiao et al., (2004), canopies consist of PAV (Photosynthetically Active Vegetation) and NPV (Non-Photosynthetic Vegetation), and only the PAR absorbed by PAV can be used for carbon uptake [20]. VPM uses the same temperature limitation factor as GLO-PEM and adopts the satellite-based $LSWI$ (land surface water index) associated with canopy and leaf moisture as the water stress. In the definition of the $LSWI$, ρ_{nir} and ρ_{swir} are the reflectance in NIR and SWIR bands from satellite data. VPM also considers the leaf growth, so during the leaf growth period, $f(P)$ is taken as $(1 + LSWI)/2$, as shown in Table 2.

2.2.5. EC-LUE

The EC-LUE model developed by Yuan et al., (2007) supposes that the capacity of vegetation photosynthetic activity is regulated by Liebig's law, which means the carbon sequestration is only calculated by the strongest constraint (i.e., the minimum of temperature and water limitation) at any given time [19]. The temperature stress, $f(T)$, is the same as in GLO-PEM. In the water stress function, LE denotes latent heat, and H represents sensible heat. As an effective index of land surface water conditions, EF can be derived from Bowen's ratio [42]. EC-LUE was taken as the GPP inversion algorithm in GLASS products and has been widely applied to simulate GPP at different spatial scales [43–45].

2.2.6. LUE Models' Parameterization

Generally, the maximum LUE (ε_0), which should be calibrated for sites or forest types, varies among different forest types, as well as other parameters such as optimal, minimum, and maximum temperatures of photosynthetic activities. In this study, a look-up table (Table 2) was adopted to define ε_0 based on the previous literature [19,34].

2.3. Machine Learning Methods

SVMs and RFs were employed in this study to fuse single LUE models and global Köppen–Geiger climate types at 45 forest FLUXNET2015 sites.

An SVM is a widely used classification method and processes multi-dimensional data according to structural risk minimization criteria and VC dimension theory. An SVM can be employed to solve nonlinear problems in a new feature space [46,47]. Considering a labeled training set $(x_i, y_i), i = 1, 2, 3, \cdots h$, where x_i denotes a vector ($x_i \in R^n$, and $y_i \in [-1, +1]$), to acquire the functional dependency $f(x) = (w \cdot x) + b$, the SVM requires the following expression to be minimized:

$$minimize\left(\frac{w^T w}{2} + C\sum_{i=1}^{h} \xi_i\right) \tag{4}$$

subject to $y_i\left(w^T\phi(x) + b\right) \geq 1 - \xi_i, \xi_i > 0$. In Equation (4), w represents a weight vector; b denotes the bias; C represents a cost parameter; $\phi(x)$ converts x_i into a higher-dimensional space; and ξ_i represents the error. In addition, $K\left(x_i, x_j\right) = \phi(x_i)^T\phi\left(x_j\right)$ is the kernel function. As a popular kernel in previous studies, the RBF (radial basis function) was chosen [33,48]. More SVM algorithm details can be found in the literature [47,49].

The RF classifier was developed by Breiman, (2001) and has been a widely used method in geographical fields for its robustness, easier implementation, and small number of hyperparameters [10,50–52]. An RF is famous as a classifier which includes many decision tree classifiers $\{h(x, \Theta_k), k = 1, 2, 3 \cdots\}$ where the item $h(x, \Theta_k)$ denotes a specific classifier; Θ_k is the random vector that is distributed in the kth place; and x is an input vector [50]. An RF is a robust machine learning method which includes many small decision trees and can be used as an ensemble method for different tasks including classification and regression.

For each forest type (i.e., DBF, EBF, ENF, and MF), the grid-search method was employed to identify the optimum parameters for the RF and SVM. We tried many trees (ntree, from 100 to 1000) and variables sampled as candidates at each split (mtry, from 1 to 5) separately for the RF across each forest type. Similarly, for the SVM, different values of the kernel parameter (gamma, from 0.1 to 10) and the cost of violating constraints (cost, from 0.1 to 10) were tried. The optimum parameters (see Table S4 in the Supporting Information File) were applied to the process of fusing CASA, GLO-PEM, C-Fix, VPM and EC-LUE models together with Köppen–Geiger climate zones. The workflow is shown in Figure 2.

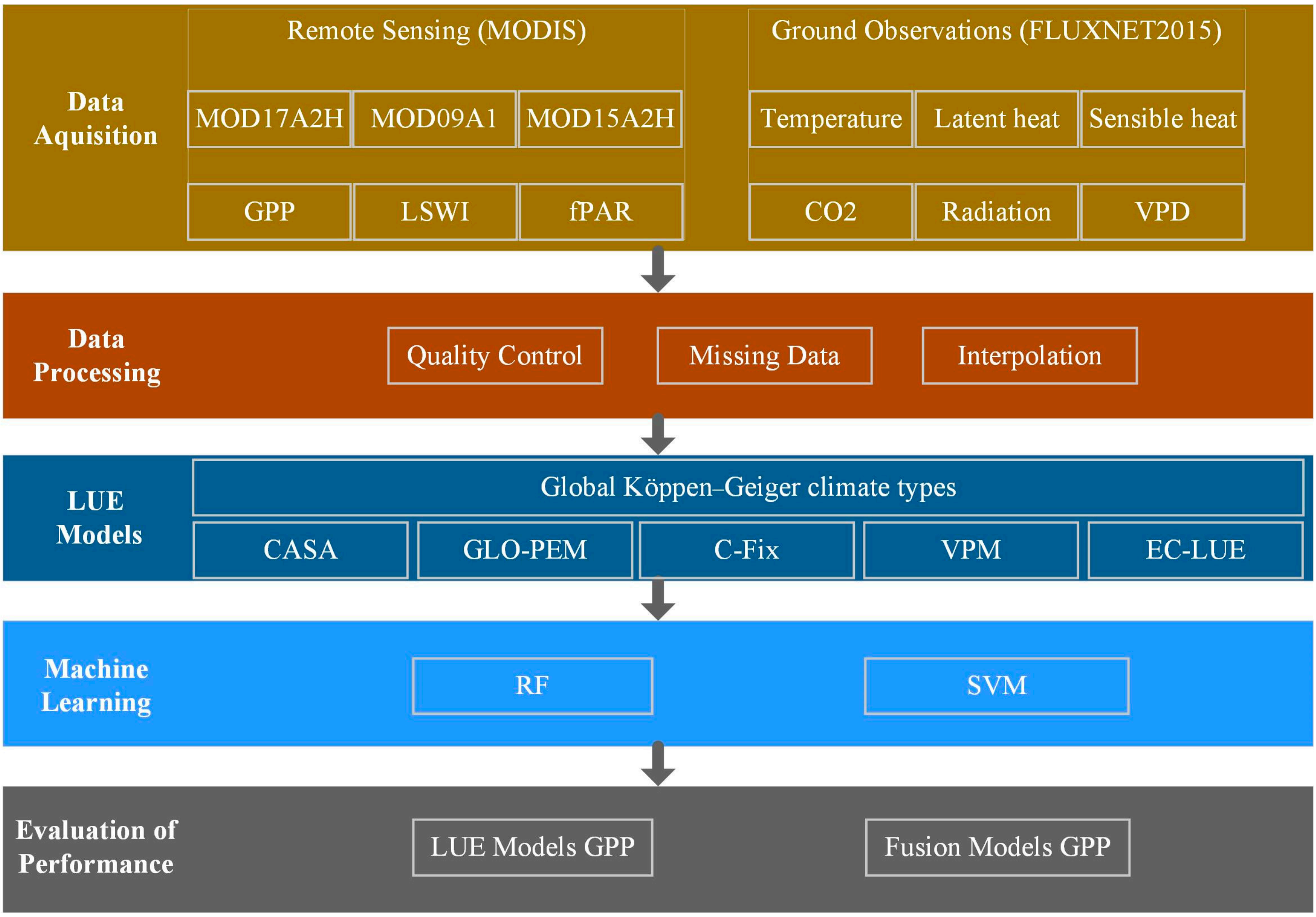

Figure 2. Workflow of GPP estimation through the integration of LUE models based on ground measurements, remote sensing observations, and Köppen–Geiger climate zones.

2.4. Statistical Analysis and Accuracy Validation

We used the metrics of R (correlation coefficient), SD (standard deviation), $RMSE$ (root-mean-square error), $RMSD$ (root-mean-square difference), R^2 (coefficient of determination), and *RPE* (relative predictive error) in Taylor diagrams [53] and scatter plots to evaluate the LUE models, SVM, and RF. The *RMSE*, *RMSD*, and *RPE* are defined as:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{m}(\hat{y}_i - y_i)^2}{m}} \tag{5}$$

$$RMSD = \sqrt{\frac{\sum_{i=1}^{m}\left[(\hat{y}_i - \overline{\hat{y}}) - (y_i - \overline{y})\right]^2}{m}} \tag{6}$$

$$RPE = \frac{\overline{\hat{y}} - \overline{y}}{\overline{y}} * 100\% \tag{7}$$

where $\hat{y}_i$ and y_i represent the GPP estimated by the LUE models or machine learning methods, and FLUXNET2015 sites' GPP; $\overline{\hat{y}}$ is the average of modeled values; $\overline{y}$ is the average of observed values; and m denotes the sample size. Better model performance is indicated by higher R or R^2, lower SD or *RMSE*/*RMSD*, and lower absolute *RPE*.

The RF and SVM were evaluated by a method named 5-fold cross validation. All data were partitioned randomly into five groups with equal samples, and one group was chosen as a validation set while the remaining four groups were taken as a training set. We trained models with four of the five groups and then validated GPP estimates of the machine learning by using the remaining group. Then, the average of R^2, *RMSE*, and *RPE* for the SVM and RF was calculated. We also used site-derived and simulated data to test two kinds of GPP variability across forest types and ground sites.

The index of *AIC* and *BIC* were used here to evaluate models from candidates. The definitions of *AIC* and *BIC* are as follows [54,55]:

$$AIC = -2 * ln(H) + 2 * k \tag{8}$$

$$BIC = -2 * ln(H) + k * ln(m) \tag{9}$$

where H denotes the maximum likelihood function, k represents how many parameters are needed, and m represents sample size. Lower values of *AIC* and *BIC* indicate a better choice.

3. Results and Discussion

3.1. GPP Estimation of LUE Models, SVM, and RF Based on Forest Types

Figure 3 shows Taylor diagrams for the site-derived GPP and estimated GPP using different LUE models at the 45 EC sites. The Taylor diagram is a graph in polar coordinate which includes the metric of standard deviation (*SD*), the correlation coefficient (*R*) and the root-mean-square difference (*RMSD*) between the estimations and the reference values [53]. In Figure 3, *SD* is the distance from the original coordinate point; R is the cosine of the azimuth angle; and *RMSD* denotes the distance from the reference value (i.e., FLUXNET2015's GPP products). Figure 3 indicates that the performance of different LUE models varied across forest types. In the deciduous broadleaf forest (DBF), the VPM estimation was the closest one to site-derived GPP compared to other LUE models including MODIS's GPP in the dimensions of *SD*, R, and *RMSD*. EC-LUE and MODIS held similar R and *RMSD* but higher and lower *SD* values compared to VPM. The *RMSD* of GLO-PEM was higher than that of VPM, EC, and MODIS, while CASA and C-Fix stayed further away from the referenced point (site-derived GPP), which means that CASA and C-Fix ranked lower compared to other LUE models. In the evergreen broadleaf forest (EBF), EC-LUE and MODIS outperformed other LUE models with a lower *RMSD*, higher R, and closer *SD* distance to site-derived GPP. In the ENF and MF types, EC-LUE, VPM, and MODIS performed better than GLO-PEM, CASA, and C-Fix.

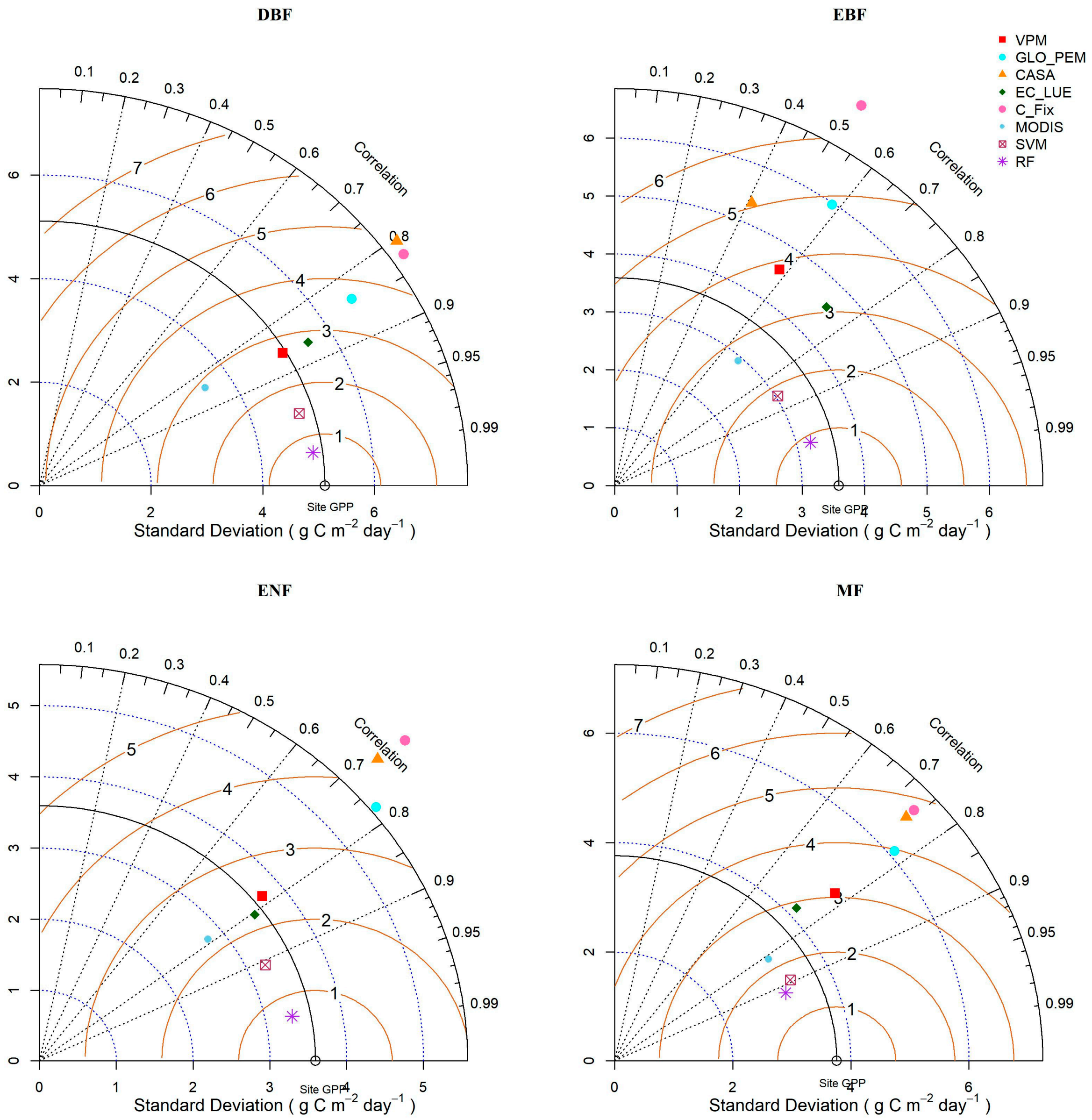

Figure 3. The Taylor diagrams for site-derived GPP and LUE models/machine learning estimates at the 45 FLUXNET2015 sites. The dotted circular lines which connect the X and Y axes denote *SD*. The dotted radial lines represent R. The brown curves are *RMSD* compared to the referenced site's GPP.

In Figure 3, the machine learning methods, SVM and RF, showed significant advantages in estimating forest GPP over individual LUE models. The *RMSD* of the SVM and RF were within 2 gC m^{-2}day^{-1} while that of individual LUE models ranged from 2 to 6 gC m^{-2}day^{-1}. The correlation coefficient (R) of the SVM and RF were obviously higher than that of individual LUE models in all four forest types. Most correlation coefficients of the SVM and RF across the DBF, EBF, ENF, and MF were higher than 0.9, while LUE models held lower R values below 0.9. The lowest R value of C-Fix was even below 0.5 with the highest *RMSD* in the EBF.

Figure 4 shows that the single LUE models, SVM, and RF exhibited substantial differences for each forest type using R^2, *RPE*, and *RMSE* compared to ground site-derived GPP. EC-LUE, VPM, and GLO-PEM explained major GPP variations across each forest type, indicated by R^2 of 0.62, 0.57 and 0.56, and C-Fix and CASA explained less variations with an average R^2 of 0.51 and 0.47. The R^2 of MODIS's GPP was 0.61, only less than EC-LUE, which indicated a good performance compared to other LUE models. The SVM and RF significantly improved the ability to explain variations for each forest type with R^2 of 0.79 and 0.80, respectively. The *RMSE* and *RPE* describe the errors from estimation, so lower values indicate better performance. In the five LUE models, EC-LUE had the smallest *RMSE* of 2.77 gC m^{-2} day^{-1}, while VPM had the least absolute *RPE* of −0.04, which means EC-LUE and VPM outperformed other LUE models in terms of simulation errors. The *RMSE* (gC m^{-2} day^{-1}) and *RPE* of C-Fix (5.79 and 0.53) and CASA (4.96 and 0.23) were much higher than those of EC-LUE and VPM. MODIS performed well with a 2.55 gC m^{-2} day^{-1} *RMSE* and *RPE* of −0.11, indicating that it was justifiably trustworthy as a popular worldwide research product. VPM, EC-LUE, and MODIS, as the better performance models, underestimated GPP compared to site observations. The SVM and RF decreased the *RMSE* (gC m^{-2} day^{-1}) to 1.75 and 1.72, and *RPE* to −0.01 and 0, respectively, which further demonstrated that machine learning could significantly improve GPP simulation capabilities.

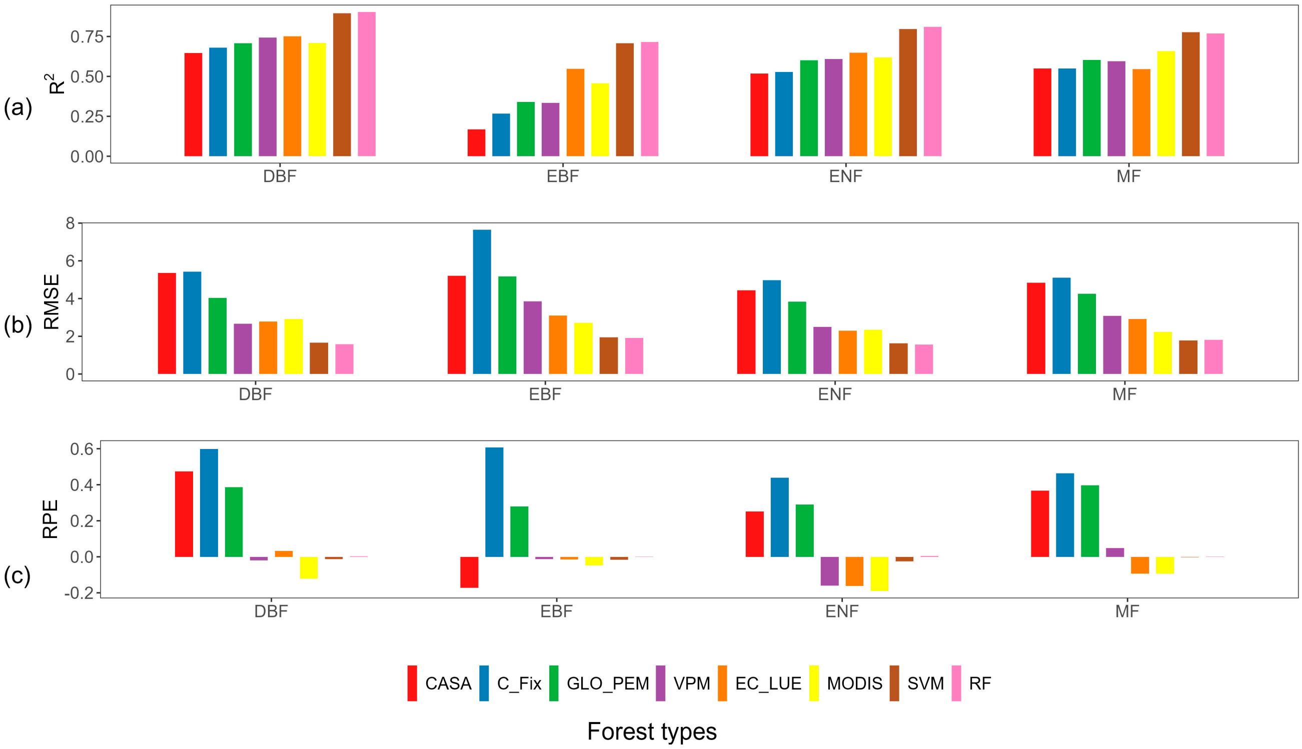

Figure 4. The R^2 (**a**), *RMSE* (**b**), and *RPE* (**c**) of 5 single models, MODIS, SVM, and RF across the DBF, EBF, ENF, and MF.

Figure 5 shows the performance of the five single LUE models, MODIS, SVM, and RF across all DBF flux sites. EC-LUE had the highest R^2 (0.75), followed by VPM (0.74), GLO-PEM (0.71), MODIS (0.71), C-Fix (0.68), and CASA (0.65), meanwhile, both SVM and RF increased the R^2 to 0.90, demonstrating the improved explanation of the GPP variation that machine learning could achieve. VPM gave the smallest *RMSE* and *RPE* (2.67 gC m^{-2} day^{-1}, and 0.02) compared to other LUE models, indicating the reliable performance of the GPP simulation in deciduous broadleaf forests. EC-LUE and MODIS also performed well with a lower *RMSE* (2.79 and 2.92) and *RPE* (0.03 and 0.12), followed by GLO-PEM, CASA, and C-Fix, as shown in Figure 5. The SVM and RF evidently narrowed the simulation scope, and the RF performed better with a lower *RMSE* and *RPE* than SVM (see Figure 5). We put the evaluation of the five LUE models, MODIS product, SVM, and RF across the EBF, ENF, and MF in the supporting information file, and similar results can be seen in Figures S1–S3.

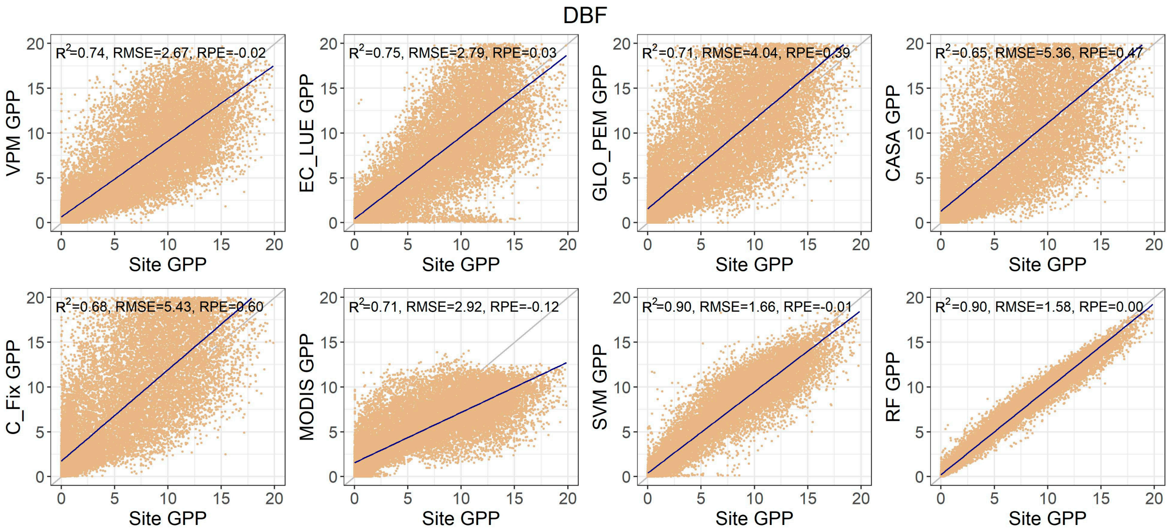

Figure 5. The scatter plots of R^2, *RMSE*, and *RPE* across the DBF between site-derived GPP and the estimates from the LUE models, MODIS, SVM, and RF.

Figure 6 shows the probability distribution of the predictive error in LUE models, MODIS, SVM, and RF compared to FLUXNET2015's GPP product. The VPM, EC-LUE, and GLO-PEM models had maximum value that centered on the x axis origin with thicker tails on the left hand, which showed that these three LUE models underestimated GPP compared to site observations. CASA and C-Fix had lower peaks compared to EC-LUE, VPM, and GLO-PEM, illustrating that the predictive bias from these two models fluctuated more strongly, which may lead to a higher probability of estimation errors. Compared to individual LUE models and MODIS's GPP product, the GPP error distributions of machine learning estimations were centered on the x axis origin closely, in a narrow range, with explicitly higher peaks and balanced biases on both sides. Therefore, these histograms indicate that the machine learning fusion strategy—especially the RF—were best at explaining the GPP variance.

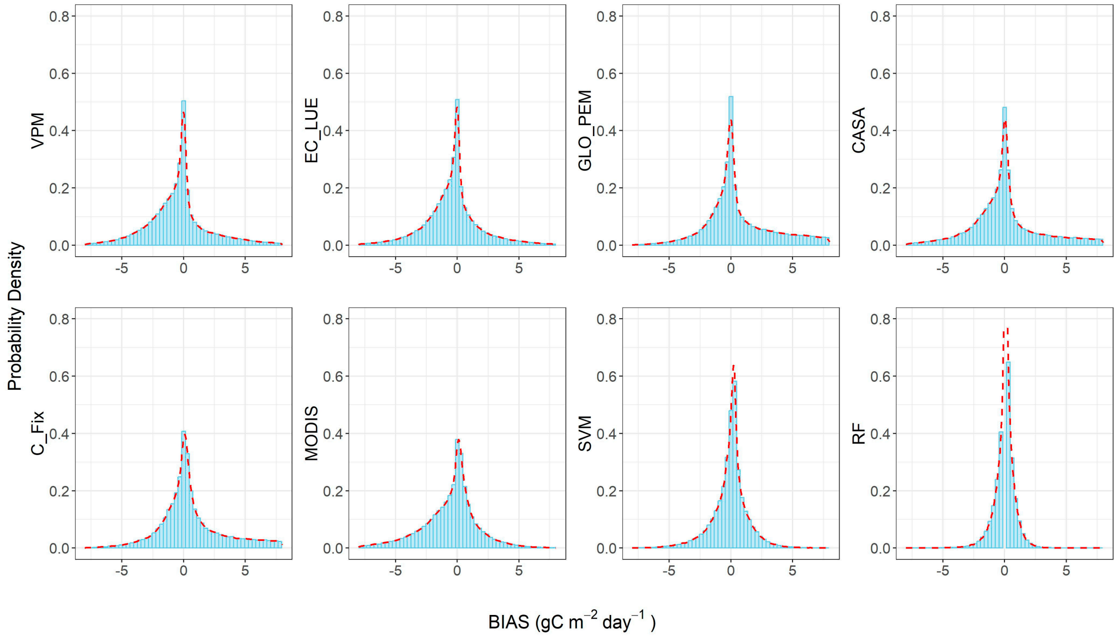

Figure 6. The probability distribution of errors from the LUE models, MODIS, SVM, and RF.

The values of the *AIC* and *BIC* across all forest types were calculated for each LUE and machine learning model based on Equations (8) and (9) in Section 2.4. Figure 7 shows that machine learning methods gave lower *AIC*/*BIC* values than individual LUE models, which indicated the SVM and RF were better alternatives than individual LUE models. EC-LUE and VPM outperformed other LUE models, which was consistent with previous results.

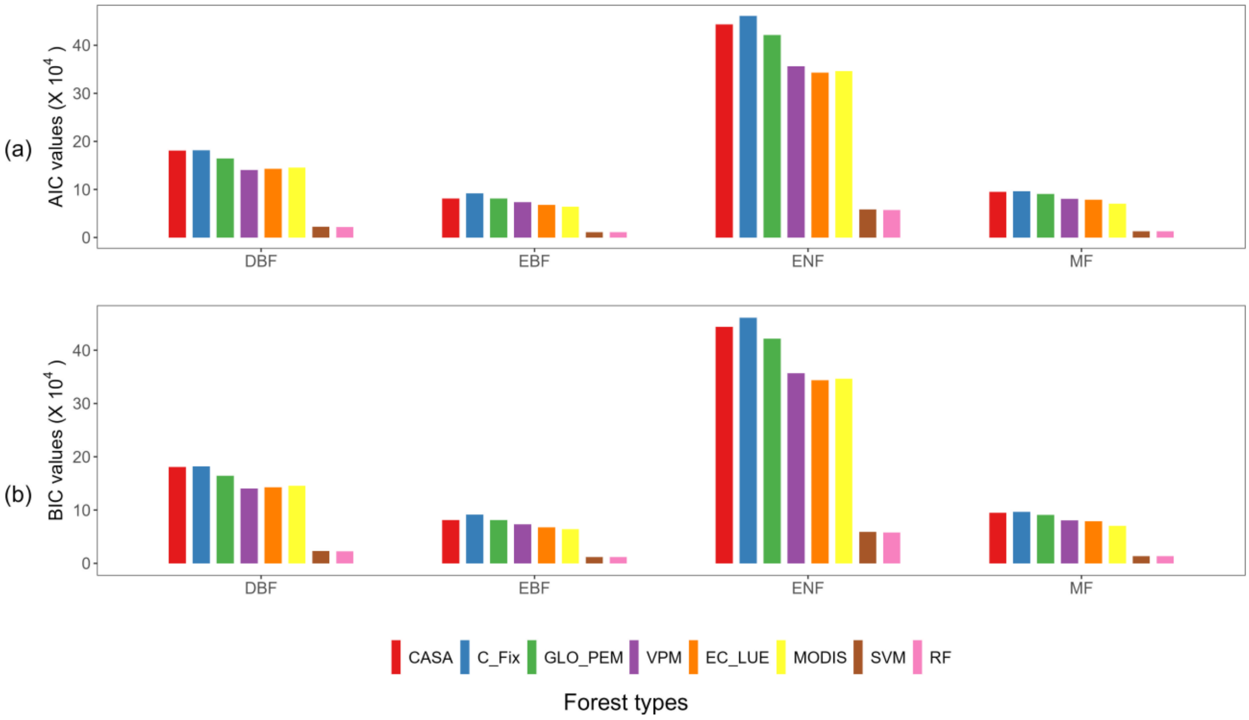

Figure 7. The *AIC* (**a**) and *BIC* (**b**) of the LUE models, MODIS, SVM, and RF across the DBF, EBF, ENF, and MF.

3.2. GPP Estimation of LUE Models, SVM, and RF on Forest Sites of FLUXNET2015

We evaluated each LUE model and machine learning-based fusion models for all sites using the R^2, *RMSE*, and *RPE* metrics, as shown in Table S3 in the supporting information file. Table S3 shows that EC-LUE and VPM outperformed other LUE models in most sites, and the RF outperformed all models involved in this study. Specifically, we picked out four typical DBF, EBF, ENF, and MF flux sites to illuminate the ability of simulating seasonal GPP variation by using the best LUE model (i.e., EC-LUE) and machine learning model (i.e., RF) compared to FLUNNET2015's GPP product. These representative sites were from Europe, the United States, and Australia for each forest type, and 6 years of data were extracted in long-term observation, as shown in Figure 8. Our results indicate that the difference between FLUXNET2015's GPP product and the RF estimation was dramatically less than the difference between FLUXNET2015's GPP and that of EC-LUE. This means that compared to EC-LUE, the RF could explain more seasonal variation in GPP.

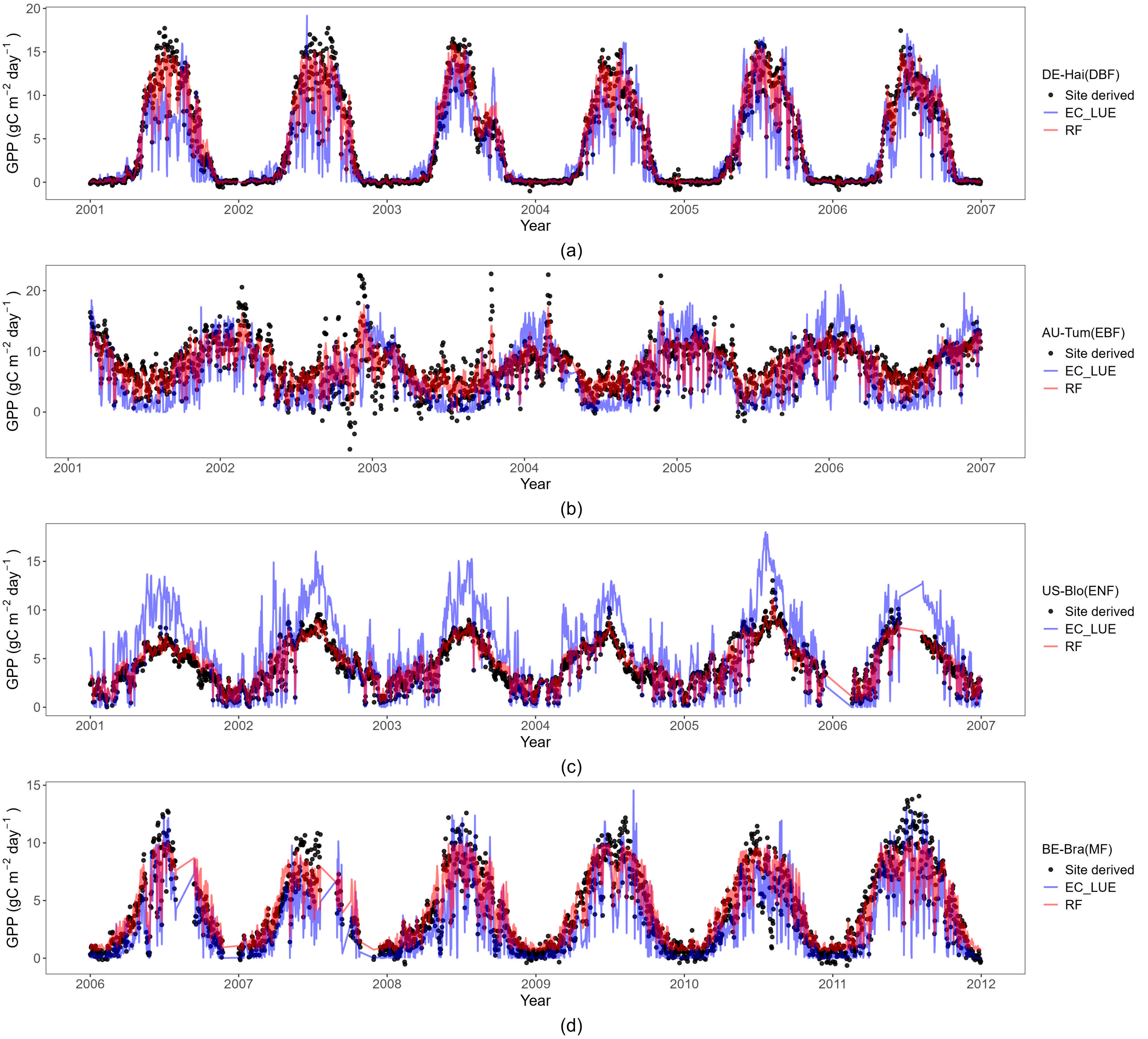

Figure 8. Daily FLUXNET2015's GPP (black dots), the best LUE model-estimated GPP (EC-LUE, line in blue), and the best fusion method-estimated GPP (RF, line in orange) at 4 sites: DE-Hai of DBF (**a**), AU-Tum of EBF (**b**), US-Blo of ENF (**c**), and BE-Bra of MF (**d**).

We drew a boxplot of R^2, *RMSE*, and *AIC* for each LUE and machine learning model across all selected FLUXNET2015 sites. A flat box and narrow range of outliers means reliable model performance in a boxplot. Figure 9 shows that EC-LUE and VPM outperformed other LUE models, while the SVM and RF could improve GPP estimation accuracy significantly. The RF had flatter boxes and a narrower outlier range than the SVM, especially an obvious distribution of a higher R^2 and a lower *RMSE* and *AIC* compared to that of the SVM, which indicated the RF was a better machine learning method to simulate forest GPP in this study.

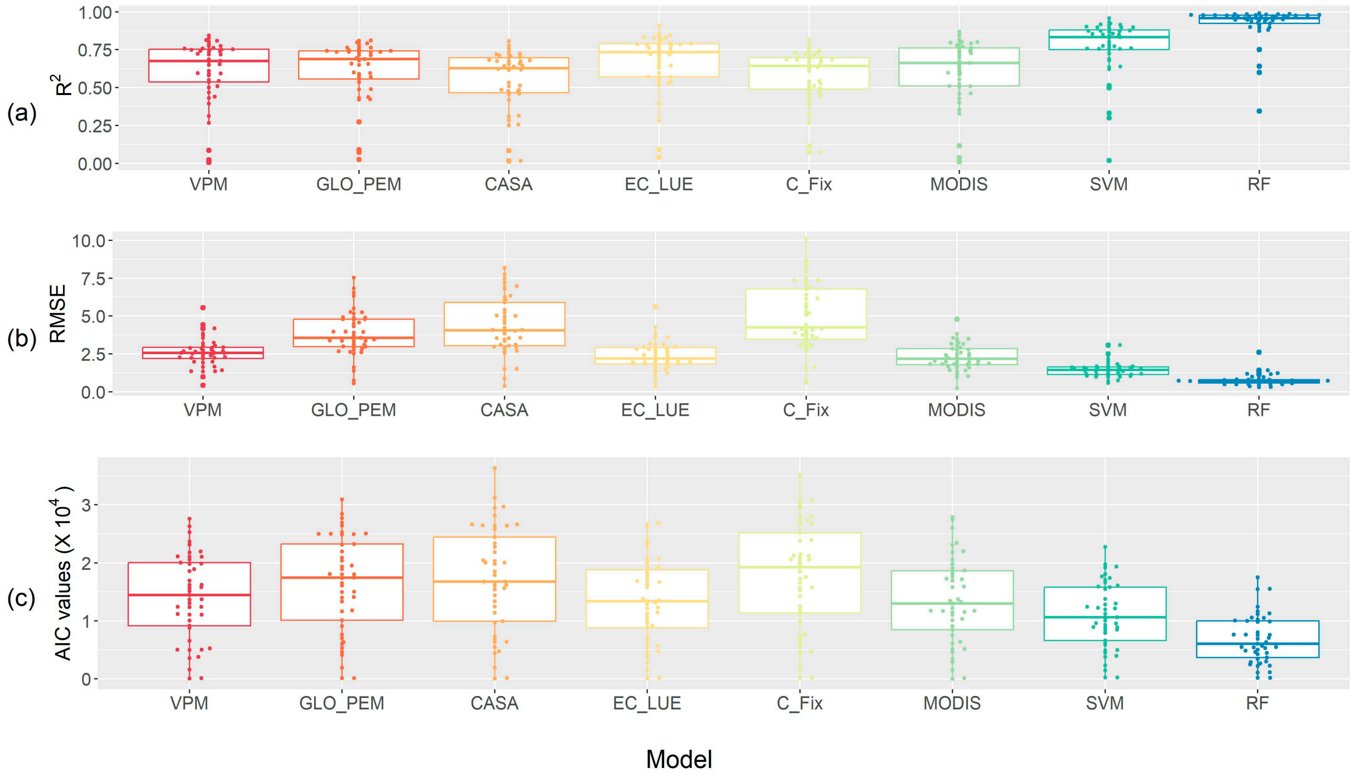

Figure 9. Boxplot of performance of R^2 (**a**), *RMSE* (**b**), and *AIC* (**c**) of LUE and machine learning methods across 45 FLUXNET2015 sites.

3.3. Performance of Individual LUE Models

The selected LUE models exhibited significant differences in the simulation of daily GPP at forest-type and flux-site scales (Figures 3 and 4, Table S3 in the Supporting Information File). Considering R^2, *RMSE*, and *RPE*, EC-LUE outperformed other models as Figure 4 and Table S3 show. Carbon fixation and evapotranspiration are two closely related processes occurring simultaneously in the same place, the stomata of leaves, during the growth of a forest. On the one hand, the sensible heat and latent heat were taken as parameters of water stress function in the EC-LUE model, which caught the key aspect in photosynthesis and evapotranspiration. Therefore, the evaporative fraction, EF, as shown in Table 1, could serve as a good representative index of soil surface moisture that was the main water limitation for forest growth [19]. On the other hand, latent heat and sensible heat were taken from FLUXNET2015 products, which means EC-LUE had the best spatial and temporal matching of input water stress parameters, while other LUE models' water stress parameters were derived from ground energy and CO_2 exchange measurements or remote sensing observations. Therefore, EC-LUE was the best model compared with other LUE models in extracting the key information in the carbon sequestration process.

As a popular LUE model, VPM gave quite good performance, especially in mixed forest terrain types. Leaf phenology and chlorophyll content in the growth period play vital roles in light harvesting reactions during forest photosynthesis [56], so the leaf phenology factor in VPM's environmental constraints could help to capture seasonal variance in photosynthetic processes, bringing an advantage in GPP simulation over other LUE models. As for the water stress function, the soil water index (LSWI) that is computed by the combination of NIR and SWIR bands from remote sensing data, is sensitive to soil moisture, which could represent the water limitation in forest growth. The ease of obtaining remote sensing data enables VPM to simulate GPP at large temporal and spatial scales across different terrestrial ecosystems [34].

GLO-PEM and CASA were proposed in the 1990s and applied to model net or gross primary production using the production efficiency approach at global or regional scales [8,12]. Although GLO-PEM and CASA yielded slightly lower accuracy compared to EC-LUE and VPM, they still performed quite well in forest GPP simulation. C-Fix was proposed to estimate net biomass in Europe using ground-site measurements and remote sensing data and has seldom been applied in other areas [15,41]. C-Fix is the only one that accounts for the effect of CO_2 fertilization in the photosynthesis process; however, the water stress is ignored, which may be the reason for a relatively poor performance compared with other LUE models in this study.

MODIS's GPP is a composite product on an 8-day timescale with a spatial resolution of 500 m based on the definition of radiation use efficiency, providing long time series data from 2000 to present on a global scale [13,18]. As a dataset used worldwide, MODIS's GPP product offered good performance in the validation using FLUXNET2015's GPP across four forest types at 45 ground sites as Figures 3, 4 and 6 show.

3.4. Performance of Machine Learning Methods

The selected five LUE models together with Köppen–Geiger climate zones were fused by the SVM and RF. This process greatly increased the accuracy of daily GPP estimation in four forest types as shown in Figures 3–6, and Figures S1–S3 in the Supporting Information File. Considering R^2 as the metric, the SVM and RF increased the accuracy of the GPP estimation by 27% and 28% compared to EC-LUE and decreased the *RMSE* (gC m^{-2} day^{-1}) from 2.77 (in EC-LUE) to 1.75 and 1.72. Both SVM and RF explained more than 90% of the GPP variability for DBF, 71% for EBF, 80% for ENF, and 77% for MF, which were significantly more than individual LUE models. Previous studies indicated that the SVM and RF outperformed other techniques and could effectively simulate land surface variables such as evapotranspiration and GPP, and our results were consistent with previous research findings [10,33,57,58].

The LUE model assumes that the ecosystem GPP is dependent on the absorbed PAR and the actual LUE that is downregulated by environmental constraints such as temperature stress and water limitation. The inputs of LUE models in this study, which were derived from the FLUXNET2015 dataset or satellite observations, could reflect the seasonal variation in environmental factors, so the LUE models could capture the seasonal cycle of forests and then contribute to the performance of the SVM and RF in GPP estimation. Previous studies showed that it was more difficult to simulate GPP in evergreen broadleaf forests because of slight fluctuations in seasonal canopy phenology and environmental conditions [17,59,60], and our results drew similar conclusions, as shown in Figure 3 and Figure S1. Although LUE models performed worse in the EBF, the SVM and RF could still significantly increase R^2 (from 0.55 to 0.71 and 0.72) and decrease *RMSE* (from 3.09 to 1.94 and 1.91), which proved that machine learning could be a feasible way to enhance the ability of simulating GPP for evergreen broadleaf forests.

LUE models are simplified equations describing how vegetation synthesizes organic matter through photosynthesis, which inevitably introduces errors for overlooking the details in the photosynthetic process. Furthermore, the uncertainties in ground measurement that drive LUE models also introduce interferences in GPP estimation. Due to the different input parameters and model structure in calculating temperature and water constraints, LUE models exhibited quite a variability in GPP simulation. Compared to LUE models, machine learning could achieve higher accuracy with the premise of providing sufficient high-quality training data. With sufficient training data and parameterization, machine learning models can automatically learn the implicit relationships between complex physical phenomena and problems without theoretical logic embedding [32,61,62]. As nonlinear models, machine learning could provide stronger regression capability than normal linear models, which can explain the outstanding performance in GPP simulations. Based on LUE modeling, machine learning models set up the relationships between inputs and ground-measurement-derived GPP and outperformed all individual LUE models. Therefore, as

Figures 3–5 show, compared to the single model, the SVM and RF vastly improved GPP estimation accuracy.

3.5. Uncertainties in Modeling Forest GPP

Our results show that individual LUE models' performance varied drastically in GPP simulation across forest types at eddy covariance sites. Considering the averages of R^2 and *RMSE*, the VPM, EC-LUE, and MODIS products outperformed GLO-PEM, CASA, and C-Fix, while machine learning significantly improved the accuracy of estimation. The uncertainties in modeling forest GPP may come from ground-site and satellite observations, which were the model input parameters, and the structure of integrating temperature and water stress in different LUE models.

First, the ground observations and subsequent processing procedures have uncertainties. GPP is not a variable that can be measured directly, and the uncertainties associated with deriving GPP products in FLUXNET2015 have been analyzed and evaluated previously in the literature [35,63]. How to divide the NEE (Net Ecosystem Exchange) of CO_2 into GPP and plants' respiration and the data series' gap-filling methods are the major uncertainty source. There are two popular ways to calculate GPP from NEE measurements: daytime and nighttime fluxes method [35]. The daytime fluxes method uses a light response curve to simulate the daytime NEE, including the consideration of the temperature impact on respiration and the VPD constraint of vegetation photosynthesis [64]. The nighttime method extrapolates night-time plants respiration into the daytime by using a temperature response function that comes from long time observations [65]. Daytime and nighttime methods are popular in separating the NEE into respiration and GPP; nevertheless, both of them have advantages and disadvantages, so we used the average of FLUXNET2015's nighttime and daytime GPP products as the referenced value to validate the accuracy of our model estimation [34]. Gap-filling methods also introduce uncertainties in the calculation of GPP. The occasional instrumental failures, adverse weather conditions, and low turbulence data induced by advection issues cause missing data and require gap-filling [63,66,67]. The friction velocity (USTAR) indicates the turbulence strength, and a USTAR threshold was applied below which the eddy covariance NEE data were considered as bad data to be filled with predictions by a regression method [65]. These GPP derivation and gap-filling methods may introduce extra uncertainties in FLUXNET2015's GPP products. Another issue was the enclosure of the energy budget at eddy covariance sites because of wind variation, footprint size, and spatial heterogeneity. The imbalance in energy budget may induce an error of about 5% to 20%, although all FLUXNET sites were requested to obey a uniform data quality control process [35,68,69]. These uncertainties in eddy covariance measurements, meteorological observations, and subsequent data processing were introduced to individual LUE models and further brought into the fusion methods, which naturally impacted the accuracy of simulations.

Second, the uncertainties in satellite observations contributed to the accuracy of LUE models that took remote sensing data as inputs. The errors in MODIS products used in this study have been reported [18,36,70] and contributed to the uncertainties in estimation of LUE models. The spatial heterogeneity was another error source. The spatial resolution of MODIS products is about 500 m, while the footprint depends on the height of the flux tower, which may be tens of meters or more, and the local terrain. Although MODIS pixels that covered the ground site were chosen carefully, mismatches still existed between MODIS pixels and footprints of ground EC sites.

Third, the differences in LUE model structure can cause significantly different performance in GPP simulations [59,71]. Temperature and water were the main constraints for most LUE models, but VPM included leaf phenology, and C-Fix considered CO_2 fertilization effects. The function of water constraint has been a long-term challenge, and many functions of water stress have been defined considering VPD, soil moisture, evapotranspiration, and water index from remote sensing [13,19,39,72]. Also, the methods used to combine water and temperature stresses were different. EC-LUE used the minimum rule, while other LUE models adopted a multiplication of environmental factors, which also contributed to the variance in GPP simulations from different LUE models.

Overall, the errors mentioned above from ground measurements and satellite observations, data processing and derivation, and the structure of LUE models (i.e., the definition of environmental stresses and integration methods) could bring uncertainties to LUE models and further contribute to the uncertainties in SVM and RF estimation.

3.6. Limitations and Future Work

The GPP modeling and fusion were implemented at the selected sites across four forest types without considering the impact of climate footprints which may not match exactly with corresponding MODIS pixels at specific sites. The land surface conditions and height of the eddy covariance towers greatly affect the GPP footprints [73,74]; therefore, the proposed machine learning methods in this study may need more validation by considering the influence of site footprints before being applied at larger scales. Although we used the special software package supplied by ORNL DAAC to identify MODIS pixels based on the geographical coordinates of each FLUXNET2015 sites, the specific site may not be located at the center of a remote sensing pixel; meanwhile, errors may also be introduced in the spatial heterogeneity which may result in mixed pixels in MODIS products of 500 m resolution.

Key factors in LUE models are ε_0, Tmin, Tmax, and Topt in Table 2. In this study, we adopted a specific look-up table to provide uniform parameters for all LUE models, and we could conduct more research on LUE model parameters optimization to leverage the advantages of each model in the future.

Another limitation is the insufficient ground observation stations for training data. Considering the data availability in long time series, the selected sites were mostly distributed in Europe and North America, as shown in Figure 1, which may make the proposed methods not applicable to each continent. Future research should pay more attention to other datasets which include eddy covariance sites across South America, Australia, Asia, and Africa.

The selection and combination of input models using machine learning contribute to the performance of model fusion [75]. Here, we chose five popular LUE models as the inputs of machine learning models; however, there are many other models to simulate GPP at different spatial scales [7]. How to select and combine these models to obtain reliable and accurate GPP estimation needs to be further studied in the future.

4. Conclusions

Based on meteorological and eddy covariance flux measurements at 45 FLUXNET2015 forest sites and remote sensing data, we ran LUE GPP models (i.e., VPM, EC-LUE, GLO-PEM, CASA, and C-Fix) to estimate forest GPP and then fused these individual models based on Köppen–Geiger climate zones to improve the ability of simulating GPP. Our results showed that EC-LUE outperformed other LUE models, giving the highest R^2 between simulated GPP and site-derived GPP on a daily scale. The performance of VPM and GLO-PEM were quite good too, while CASA and C-Fix showed extra uncertainties in simulation accuracy. The five-fold cross-validation showed that the SVM and RF could greatly improve estimation accuracy by 27% and 28%, respectively. A further analysis at each FLUXNET2015 site indicated that the RF could capture more magnitudes of GPP variations with a lower *RMSE*, so the RF was the best fusion method for simulating forest GPP. Despite the possible errors from ground measurements, satellite observations, and LUE model structure, machine learning, especially RFs, could be a good alternative option to enhance the ability of GPP simulation beyond that of individual LUE models at site scale.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/f15091615/s1, Figure S1: The scatter plots of the coefficient of determination (R^2), root-mean-square error ($RMSE$, gC m^{-2}day^{-1}), and relative predictive error (RPE) across the evergreen broadleaf forest (EBF) biome between daily site-derived GPP (gC m^{-2}day^{-1}) at flux sites and the estimates from five individual models, MODIS's GPP product, and machine learning methods (SVM and RF); Figure S2: The scatter plots of the coefficient of determination (R^2), root-mean-square error ($RMSE$, gC m^{-2}day^{-1}), and relative predictive error (RPE) across the evergreen needleleaf forest (ENF) biome between daily site-derived GPP (gC m^{-2}day^{-1}) at flux sites and the estimates from five individual models, MODIS GPP's product, and machine learning methods (SVM and RF); Figure S3: The scatter plots of the coefficient of determination (R^2), root-mean-square error ($RMSE$, gC m^{-2}day^{-1}), and relative predictive error (RPE) across the mixed forest (MF) biome between daily site-derived GPP (gC m^{-2}day^{-1}) at flux sites and the estimates from five individual models, MODIS's GPP product, and machine learning methods (SVM and RF); Table S1: The information of the 45 selected FLUXNET2015 forest sites; Table S2: The Köppen–Geiger climate classes; Table S3: The summary of the statistics (R^2, $RMSE$ (gC m^{-2} day^{-1}), and RPE) between modeled and FLUXNET2015's GPP across the 45 forest sites; Table S4: The optimum parameters of the RF and SVM across the DBF, EBF, ENF, and MF.

Author Contributions: Conceptualization, Y.F. and Z.T.; methodology, Z.T.; software, Z.T. and Y.F.; validation, C.Y. and Q.Z.; formal analysis, Z.T.; investigation, Z.T.; resources, Z.T.; data curation, C.Y.; writing—original draft preparation, Z.T. and Y.F.; writing—review and editing, C.Y., T.Z., E.K. and N.Y.K.; visualization, Y.F. and Z.T.; supervision, Y.F.; project administration, Z.T.; funding acquisition, Z.T. and Y.F. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Bureau of Statistics of China (grant number 2023LY063), China University of Labor Relations (grant number 24XYJS020), and the Disciplinary Funding of Beijing Technology and Business University (grant number STKY202311).

Data Availability Statement: The data used in this study were downloaded from public datasets: FLUXNET2015 website (https://fluxnet.org/data/fluxnet2015-dataset/, accessed on 20 February 2024) and MODIS website (https://modis.ornl.gov/, accessed on 25 February 2024).

Acknowledgments: We would like to thank Oak Ridge National Laboratory's Distributed Active Archive Center (ORNL DAAC) for providing the MODIS products. We also thank the regional flux networks of FLUXNET for obtaining the eddy covariance data. We acknowledge the financial and technical support to the eddy covariance data harmonization from Lawrence Berkeley National Laboratory, Max Planck Institute of Biogeochemistry, Universities (Tuscia, Virginia, California-Berkeley), Oak Ridge National Laboratory Distributed Active Archive Center, US National Science Foundation, Microsoft Research, US Department of Agriculture (Forest Service), and US Department of Energy.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Friedlingstein, P.; O'Sullivan, M.; Jones, M.W.; Andrew, R.M.; Bakker, D.C.E.; Hauck, J.; Landschützer, P.; Le Quéré, C.; Luijkx, I.T.; Peters, G.P.; et al. Global Carbon Budget 2023. *Earth Syst. Sci. Data* **2023**, *15*, 5301–5369. [CrossRef]
2. Pan, Y.; Birdsey, R.A.; Fang, J.; Houghton, R.; Kauppi, P.E.; Kurz, W.A.; Phillips, O.L.; Shvidenko, A.; Lewis, S.L.; Canadell, J.G.; et al. A Large and Persistent Carbon Sink in the World's Forests. *Science* **2011**, *333*, 988–993. [CrossRef] [PubMed]
3. Pugh, T.A.M.; Lindeskog, M.; Smith, B.; Poulter, B.; Arneth, A.; Haverd, V.; Calle, L. Role of Forest Regrowth in Global Carbon Sink Dynamics. *Proc. Natl. Acad. Sci. USA* **2019**, *116*, 4382–4387. [CrossRef] [PubMed]
4. Harris, N.L.; Gibbs, D.A.; Baccini, A.; Birdsey, R.A.; De Bruin, S.; Farina, M.; Fatoyinbo, L.; Hansen, M.C.; Herold, M.; Houghton, R.A.; et al. Global Maps of Twenty-First Century Forest Carbon Fluxes. *Nat. Clim. Chang.* **2021**, *11*, 234–240. [CrossRef]
5. Landsberg, J.J.; Waring, R.H. A Generalised Model of Forest Productivity Using Simplified Concepts of Radiation-Use Efficiency, Carbon Balance and Partitioning. *For. Ecol. Manag.* **1997**, *95*, 209–228. [CrossRef]
6. Monteith, J.L. Solar Radiation and Productivity in Tropical Ecosystems. *J. Appl. Ecol.* **1972**, *9*, 747. [CrossRef]
7. Pei, Y.; Dong, J.; Zhang, Y.; Yuan, W.; Doughty, R.; Yang, J.; Zhou, D.; Zhang, L.; Xiao, X. Evolution of Light Use Efficiency Models: Improvement, Uncertainties, and Implications. *Agric. For. Meteorol.* **2022**, *317*, 108905. [CrossRef]
8. Potter, C.S.; Randerson, J.T.; Field, C.B.; Matson, P.A.; Vitousek, P.M.; Mooney, H.A.; Klooster, S.A. Terrestrial Ecosystem Production: A Process Model Based on Global Satellite and Surface Data. *Glob. Biogeochem. Cycles* **1993**, *7*, 811–841. [CrossRef]
9. Running, S.; Nemani, R.R.; Heinsch, F.A.; Zhao, M.; Reeves, M.; Hashimoto, H. A Continuous Satellite-Derived Measure of Global Terrestrial Primary Production. *BioScience* **2004**, *54*, 547. [CrossRef]

10. Wei, S.; Yi, C.; Fang, W.; Hendrey, G. A Global Study of GPP Focusing on Light-Use Efficiency in a Random Forest Regression Model. *Ecosphere* **2017**, *8*, e01724. [CrossRef]
11. Thornton, P.E.; Rosenbloom, N.A. Ecosystem Model Spin-up: Estimating Steady State Conditions in a Coupled Terrestrial Carbon and Nitrogen Cycle Model. *Ecol. Model.* **2005**, *189*, 25–48. [CrossRef]
12. Prince, S.D.; Goward, S.N. Global Primary Production: A Remote Sensing Approach. *J. Biogeogr.* **1995**, *22*, 815–835. [CrossRef]
13. Running, S.; Thornton, P.; Nemani, R.; Glassy, J. Global Terrestrial Gross and Net Primary Productivity from the Earth Observing System. In *Methods in Ecosystem Science*; Springer: New York, NY, USA, 2000. [CrossRef]
14. Sims, D.A.; Rahman, A.F.; Cordova, V.D.; El-Masri, B.Z.; Baldocchi, D.D.; Flanagan, L.B.; Goldstein, A.H.; Hollinger, D.Y.; Misson, L.; Monson, R.K.; et al. On the Use of MODIS EVI to Assess Gross Primary Productivity of North American Ecosystems. *J. Geophys. Res.* **2006**, *111*, G04015. [CrossRef]
15. Veroustraete, F.; Sabbe, H.; Eerens, H. Estimation of Carbon Mass Fluxes over Europe Using the C-Fix Model and Euroflux Data. *Remote Sens. Environ.* **2002**, *83*, 376–399. [CrossRef]
16. Wu, C.; Niu, Z.; Gao, S. Gross Primary Production Estimation from MODIS Data with Vegetation Index and Photosynthetically Active Radiation in Maize. *J. Geophys. Res.* **2010**, *115*, D12127. [CrossRef]
17. Xiao, X.; Zhang, Q.; Saleska, S.; Hutyra, L.; De Camargo, P.; Wofsy, S.; Frolking, S.; Boles, S.; Keller, M.; Moore, B. Satellite-Based Modeling of Gross Primary Production in a Seasonally Moist Tropical Evergreen Forest. *Remote Sens. Environ.* **2005**, *94*, 105–122. [CrossRef]
18. Zhao, M.; Heinsch, F.A.; Nemani, R.R.; Running, S.W. Improvements of the MODIS Terrestrial Gross and Net Primary Production Global Data Set. *Remote Sens. Environ.* **2005**, *95*, 164–176. [CrossRef]
19. Yuan, W.; Liu, S.; Zhou, G.; Zhou, G.; Tieszen, L.L.; Baldocchi, D.; Bernhofer, C.; Gholz, H.; Goldstein, A.H.; Goulden, M.L.; et al. Deriving a Light Use Efficiency Model from Eddy Covariance Flux Data for Predicting Daily Gross Primary Production across Biomes. *Agric. For. Meteorol.* **2007**, *143*, 189–207. [CrossRef]
20. Xiao, X.; Hollinger, D.; Aber, J.; Goltz, M.; Davidson, E.A.; Zhang, Q.; Moore, B. Satellite-Based Modeling of Gross Primary Production in an Evergreen Needleleaf Forest. *Remote Sens. Environ.* **2004**, *89*, 519–534. [CrossRef]
21. Jordan, M.I.; Mitchell, T.M. Machine Learning: Trends, Perspectives, and Prospects. *Science* **2015**, *349*, 255–260. [CrossRef]
22. Arun Bhavsar, K.; Singla, J.; Al-Otaibi, Y.D.; Song, O.-Y.; Bin Zikriya, Y.; Kashif Bashir, A. Medical Diagnosis Using Machine Learning: A Statistical Review. *Comput. Mater. Contin.* **2021**, *67*, 107–125. [CrossRef]
23. Iqbal, S.; Hassan, S.-U.; Aljohani, N.R.; Alelyani, S.; Nawaz, R.; Bornmann, L. A Decade of In-Text Citation Analysis Based on Natural Language Processing and Machine Learning Techniques: An Overview of Empirical Studies. *Scientometrics* **2021**, *126*, 6551–6599. [CrossRef]
24. Lary, D.J.; Alavi, A.H.; Gandomi, A.H.; Walker, A.L. Machine Learning in Geosciences and Remote Sensing. *Geosci. Front.* **2016**, *7*, 3–10. [CrossRef]
25. Mahadevkar, S.V.; Khemani, B.; Patil, S.; Kotecha, K.; Vora, D.R.; Abraham, A.; Gabralla, L.A. A Review on Machine Learning Styles in Computer Vision—Techniques and Future Directions. *IEEE Access* **2022**, *10*, 107293–107329. [CrossRef]
26. Li, Y.; Li, C.; Li, M.; Liu, Z. Influence of Variable Selection and Forest Type on Forest Aboveground Biomass Estimation Using Machine Learning Algorithms. *Forests* **2019**, *10*, 1073. [CrossRef]
27. Li, M.; Im, J.; Beier, C. Machine Learning Approaches for Forest Classification and Change Analysis Using Multi-Temporal Landsat TM Images over Huntington Wildlife Forest. *GISci. Remote Sens.* **2013**, *50*, 361–384. [CrossRef]
28. Tamm, T.; Remm, K. Estimating the Parameters of Forest Inventory Using Machine Learning and the Reduction of Remote Sensing Features. *Int. J. Appl. Earth Obs. Geoinf.* **2009**, *11*, 290–297. [CrossRef]
29. Chen, Y.; Yuan, W.; Xia, J.; Fisher, J.B.; Dong, W.; Zhang, X.; Liang, S.; Ye, A.; Cai, W.; Feng, J. Using Bayesian Model Averaging to Estimate Terrestrial Evapotranspiration in China. *J. Hydrol.* **2015**, *528*, 537–549. [CrossRef]
30. Duan, Q.; Phillips, T.J. Bayesian Estimation of Local Signal and Noise in Multimodel Simulations of Climate Change. *J. Geophys. Res.* **2010**, *115*, D18123. [CrossRef]
31. Wu, H.; Zhang, X.; Liang, S.; Yang, H.; Zhou, G. Estimation of Clear-Sky Land Surface Longwave Radiation from MODIS Data Products by Merging Multiple Models. *J. Geophys. Res.* **2012**, *117*, D22107. [CrossRef]
32. Yang, L.; Jia, K.; Liang, S.; Liu, J.; Wang, X. Comparison of Four Machine Learning Methods for Generating the GLASS Fractional Vegetation Cover Product from MODIS Data. *Remote Sens.* **2016**, *8*, 682. [CrossRef]
33. Yao, Y.; Liang, S.; Li, X.; Chen, J.; Liu, S.; Jia, K.; Zhang, X.; Xiao, Z.; Fisher, J.B.; Mu, Q.; et al. Improving Global Terrestrial Evapotranspiration Estimation Using Support Vector Machine by Integrating Three Process-Based Algorithms. *Agric. For. Meteorol.* **2017**, *242*, 55–74. [CrossRef]
34. Zhang, Y.; Xiao, X.; Wu, X.; Zhou, S.; Zhang, G.; Qin, Y.; Dong, J. A Global Moderate Resolution Dataset of Gross Primary Production of Vegetation for 2000–2016. *Sci. Data* **2017**, *4*, 170165. [CrossRef]
35. Pastorello, G.; Trotta, C.; Canfora, E.; Chu, H.; Christianson, D.; Cheah, Y.-W.; Poindexter, C.; Chen, J.; Elbashandy, A.; Humphrey, M.; et al. The FLUXNET2015 Dataset and the ONEFlux Processing Pipeline for Eddy Covariance Data. *Sci. Data* **2020**, *7*, 225. [CrossRef]
36. Zhao, M.; Running, S.W.; Nemani, R.R. Sensitivity of Moderate Resolution Imaging Spectroradiometer (MODIS) Terrestrial Primary Production to the Accuracy of Meteorological Reanalyses. *J. Geophys. Res.* **2006**, *111*, G01002. [CrossRef]
37. Beck, H.E.; Zimmermann, N.E.; McVicar, T.R.; Vergopolan, N.; Berg, A.; Wood, E.F. Present and Future Köppen-Geiger Climate Classification Maps at 1-Km Resolution. *Sci. Data* **2018**, *5*, 180214. [CrossRef]

38. Priestley, C.H.B.; Taylor, R.J. On the Assessment of Surface Heat Flux and Evaporation Using Large-Scale Parameters. *Mon. Wea. Rev.* **1972**, *100*, 81–92. [CrossRef]
39. Cao, M.; Prince, S.D.; Small, J.; Goetz, S.J. Remotely Sensed Interannual Variations and Trends in Terrestrial Net Primary Productivity 1981–2000. *Ecosystems* **2004**, *7*, 233–242. [CrossRef]
40. Goetz, S.J.; Prince, S.D.; Small, J.; Gleason, A.C.R. Interannual Variability of Global Terrestrial Primary Production: Results of a Model Driven with Satellite Observations. *J. Geophys. Res.* **2000**, *105*, 20077–20091. [CrossRef]
41. Veroustraete, F.; Patyn, J.; Myneni, R.B. Estimating Net Ecosystem Exchange of Carbon Using the Normalized Difference Vegetation Index and an Ecosystem Model. *Remote Sens. Environ.* **1996**, *58*, 115–130. [CrossRef]
42. Lewis, J.M. The Story behind the Bowen Ratio. *Bull. Am. Meteor. Soc.* **1995**, *76*, 2433–2443. [CrossRef]
43. Li, X.; Liang, S.; Yu, G.; Yuan, W.; Cheng, X.; Xia, J.; Zhao, T.; Feng, J.; Ma, Z.; Ma, M.; et al. Estimation of Gross Primary Production over the Terrestrial Ecosystems in China. *Ecol. Model.* **2013**, *261–262*, 80–92. [CrossRef]
44. Liang, S.; Zhao, X.; Liu, S.; Yuan, W.; Cheng, X.; Xiao, Z.; Zhang, X.; Liu, Q.; Cheng, J.; Tang, H.; et al. A Long-Term Global LAnd Surface Satellite (GLASS) Data-Set for Environmental Studies. *Int. J. Digit. Earth* **2013**, *6*, 5–33. [CrossRef]
45. Yuan, W.; Zheng, Y.; Piao, S.; Ciais, P.; Lombardozzi, D.; Wang, Y.; Ryu, Y.; Chen, G.; Dong, W.; Hu, Z.; et al. Increased Atmospheric Vapor Pressure Deficit Reduces Global Vegetation Growth. *Sci. Adv.* **2019**, *5*, eaax1396. [CrossRef]
46. Suykens, J.A.K. Support Vector Machines: A Nonlinear Modelling and Control Perspective. *Eur. J. Control* **2001**, *7*, 311–327. [CrossRef]
47. Vapnik, V. Measuring the VC-Dimension of a Learning Machine. *Neural Comput.* **1994**, *6*, 851–876. [CrossRef]
48. Khalil, A.F.; McKee, M.; Kemblowski, M.; Asefa, T.; Bastidas, L. Multiobjective Analysis of Chaotic Dynamic Systems with Sparse Learning Machines. *Adv. Water Resour.* **2006**, *29*, 72–88. [CrossRef]
49. Cortes, C.; Vapnik, V. Support-Vector Networks. *Mach. Learn.* **1995**, *20*, 273–297. [CrossRef]
50. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
51. Amini, S.; Saber, M.; Rabiei-Dastjerdi, H.; Homayouni, S. Urban Land Use and Land Cover Change Analysis Using Random Forest Classification of Landsat Time Series. *Remote Sens.* **2022**, *14*, 2654. [CrossRef]
52. Gyamerah, S. Probabilistic Forecasting of Crop Yields via Quantile Random Forest and Epanechnikov Kernel Function. *Agric. For. Meteorol.* **2020**, *280*, 107808. [CrossRef]
53. Taylor, K.E. Summarizing Multiple Aspects of Model Performance in a Single Diagram. *J. Geophys. Res.* **2001**, *106*, 7183–7192. [CrossRef]
54. Akaike, H. A New Look at the Statistical Model Identification. *IEEE Trans. Automat. Contr.* **1974**, *19*, 716–723. [CrossRef]
55. Schwarz, G. Estimating the Dimension of a Model. *Ann. Stat.* **1978**, *6*, 461–464. [CrossRef]
56. Croft, H.; Chen, J.M.; Froelich, N.J.; Chen, B.; Staebler, R.M. Seasonal Controls of Canopy Chlorophyll Content on Forest Carbon Uptake: Implications for GPP Modeling. *J. Geophys. Res. Biogeosci.* **2015**, *120*, 1576–1586. [CrossRef]
57. Tejada, A.T.; Ella, V.B.; Lampayan, R.M.; Reaño, C.E. Modeling Reference Crop Evapotranspiration Using Support Vector Machine (SVM) and Extreme Learning Machine (ELM) in Region IV-A, Philippines. *Water* **2022**, *14*, 754. [CrossRef]
58. Tian, Z.; Yi, C.; Fu, Y.; Kutter, E.; Krakauer, N.Y.; Fang, W.; Zhang, Q.; Luo, H. Fusion of Multiple Models for Improving Gross Primary Production Estimation with Eddy Covariance Data Based on Machine Learning. *JGR Biogeosci.* **2023**, *128*, e2022JG007122. [CrossRef]
59. Raczka, B.M.; Davis, K.J.; Huntzinger, D.; Neilson, R.P.; Poulter, B.; Richardson, A.D.; Xiao, J.; Baker, I.; Ciais, P.; Keenan, T.F.; et al. Evaluation of Continental Carbon Cycle Simulations with North American Flux Tower Observations. *Ecol. Monogr.* **2013**, *83*, 531–556. [CrossRef]
60. Yuan, W.; Cai, W.; Xia, J.; Chen, J.; Liu, S.; Dong, W.; Merbold, L.; Law, B.; Arain, A.; Beringer, J.; et al. Global Comparison of Light Use Efficiency Models for Simulating Terrestrial Vegetation Gross Primary Production Based on the LaThuile Database. *Agric. For. Meteorol.* **2014**, *192–193*, 108–120. [CrossRef]
61. Pande, C.B.; Egbueri, J.C.; Costache, R.; Sidek, L.M.; Wang, Q.; Alshehri, F.; Din, N.M.; Gautam, V.K.; Chandra Pal, S. Predictive Modeling of Land Surface Temperature (LST) Based on Landsat-8 Satellite Data and Machine Learning Models for Sustainable Development. *J. Clean. Prod.* **2024**, *444*, 141035. [CrossRef]
62. Sharifi, E.; Saghafian, B.; Steinacker, R. Downscaling Satellite Precipitation Estimates with Multiple Linear Regression, Artificial Neural Networks, and Spline Interpolation Techniques. *JGR Atmos.* **2019**, *124*, 789–805. [CrossRef]
63. Aubinet, M.; Vesala, T.; Papale, D. (Eds.) *Eddy Covariance: A Practical Guide to Measurement and Data Analysis*; Springer: Dordrecht, The Netherlands, 2012; ISBN 978-94-007-2350-4.
64. Lasslop, G.; Reichstein, M.; Papale, D.; Richardson, A.D.; Arneth, A.; Barr, A.; Stoy, P.; Wohlfahrt, G. Separation of Net Ecosystem Exchange into Assimilation and Respiration Using a Light Response Curve Approach: Critical Issues and Global Evaluation. *Glob. Chang. Biol.* **2010**, *16*, 187–208. [CrossRef]
65. Reichstein, M.; Falge, E.; Baldocchi, D.; Papale, D.; Aubinet, M.; Berbigier, P.; Bernhofer, C.; Buchmann, N.; Gilmanov, T.; Granier, A.; et al. On the Separation of Net Ecosystem Exchange into Assimilation and Ecosystem Respiration: Review and Improved Algorithm. *Glob. Chang. Biol.* **2005**, *11*, 1424–1439. [CrossRef]
66. Goulden, M.L.; Munger, J.W.; Fan, S.-M.; Daube, B.C.; Wofsy, S.C. Measurements of Carbon Sequestration by Long-Term Eddy Covariance: Methods and a Critical Evaluation of Accuracy. *Glob. Chang. Biol.* **1996**, *2*, 169–182. [CrossRef]
67. Yi, C.; Monson, R.K.; Zhai, Z.; Anderson, D.E.; Lamb, B.; Allwine, G.; Turnipseed, A.A.; Burns, S.P. Modeling and Measuring the Nocturnal Drainage Flow in a High-Elevation, Subalpine Forest with Complex Terrain. *J. Geophys. Res.* **2005**, *110*, D22303. [CrossRef]
68. Foken, T. The Energy Balance Closure Problem: An Overview. *Ecol. Appl.* **2008**, *18*, 1351–1367. [CrossRef]

69. Stoy, P.C.; Mauder, M.; Foken, T.; Marcolla, B.; Boegh, E.; Ibrom, A.; Arain, M.A.; Arneth, A.; Aurela, M.; Bernhofer, C.; et al. A Data-Driven Analysis of Energy Balance Closure across FLUXNET Research Sites: The Role of Landscape Scale Heterogeneity. *Agric. For. Meteorol.* **2013**, *171–172*, 137–152. [CrossRef]
70. Serbin, S.P.; Ahl, D.E.; Gower, S.T. Spatial and Temporal Validation of the MODIS LAI and FPAR Products across a Boreal Forest Wildfire Chronosequence. *Remote Sens. Environ.* **2013**, *133*, 71–84. [CrossRef]
71. Zhang, L.; Zhou, D.; Fan, J.; Hu, Z. Comparison of Four Light Use Efficiency Models for Estimating Terrestrial Gross Primary Production. *Ecol. Model.* **2015**, *300*, 30–39. [CrossRef]
72. Xiao, X. Modeling Gross Primary Production of Temperate Deciduous Broadleaf Forest Using Satellite Images and Climate Data. *Remote Sens. Environ.* **2004**, *91*, 256–270. [CrossRef]
73. Schmid, H.P. Footprint Modeling for Vegetation Atmosphere Exchange Studies: A Review and Perspective. *Agric. For. Meteorol.* **2002**, *113*, 159–183. [CrossRef]
74. Yu, T.; Zhang, Q.; Sun, R. Comparison of Machine Learning Methods to Up-Scale Gross Primary Production. *Remote Sens.* **2021**, *13*, 2448. [CrossRef]
75. Tian, Z.; Fu, Y.; Liu, S. Remote Sensing Image Classification Based on Heterogeneous Machine Learning Algorithm Fusion. *Comput. Sci.* **2019**, *46*, 235–240.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.