Article

Exploring the Best-Matching Plant Traits and Environmental Factors for Vegetation Indices in Estimates of Global Gross Primary Productivity

Weiqing Zhao [1,2] and Zaichun Zhu [1,2,*]

[1] School of Urban Planning and Design, Shenzhen Graduate School, Peking University, Shenzhen 518055, China
[2] Key Laboratory of Earth Surface System and Human-Earth Relations, Ministry of Natural Resources of China, Shenzhen Graduate School, Peking University, Shenzhen 518055, China
* Correspondence: zhu.zaichun@pku.edu.cn

Abstract: As the largest source of uncertainty in carbon cycle studies, accurate quantification of gross primary productivity (GPP) is critical for the global carbon budget in the context of global climate change. Numerous vegetation indices (VIs) based on satellite data have participated in the construction of GPP models. However, the relative performance of various VIs in predicting GPP and what additional factors should be combined with them to reveal the photosynthetic capacity of vegetation mechanistically better are still poorly understood. We constructed two types of models (universal and plant functional type [PFT]-specific) for solar-induced chlorophyll fluorescence (SIF), near-infrared reflectance of vegetation (NIR$_v$), and Leaf Area Index (LAI) based on two widely used machine learning algorithms, i.e., the random forest (RF) and back propagation neural network (BPNN) algorithms. A total of thirty plant traits and environmental factors with legacy effects are considered in the model. We then systematically investigated the ancillary variables that best match each vegetation index in estimating global GPP. Four types of models (universal and PFT-specific, RF and BPNN) consistently show that SIF performs best when modeled using a single vegetation index ($R^2 = 0.67$, RMSE = 2.24 g C·m^{-2}·d^{-1}); however, NIR$_v$ combined with CO$_2$, plant traits, and climatic factors can achieve the highest prediction accuracy ($R^2 = 0.87$, RMSE = 1.40 g C·m^{-2}·d^{-1}). Plant traits effectively enhance all prediction models' accuracy, and climatic variables are essential factors in improving the accuracy of NIRv- or LAI-based GPP models, but not the accuracy of SIF-based models. Our findings provide valuable information for the configuration of the data-driven models to improve the accuracy of predicting GPP and provide insights into the physiological and ecological mechanisms underpinning GPP prediction.

Keywords: gross primary productivity; solar-induced chlorophyll fluorescence; leaf area index; NIRv; machine learning

1. Introduction

With the massive fossil fuel burning and land use/land cover changes due to human activities since the industrial revolution, an imbalance between carbon sources and sinks has been discovered in the investigation of the global carbon budget [1]. Gross primary productivity (GPP) is primarily used to describe the photosynthetic capability of terrestrial ecosystems and is the largest source of uncertainty in carbon cycle studies [2]. Improving the accuracy of GPP estimates is inextricably linked to gaining a better understanding of the relationship between ecosystems and climate change, as well as laying the foundation for understanding global carbon cycle processes and ecosystem functions [3–5].

GPP cannot be measured directly at the ecosystem scale, and model simulations, including process-based models, light-use efficiency (LUE) models, and data-driven models, are seen as an effective way to overcome this challenge [6]. Process-based models are developed with rigorous plant physiology and ecological principles coupled with the

dynamic processes of ecosystems and can accurately simulate vegetation photosynthetic capacity [7,8]. However, the method is limited by the complexity of the required input parameters and parameterization over large-scale regions [9,10]. Along with the generation of eddy covariance (EC) techniques and the advancement of remote sensing technology, LUE models based on satellite data and flux data are widely popular in ecological studies [11–14]. However, such models typically do not account for the variation in LUE caused by radiation intensity and assume that maximum LUE is only related to plant functional types (PFTs), introducing great uncertainty into GPP simulations [15,16]. Another approach to estimating GPP is data-driven, generally by establishing relationships between local vegetation parameters, climate factors, and flux tower GPP and then upscaling to regional and global scales [17,18]. Data-driven models based on machine learning are widely used in large-scale studies for predicting vegetation productivity due to their higher accuracy than traditional regression methods [19,20]. The most widely known MTE GPP model [21] often serves as a reference to validate and evaluate the simulation of other models [22–24].

Remote sensing vegetation indices (VIs), allowing for the effective detection of changes in large-scale vegetation growth, have been widely used to construct GPP models. Compared to simple VIs (e.g., the Normalized Difference Vegetation Index [NDVI] [25] and Enhanced Vegetation Index [EVI] [26]), Leaf Area Index (LAI) quantifies the structure and growth of vegetation and the Fraction of Photosynthetically Active Radiation absorbed by vegetation (FPAR) reflects the potential energy utilized by the canopy in photosynthesis, both of which are critical biophysical parameters in carbon cycle research [27–29]. They are solely driven by leaf development, are more closely related to actual ground observations, and are more sensitive to high-density biomes than simple VIs when observing vegetation phenology [30,31]. The near-infrared reflectance of vegetation (NIR_v), a recently proposed novel vegetation index, has been shown to correlate well with GPP [32]. It has the advantage of overcoming the sensitivity of the NDVI to the vegetation fraction and can refine vegetation's contribution to NIR reflectance no matter how sparse the canopies are and how bright the soil background is. Nevertheless, the common deficiency of the above VIs constructed based on the spectral characteristics of vegetation is that they are insensitive to photosynthetic processes that have not yet caused reflectance changes and do not correlate well with the photosynthetic activities of plants in the short term, making it difficult for them to accurately and timely reflect the dynamic changes in vegetation photosynthesis [33]. Several studies have proven that these indices perform poorly in ecosystems without high variability in greenness [34,35]. The absorption of sunlight by vegetation for photosynthesis is accompanied by the re-emission of red and near-infrared photons, called solar-induced chlorophyll fluorescence (SIF). SIF, which is closely related to the two processes of nonphotochemical quenching and photochemical quenching, can capture transient photosynthesis in vegetation even if there is no change in greenness or structure. One study surprisingly discovered a strong spatiotemporal correlation between satellite-derived SIF and flux tower-derived GPP estimates [36]. This finding provides new insight for estimating ecosystem-scale GPP and encourages us to utilize SIF in carbon cycle research as a reliable proxy for GPP. However, as SIF products have only been available since 2000, the reasons for the variations in the relationship between SIF and GPP over time and space need to be clarified [37–39].

Accumulating evidence suggests that ecosystem-scale GPP is influenced by a combination of vegetation's biophysical properties [37,38,40–43] and environmental factors [12,44,45]. Data-driven models based on machine learning are good at handling the interaction of individual factors and can achieve excellent prediction accuracy. It has been shown that supplementing the vegetation index with other information about vegetation can improve the prediction results of GPP models [35,46]. On the other hand, the black-box nature of machine learning results in model performance that depends heavily on how the explanatory variables are combined. The relative performance of various VIs in predicting GPP and what additional factors should be combined with them to mechanistically better reveal the photosynthetic capacity of vegetation are still poorly understood.

In this study, we used a remote sensing time series, including the GOSIF, MODIS NIR$_v$, and MODIS LAI, and GPP observations from 197 flux towers to examine their ability to quantify the spatiotemporal variation in GPP. The objectives of this study are to (1) evaluate the performance of models in GPP estimation based on two widely used machine learning models, i.e., the random forest (RF) and back propagation neural network (BPNN) models; (2) generate several sets of global gridded GPP products for 2003–2018 based on the optimal machine learning models and compare them to previous products; and (3) systematically investigate the environmental factors and plant traits that best match each vegetation index in estimating global GPP.

2. Materials and Methods

2.1. Vegetation Indices

We used three vegetation indices (SIF, NIR$_v$, and LAI) to investigate their performance in estimating GPP when used in combination with machine learning algorithms. We used the GOSIF product developed by Li and Xiao [47], which provided global monthly SIF observations with $0.05° \times 0.05°$ spatial resolution for the 2001 to 2020 period. This product extended OCO-2 raw data to a longer period and global coverage through the Cubist regression tree model. The NIR$_v$ data were generated based on the BRDF-adjusted reflectance data (MODIS MCD43C4, Collection 6) according to the methodology described in [32]. The NIR$_v$ data we generated have global coverage from 2001 to 2018, with monthly temporal resolution and $0.05°$ spatial resolution. LAI data were derived from the Reprocessed MODIS Leaf Area Index datasets from the Land-Atmosphere Interaction Research Group at Sun Yat-Sen University, with global coverage at $0.05° \times 0.05°$ spatial resolution and monthly temporal resolution for 2003 to 2020 (http://globalchange.bnu.edu.cn/research/laiv6). The LAI data are more continuous and consistent in time and spatial domains than the original MODIS LAI product [48].

2.2. FLUXNET Data

The GPP measurements used in this study were derived from FLUXNET2015 Tier1 data (http://fluxnet.fluxdata.org, accessed on 24 December 2021), which provide a collection of EC flux data from 212 sites across multiple regional networks (Figure 1) [49]. The flux data were processed in a standardized protocol to promote consistency and intercompatibility among sites. We used monthly GPP estimated by day–time partitioning of the Net Ecosystem Exchange (NEE) with the variable USTAR threshold following [32,49]. At least 75% of valid GPP observations were required for each site–month and a minimum of 9 months for each site–year. Detailed information about each site used in this study and their PFTs can be found in Table S1.

2.3. Plant Traits

Plant traits are closely related to ecosystem functions. Leaf economic traits and hydraulic traits, which have been shown to affect photosynthetic capacity, have received much attention [50,51]. In this study, specific leaf area (SLA) and leaf nitrogen content (Nm), which are related to the carbon economy [52], and canopy height (Hc), which reflects ecosystem water use strategy, were taken into account [41]. SLA and Nm were derived from the TRY database [53] and were developed by [54]. The canopy height data were obtained from ICESat/GLAS LiDAR data (1km spatial resolution) and combined with other ancillary variables to predict uncovered areas [55].
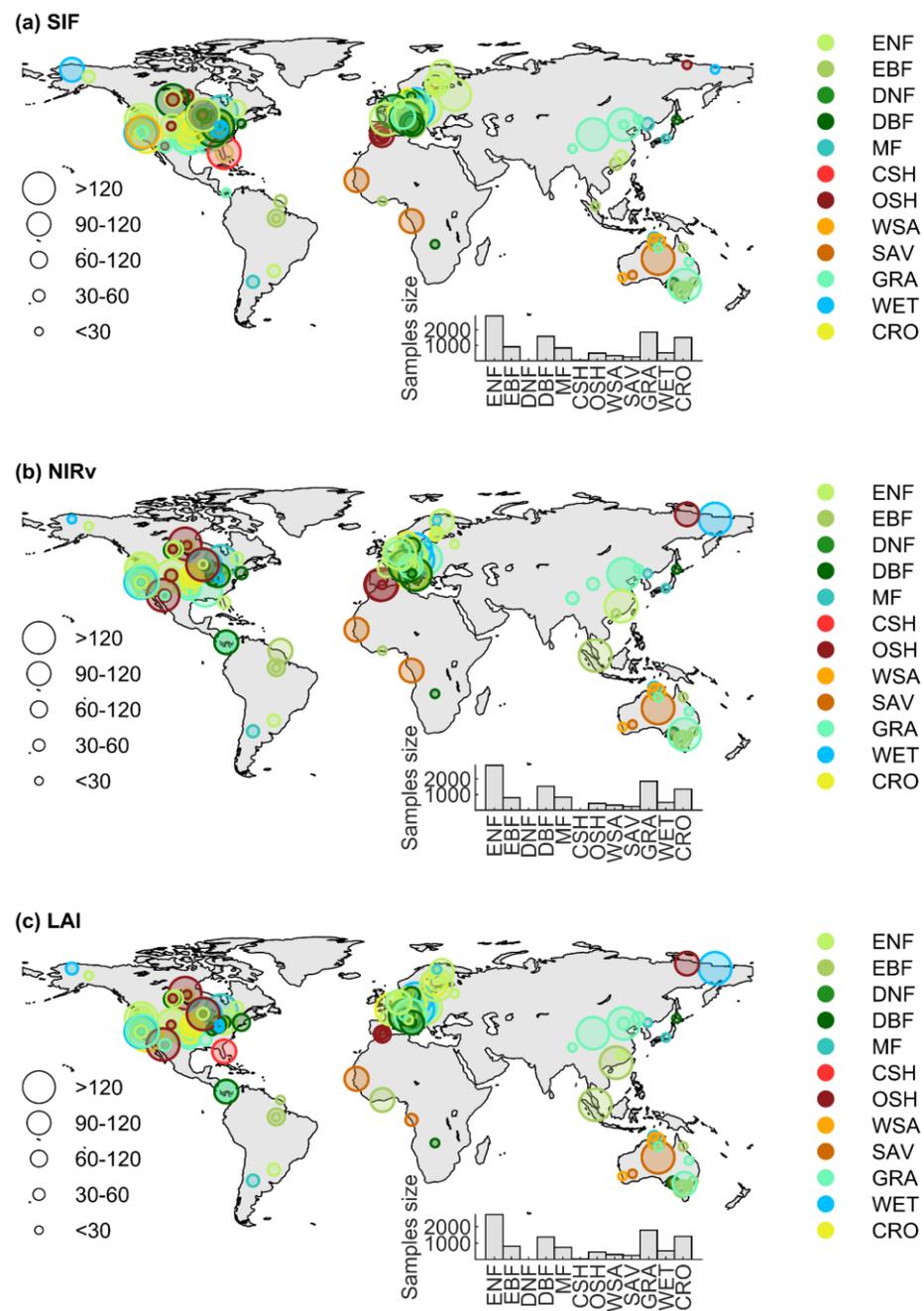
Figure 1. Map of the sites in the FLUXNET2015 dataset used for the models based on each vegetation characteristic parameter. All sites were classified into 12 PFTs according to the IGBP classification scheme. The size of the circles represents the number of samples available at the site. The histogram represents the number of samples for different PFTs. ENF: evergreen needleleaf forests, EBF: evergreen broadleaf forests, DNF: deciduous needleleaf forests, DBF: deciduous broadleaf forests, MF: mixed forests, OSH: open shrublands, WSA: woody savannas, SAV: savannas, GRA: grasslands, CRO: croplands, WET: wetlands.

2.4. Climatic Data

The TerraClimate dataset provides monthly climate measures with a high spatial resolution (1/24°) covering global land surfaces [11]. The primary climate variables used in this study are shown in Table 1, either directly from this product or calculated from the variables provided by this product. It has been demonstrated that climatic circumstances might influence vegetation growth [38,56–59], thus the climate factors for the current month,

the previous month, and the previous two months were considered as input variables for the model.

Table 1. Details of the datasets used in the data-driven models.

Category	Variable	Description	Source
Vegetation Indices	SIF	Solar-induced chlorophyll fluorescence	[47]
	NIR_v	Near-infrared reflectance of vegetation	MCD43C4
	LAI	Leaf Area Index	[48]
Plant Traits	Hc	Canopy height	[55]
	SLA	Specific leaf area	[54]
	Nm	Foliar nitrogen concentration per unit dry mass	[54]
Climatic Factors	Tmp	Air temperature	[11]
	Tmax	Maximum air temperature	[11]
	Tmin	Minimum air temperature	[11]
	DTR	Diurnal temperature range	[11]
	Prec	Precipitation	[11]
	SRAD	Downward shortwave radiation flux at the surface	[11]
	SWC	Soil water content	[11]
	VPD	Vapor Pressure Deficit	[11]
Other Factors	PFT	Plant functional type	MCD12Q1
	CO_2	Atmospheric carbon dioxide concentration	[60]

2.5. Land Cover Data

For the global-scale study, we used the land cover product from MCD12Q1 with International Geosphere-Biosphere Programme (IGBP) classes at a spatial resolution of 0.05° [61]. The time period covered by the selected data ranges from 2001 to 2020, with an image each year. The twelve major land cover classes, or PFTs, used for model construction and analysis of results in this study are shown in Figure S1. Note that we did not employ sites with the land cover class of Snow/Ice.

All data as explanatory variables were resampled into a common spatial resolution (1/24°) using the nearest neighbor algorithm.

2.6. Other GPP Products

Three GPP datasets covering the 2003–2018 period (FLUXCOM GPP and GOSIF GPP based on the data-driven method and TRENDY GPP based on the process-based models) were used to evaluate the performance of our gridded GPP products. FLUXCOM GPP (version RS+METEO) was upscaled from EC tower measurements using three machine learning algorithms with combined remote sensing and meteorological data as inputs [20]. Here, we used the average of three sets of GPP products. The monthly GOSIF-GPP dataset at 0.05° spatial resolution was generated using a robust linear relationship between tower GPP and GOSIF to estimate regional and global terrestrial photosynthesis [62]. TRENDY GPP is an ensemble of 10 state-of-the-art ecosystem models (CABLE-POP, CLM5.0, ISAM, ISBA-CTRIP, JULES, LPJ-GUESS, ORCHIDEE, ORCHIDEE-v3, SDGVM, VISIT) that participated in the TRENDY (v9) multi-model inter-comparison and followed a standard protocol [1].

2.7. Estimation of GPP Based on Machine Learning Algorithms

The Random forest (RF) algorithm is an ensemble learning approach for classification or regression that integrates many decision trees [63]. The number of decision trees, the number of features to select for each split, and the minimum number of observations per leaf are hyper-parameters that need to be modified. The RF algorithm subsamples features according to the user's settings before growing each tree, which can reduce the correlation among individual decision trees and thus improve model accuracy. Furthermore, its ability to handle high-dimensional information helps us analyze climatic and environmental factors [64].

Artificial neural networks (ANNs) are a type of machine learning algorithm inspired by the structure and function of biological neural networks [65]. A typical ANN usually consists of input, output, and hidden layers, each containing several artificial neurons. The back propagation neural network (BPNN) is one of the most popular and proven ANN algorithms being used for ecological studies [66–68]. During the model training process, signals flow from the input layer to the output layer, perhaps after passing through several hidden layers. Errors in the output layer propagate backward to the previous layers until they meet the user-defined threshold. The network attempts to minimize the discrepancies between observations and predictions.

We used the RF model to identify the factors with the most powerful explanation of GPP and to explore the importance of these predictors. We trained six RF models to select features using the 'TreeBagger' function of MATLAB 2021b, i.e., two types of models (PFT-specific and universal) for each vegetation index (SIF, NIR_v, and LAI). RF was chosen due to the robustness of ranking feature importance under different hyper-parameters and high interpretability. For each of the six models, one vegetation index as well as all environmental factors (plant functional type [PFT] and atmospheric carbon dioxide concentration [CO_2]), climatic constraints (air temperature [Tmp], maximum air temperature [Tmax], minimum air temperature [Tmin], diurnal temperature range [DTR], precipitation [Prec], downward shortwave radiation flux at the surface [SRAD], soil water content [SWC], and vapor pressure deficit [VPD] for the current month, the previous month, and the previous two months), and plant traits (canopy height [Hc], specific leaf area [SLA], and foliar nitrogen concentration per unit dry mass [Nm]) were initially selected for training (Table 1). Each model comprised 100 decision trees, was sampled without replacement, and was trained using 70% of the data. Model performance was evaluated using out-of-bag (OOB) R-squared (R^2) and root mean square error (RMSE) values. The predictor with the lowest importance score in the iteration was removed and the whole procedure was then repeated until only the vegetation index, CO_2, and PFTs were left. The predictors used to estimate GPP were identified based on the performance curve of OOB R^2 and RMSE. The determination of the model is based on the principle that further reductions in the number of predictors would considerably reduce model performance, while increasing the number of predictors would not significantly improve model performance.

The predictors for each of the six types of models were determined and the construction of the corresponding types of BPNN models was implemented with the 'Deep Learning Toolbox' in MATLAB 2021b. We divided the training dataset into training, validation, and test data with proportions of 70%, 15%, and 15%, respectively. The training set and validation set were used to tune the hyper-parameters, and the test set was used to evaluate the performance of the network. Over-fitting was defined as a loss of more than 3% in performance between the training set and validation set in this study, and an early stopping method was utilized to prevent it.

Bayesian optimization determined the optimum hyper-parameters for each machine learning method. We ran each routine 50 times to minimize the effects of random model initialization. Finally, we estimated the predicted R^2 and RMSE values using the optimal model and evaluated the performance of different models by calculating the averages of the predicted R^2 and RMSE values. All the results were produced on a PC running Windows 10 which had a 3.0 GHz CPU and 40.00GB of RAM.

We evaluated the performance of six types of models based on two machine learning algorithms in terms of GPP estimation for different vegetation types along latitudinal bands. We then investigated the factors that are the best candidates for developing GPP estimation models with different VIs. Ten combinations of VIs, CO_2, PFT, plant traits, and climate factors were tested (Table S2).

3. Results

3.1. The Performance of the Optimal VI-Based GPP Estimation Models

We found that all optimal VI-based GPP estimation models show reasonable accuracy in predicting GPP at a global scale (R^2 ranges from 0.79 to 0.87, RMSE ranges from 1.40 g C·m^{-2}·d^{-1} to 1.65 g C·m^{-2}·d^{-1}), but the NIR$_v$-based PFT-specific RF models performed the best (R^2 = 0.87, RMSE = 1.40 g C·m^{-2}·d^{-1}). Our results also suggest that the RF models generally performed better than the BPNN models, and the PFT-specific models performed better than the universal models, although their differences were minor (Table 2). For the three VIs, both RF and BPNN methods show that NIR$_v$-based models outperformed SIF-based models, followed by LAI-based ones.

Table 2. Performance of optimal models for each vegetation index. Note that PFT was used as a predictor in the PFT-specific models, while it was not included in the universal models.

Type	Vegetation Index	RF		BPNN	
		R2	RMSE (g C·m^{-2}·d^{-1})	R2	RMSE (g C·m^{-2}·d^{-1})
PFT-Specific	SIF	0.86	1.46	0.84	1.43
	NIR$_v$	0.87	1.40	0.85	1.43
	LAI	0.86	1.45	0.83	1.50
Universal	SIF	0.85	1.54	0.81	1.60
	NIR$_v$	0.85	1.51	0.83	1.54
	LAI	0.84	1.54	0.79	1.65

Figure 2 shows the latitudinal distribution of universal models' performance using the RF method. Prediction accuracy discrepancies are mostly present among various latitudinal zones and different PFTs, rather than different models. The GPP estimation models derived from the different combinations of BPNN or RF methods and universal or PFT-specific configurations show similar performance (Figures 2 and S5–S7), indicating that rather than the machine learning algorithms, the uncertainties and sample size of the training data seem to be the main factors that affect the performance of the GPP estimation models when estimating regional GPP.

The R^2 values of the models appear to be greater than or near to 70% in all latitudinal zones except 70°N–80°N (R^2 = 0.27–0.34), with minor differences between the models based on different VIs and machine learning algorithms (Figures 2 and S2–S4). Overall, the regional and PFT-specific performance of the GPP models, which is also closely related to the sample size, is much better in the northern latitudes (the median of R^2 is 0.85; the median of RMSE is 1.12 g C·m^{-2}·d^{-1}) than latitudes near the equator (the median of R^2 is 0.74; the median of RMSE is 1.19 g C·m^{-2}·d^{-1}). With a dense distribution of flux tower sites (Figure 1) in the northern mid-high latitudes (30°N–70°N), diverse PFTs, and a large amount of data available, the models performed well on all types of vegetation (forests, shrublands, grasslands, croplands, and wetlands) in this region (R^2 ranges from 0.48 to 0.97, and RMSE ranges from 0.51 to 3.13 g C·m^{-2}·d^{-1}). However, the poor performance in the northern high latitudes (70°N–80°N) is likely because only one flux tower was available (the open shrublands site, Ru-Cok). Similarly, the poor model performance for EBF near the equator (R^2 = 0.10–0.22) is also due to the limited flux tower observations (Figure 1).
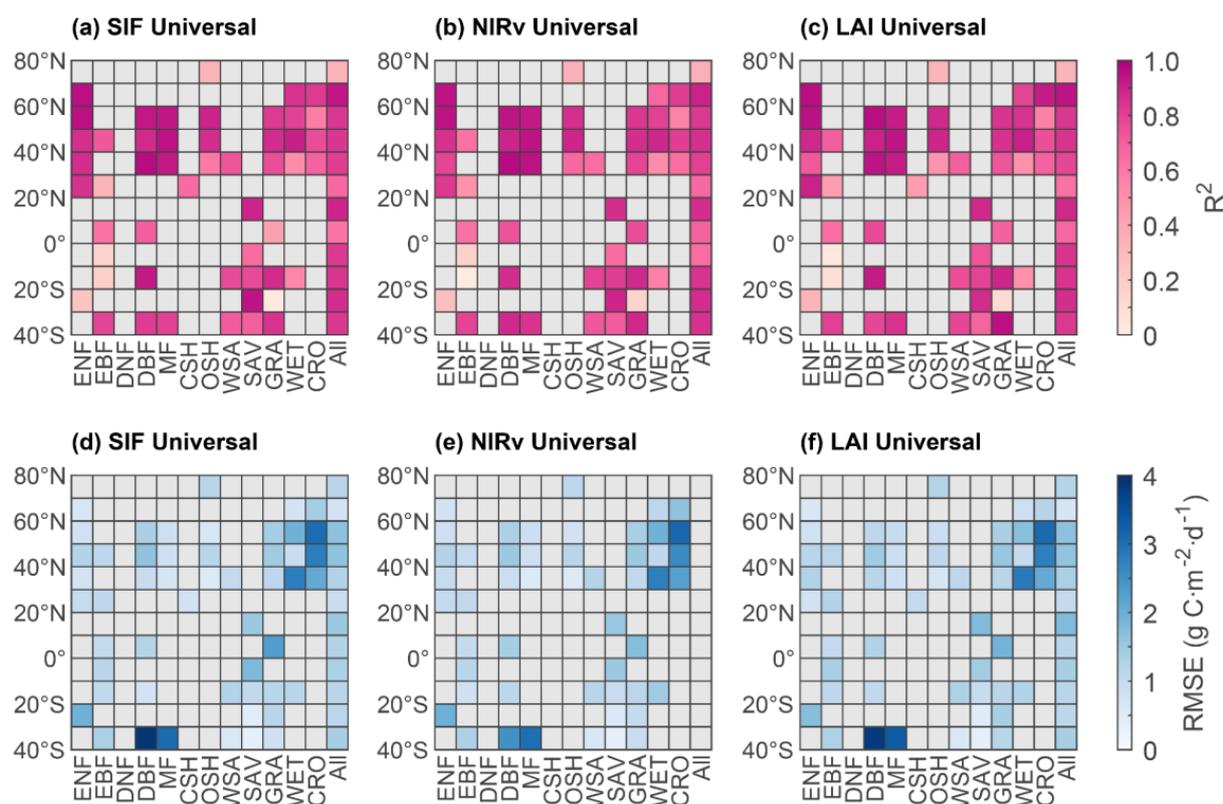
Figure 2. Universal models' performance for each PFT along different latitude gradients using the RF method. (**a**–**c**) The R^2 values of the universal models that each vegetation index used as one of the model predictors. (**d**–**f**) The RMSE values of the universal models that each vegetation index used as one of the model predictors. The gray squares represent that there is no flux tower for the corresponding PFT and latitudinal band.

3.2. Comparison between VIs-Based and Ecosystem Model-Simulated GPP Datasets

A set of three global gridded GPP products for 2003–2018 were generated for each vegetation index based on universal models with optimal configurations, i.e., the optimal combination of the VIs and the other associated variables using the RF algorithm. We compared our Vis- and machine learning-based GPP datasets with three other GPP datasets (FLUXCOM GPP, GOSIF GPP, and TRENDY GPP).

The spatial patterns of annual mean GPP for the six GPP products are similar: the highest values are found in tropical rainforest regions, with generally lower GPP values found in arid regions such as Australia, Central Asia, the southwestern United States, and southwestern Africa, as well as the cold regions at high northern latitudes (Figure 3). All products' latitudinal profiles also have good consistency (Figure S6).

Global land is divided into the northern hemisphere (NH: 30°N–90°N) and tropical and southern hemisphere (SH+Trop: 90° S–30°N). Among our three sets of GPP products, both globally and regionally, LAI-based GPP had the highest annual mean GPP, whereas NIR$_\text{v}$-based was the lowest; the interannual variability of the three sets is quite similar (Figure S7). Both SIF-based GPP and GOSIF-GPP were generated from GOSIF, and their annual mean GPP values are relatively close in the SH+Trop area; however, the former has higher annual mean and interannual variability in GPP than the latter, both globally and over the NH (Figures 4 and S7). The rest of the products are within the range of the 10 process-based models for each study scale, except for global LAI-based annual mean GPP (Figure 3). Interannual variability in FLUXCOM is substantially lower than that of other products because it did not account for the CO_2 fertilization effect (CFE) (Figure S8). Despite GOSIF-GPP not considering rising CO_2 concentrations, its interannual variability

was closer to TRENDY MMEM and our three GPP products. GOSIF may reflect the role of CO_2 in GPP to some extent.
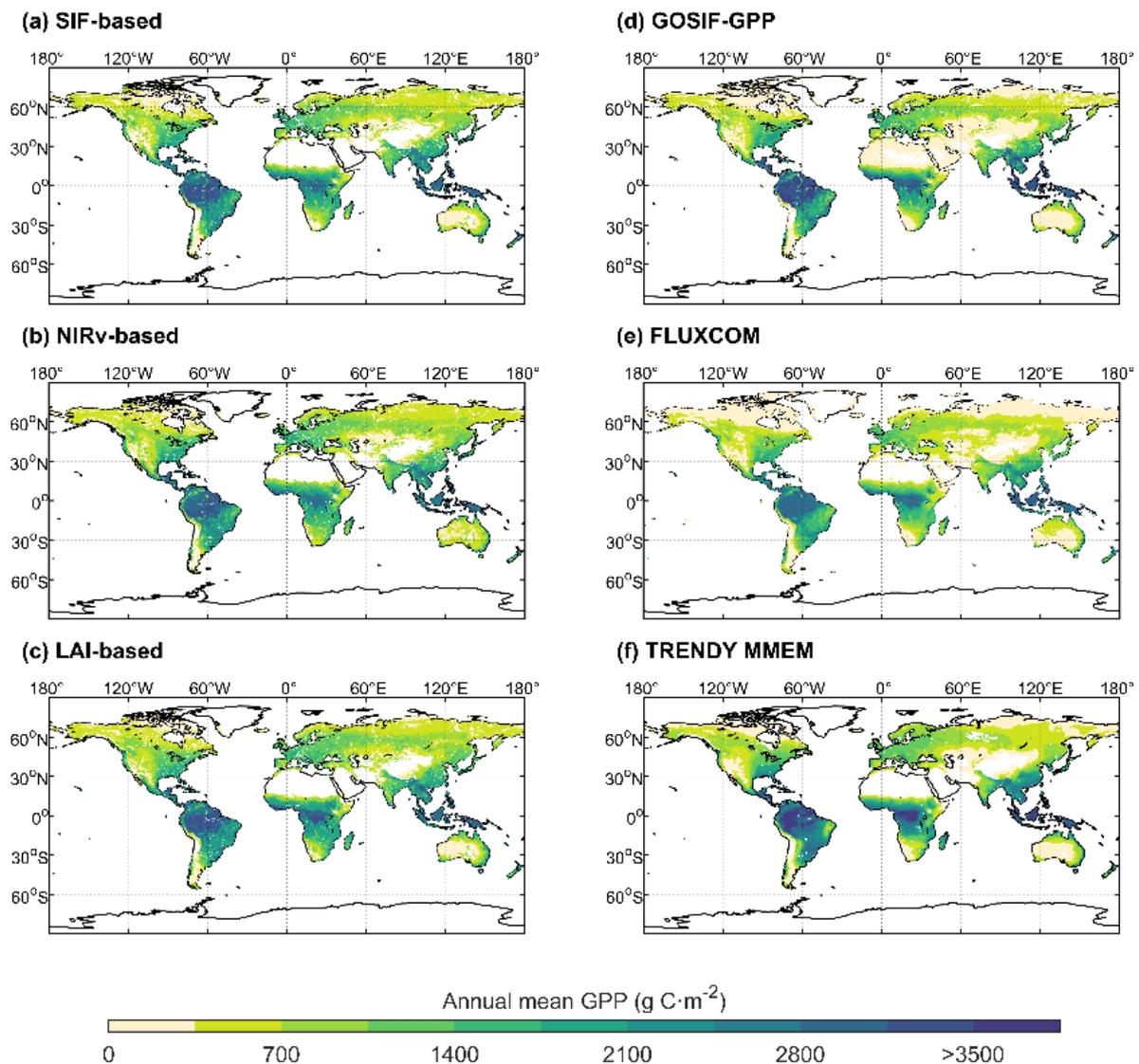

Figure 3. Spatial patterns of global annual mean GPP for 2003–2018. (**a**–**c**) represent GPP estimates based on each vegetation index with optimal configurations. (**d**) GOSIF-GPP; (**e**) FLUXCOM GPP. (**f**) represents the average of the multi-model ensemble mean.

On the other hand, the linear trends of the various GPP products are quite different (Figure S7). Except for FLUXCOM, the remaining GPP products have an increasing trend in annual mean values (Figure S7a–c). Regarding interannual variability of anomalies and trends, our three GPP products are very close to each other (Figure S7d—Global: 0.52–0.54Pg C·yr^{-2}; NH: 0.27–0.29 Pg C·yr^{-2}; SH+Trop: 0.24–0.26 Pg C·yr^{-2}). GOSIF-GPP and TRENDY MMEM all capture the significant increase in GPP in northern Europe, East Asia, South Asia, and northern North America (Figure S9). However, the Amazon region displays a wide divergence: our products and GOSIF-GPP show a decreasing trend in GPP (Figure S9a–d), but TRENDY MMEM shows a significant increase (Figure S9f). Furthermore, for eastern Siberia, LAI-based GPP, GOSIF-GPP, and TRENDY-MMEM consistently indicate a decreasing GPP trend (Figure S9c,d,f), but SIF- and NIR$_v$-based GPP regularly show considerable GPP increases (Figure S9a,b).
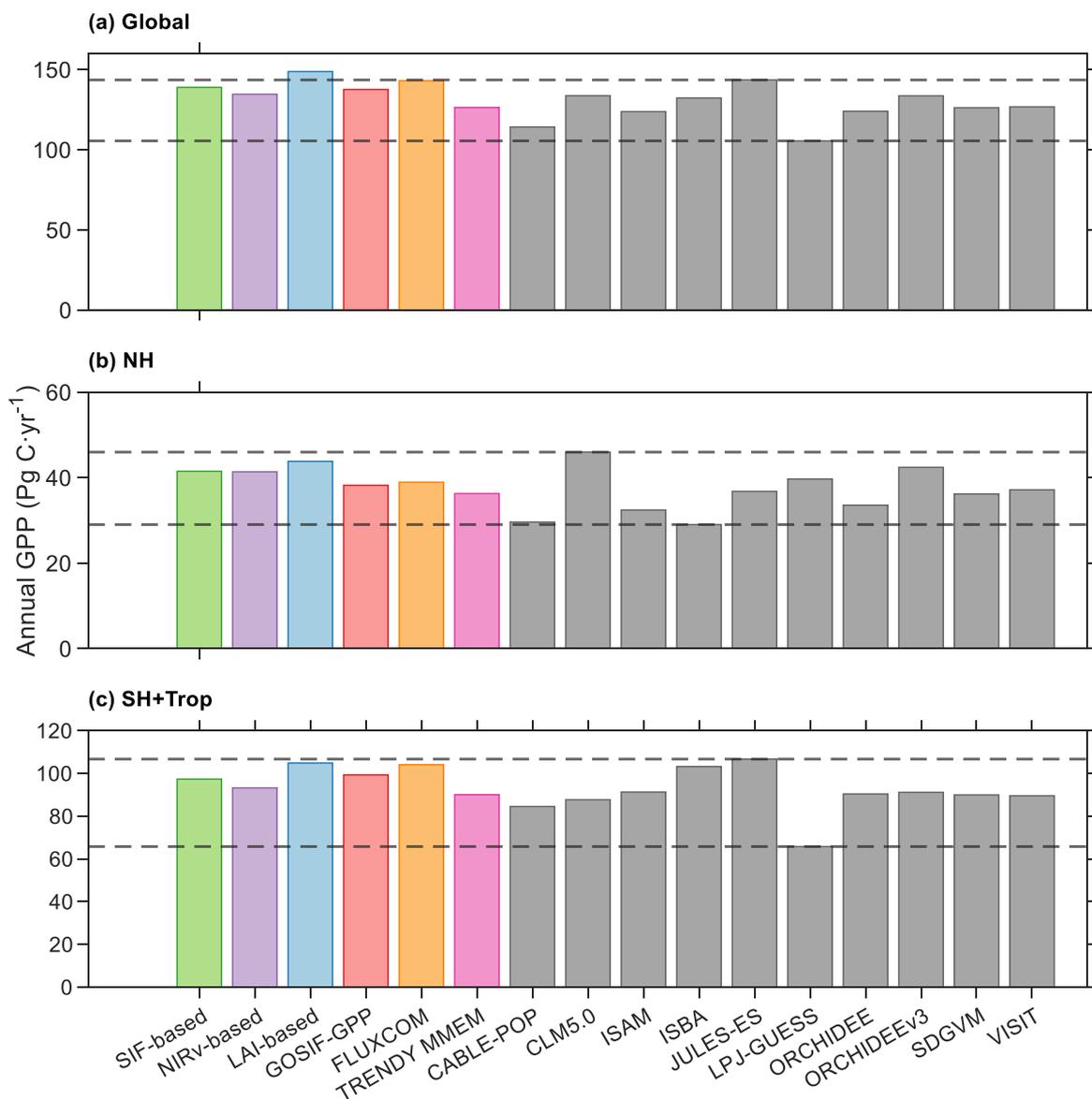
Figure 4. Comparisons of annual mean GPP among different GPP products. Global land is divided into the northern hemisphere (NH: 30°N–90°N) and tropical and southern hemisphere (SH+Trop: 90°S–30°N). The SIF-based, NIR$_v$-based, and LAI-based labels represent the GPP estimates based on each vegetation index with optimal configurations. The dashed lines represent the maximum and minimum values of the ten processed models in the TRENDY group ('S3' scenarios). TRENDY MMEM represents the average of the multi-model ensemble mean.

3.3. The Critical Factors for the Machine Learning Algorithms-Based GPP Estimation Model

We further investigated the factors that are the best candidates for developing the GPP estimation models with different VIs. The GPP models developed based only on VIs suggest that SIF is the best proxy for the estimation of global GPP using the RF algorithm ($R^2 = 0.67$, RMSE = 2.24 g C·m^{-2}·d^{-1}), followed by NIR$_v$ ($R^2 = 0.61$, RMSE = 2.45 g C·m^{-2}·d^{-1}) and LAI ($R^2 = 0.50$, RMSE = 2.79 g C·m^{-2}·d^{-1}). Likewise, the models developed based on the BPNN algorithm also suggest that SIF is the best VI for global GPP estimation if other variables, i.e., CO_2, PFT, plant traits, and climate factors, were not included ($R^2 = 0.70$, RMSE = 2.13 g C·m^{-2}·d^{-1}, Figure S10). Nevertheless, VIs are the most critical input when compared to other variables, confirming the usefulness of these VIs in estimating global GPP.

Incorporating atmospheric CO_2 concentration helps the estimation of GPP based on VIs and machine learning algorithms (Figure 5). However, the benefits of incorporating CO_2 are different for different VIs. We found that improvement in model performance was more evident for LAI-based models, with the R^2 value increasing from 0.50 to 0.58 and RMSE decreasing from 2.79 to 2.52 g C·m^{-2}·d^{-1}. The improvements for NIR$_v$-based and SIF-based models were not that significant, with the R^2 value increasing from 0.61 to 0.66 and from 0.67 to 0.70 for NIR$_v$ and SIF, respectively, while the RMSE values respectively decreased from 2.45 to 2.30 g C·m^{-2}·d^{-1} and from 2.24 to 2.13 g C·m^{-2}·d^{-1} (Figure 5). We also found that explicitly including PFT in the RF model achieved similar improvements to including CO_2. Nevertheless, incorporating both CO_2 and PFT did not achieve an evident improvement in model performance compared to the models that incorporated CO_2 or PFT.
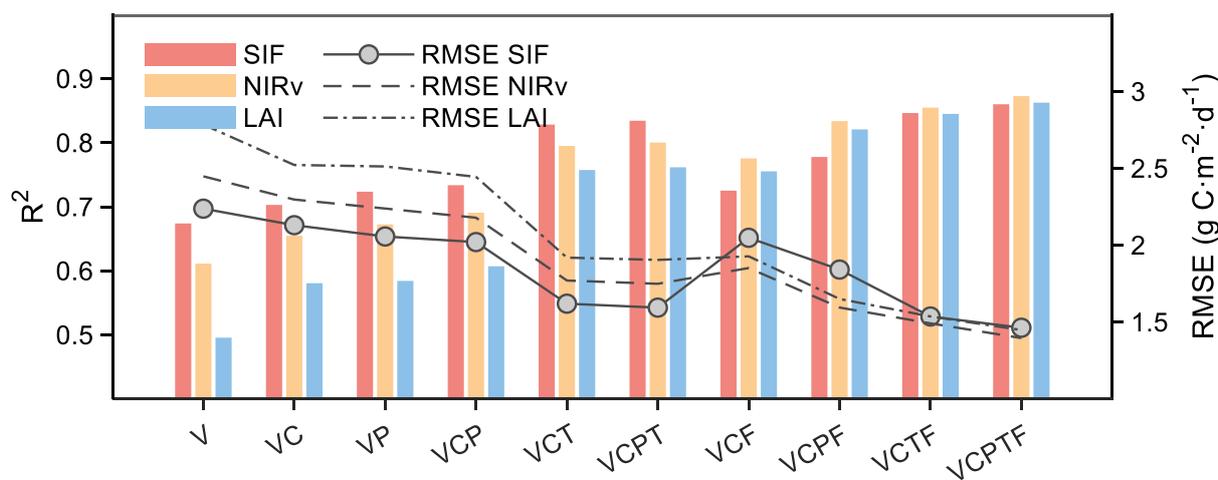

Figure 5. The performance of RF models using different combinations of predictors. The hyperparameters of the models using other combinations of inputs were consistent with the optimal model. The left and right axes represent the mean predicted R2 and RMSE values of the models over 50 repetitions, with negligible differences each time. The variables used to develop the GPP estimation models are represented by the combination of their abbreviations, i.e., VIs (V), CO_2 (C), PFT (P), plant traits (T), and climate factors (F) (Table S2). For example, 'VCT' denotes the RF GPP model that used VIs, CO_2, and plant traits as input.

Including plant traits and/or climate factors led to a further notable improvement in model performance (Figure 5). For the SIF models, plant traits significantly improved model performance (R^2 increased by 0.13 and RMSE decreased by 0.51 g C·m^{-2}·d^{-1} from the SC models to the SCT models) while climate factors did not (R^2 and RMSE were 0.70 and 2.13 g C·m^{-2}·d^{-1} for the SC models and 0.73 and 2.05 g C·m^{-2}·d^{-1} for the SCF models). For the NIR$_v$ and LAI models (VC models), considering plant traits (VCT models) or climate factors (VCF models) significantly contributed to model performance. Furthermore, models that incorporated both plant traits and climatic factors (VCPTF models) achieved the best performance (R^2 = 0.87 and RMSE = 1.40 g C·m^{-2}·d^{-1} for NIR$_v$; R^2 = 0.86 and RMSE = 1.45 g C·m^{-2}·d^{-1} for LAI). We found that results did not change for the machine learning algorithms because the combination test of VIs, PFT, plant traits, and climate factors based on the BPNN algorithm showed similar results to tests based on the RF algorithm (Figure S10).

Figure 6 shows the relative importance of the factors in the optimal machine learning GPP estimation models (see Methods). For the SIF-based and NIR$_v$-based models (both PFT-specific and universal), vegetation index was the most powerful explanatory variable with the most significant importance score for estimating GPP. SIF plays a dominant role in reconstructing GPP (Figure 6a,b). However, a notable decrease in performance was observed when climate factors and plant traits were excluded from the SIF-based models (OOB R^2 = 0.82 to 0.73 while OOB RMSE = 1.66 to 2.02 g C·m^{-2}·d^{-1} from four

to three predictors in Figure S11a and OOB R^2 = 0.78 to 0.70 while OOB RMSE = 1.83 to 2.13 g C·m^{-2}·d^{-1} from three to two predictors in Figure S11b). On the contrary, for the PFT-specific LAI-based models, PFT is the most critical predictor (Figure 6e), while the major factor in the universal model is leaf N content (Figure 6f). For the NIR$_v$-based and LAI-based models, SRAD is the most significant scoring climate variable, while it was not selected for the optimal SIF-based model. In all six types of models, we consistently found that preseason temperature was selected as an important factor in estimating GPP.
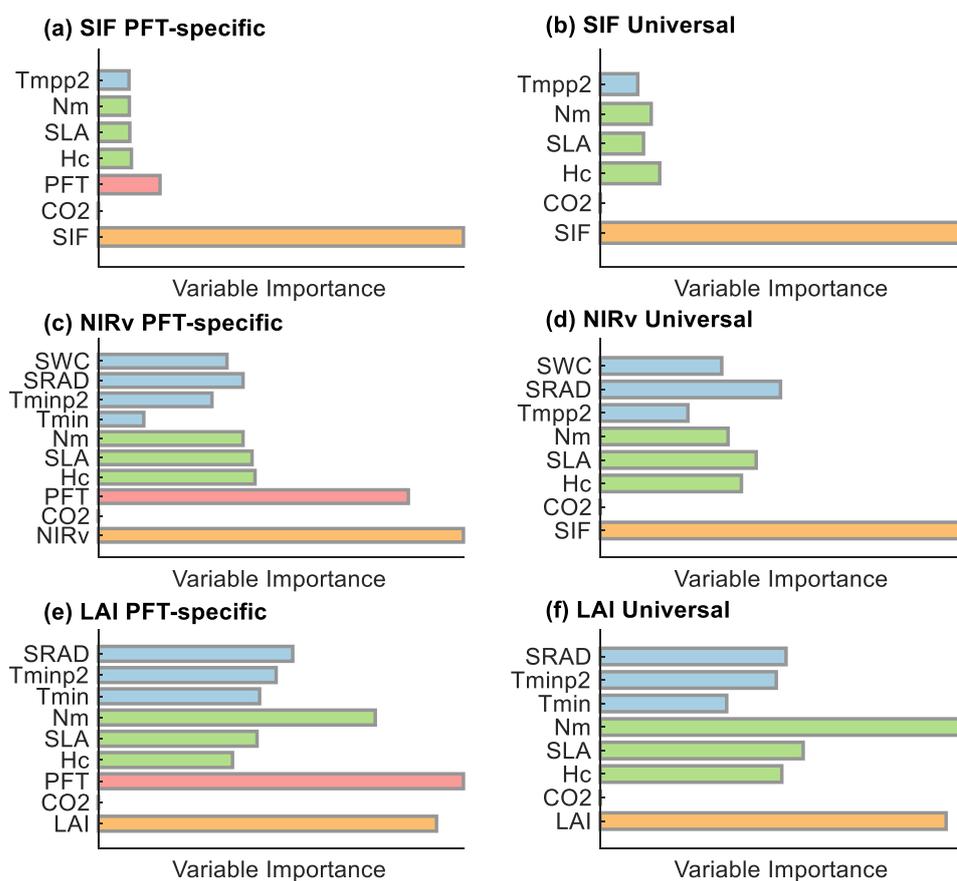
Figure 6. Importance of variables for optimal performance in random forest models. (**a,c,e**) represent PFT-specific models. (**b,d,f**) represent universal models. Rows one through three represent the models constructed based on SIF, NIR$_v$, and LAI, respectively. Variable importance scores were normalized using the minimum–maximum method. The explanatory variables are marked with "px" at the end of the abbreviation, indicating the climate factors for the previous x months (e.g., Tminp2 represents the minimum air temperature two months ago).

4. Discussion

4.1. Different Performance of VIs in GPP Estimation

SIF has the strongest linear correlation with flux tower GPP estimates among the three VIs. When using individual vegetation indices to construct GPP models, the performance of the SIF-based model is higher than that of the NIR$_v$- and LAI-based models. Unlike NIRv and LAI, SIF responds to changes in canopy structure and photosynthetically active radiation absorbed by chlorophyll [69]. In addition, SIF is sensitive to vegetation with slight interannual variation in greenness and can be retrieved under some clouds and aerosols [70]. Several studies have shown that SIF measurements are superior for detecting photosynthetic activity in evergreen forests [34,71,72]. Wang et al. [73] found that SIF is more accurate than NIR$_v$ and EVI in monitoring the phenology of drylands since it is not contaminated by soil background. Another study found that SIF is superior to traditional VIs and the new vegetation index NIR$_v$ in estimating large-scale GPP [74]. However, the

shortcomings of SIF are apparent. There is no satellite designed explicitly for SIF, and SIF retrievals have a relatively short time span and coarse spatial resolution, limiting long-term global GPP estimation and prediction [75].

Nevertheless, when combined with vegetation's biophysical properties and environmental and climate factors, the NIR_v-based models outperformed SIF-based and LAI-based models. NIR_v is a spectral reflectance-based vegetation index, and LAI is a vegetation characteristic parameter with actual physical significance, both of which can capture structural changes in vegetation canopies. LAI and leaf level CO_2 are almost equally crucial for GPP according to Hinojo-Hinojo et al. [76]. SLA is a critical trait that reflects the photosynthetic capacity of plants, which is related to leaf-scale CO_2 uptake and can manifest the growth and reproduction strategy under certain environmental conditions. With the combined influence of CO_2 and SLA, NIR_v can provide more universal predictions of GPP than LAI. Furthermore, NIR_v captures the fraction of sunlit and shaded leaves in vegetation canopies, while LAI ignores the difference in photosynthetic capacity between these two types of leaves [77,78]. Our findings suggest that NIR_v can better predict GPP when combined with other predictors and compensate for the spatial–temporal limitations of SIF measurements.

4.2. Environmental Factors and Plant Traits Paired with VIs

Most machine learning-based GPP estimation models indirectly represent the effects of CO_2 fertilization on GPP by incorporating FPAR or other vegetation indices [21,79,80], although multiple pieces of evidence show that the rising atmospheric CO_2 concentration is one of the dominant factors driving vegetation growth [81–83]. We explicitly tested the contribution of incorporating the atmospheric CO_2 concentration into our study's GPP models and how it affected performance. We found that incorporating CFE did contribute to the performance of the VI-based machine learning GPP model, especially for models developed based on LAI (Figure 5).

Our analysis (Figure 2) shows that preseason temperature is pivotal for all three VIs-based models. Temperature affects photosynthetic intensity at the leaf scale by altering the activity of enzymes in the chloroplast, with either high or low temperatures inhibiting enzyme activity [84]. Niu et al. [85] evaluated the flux tower data and found that the relationship between vegetation photosynthesis and temperature at the ecosystem scale followed the same pattern as the leaf scale. Numerous studies have revealed that temperature is one of the dominant limiting factors in high northern latitudes [86–89]. Temperature has been confirmed as the key environmental factor controlling the rising GPP of northern Eurasia [90]. Furthermore, preseason temperature also strongly regulates the phenology of plants, which essentially determines the length of the photosynthetic active period and, subsequently, annual GPP [86,91–94]. However, it has also been found that warming leads to earlier spring phenology, reducing vegetation's peak growth in North American boreal forests [95].

Solar radiation is crucial for providing energy for photosynthesis. On the one hand, incoming shortwave radiation regulates leaf development and controls the process of leaf senescence by affecting plant hormones (e.g., ethylene and abscisic acid). On the other hand, radiation intensity determines the photosynthetic rate of vegetation [96]. Several recent reports have suggested that when reflectance-based VIs incorporating biophysical and biochemical properties are paired with incoming shortwave radiation, changes in GPP are more easily captured [97,98]. This is in line with our findings that radiation participates as the most important climatic variable in NIR_v- and LAI-based models (Figure 2). In the SIF-based models, we also discovered that radiation is insignificant. Solar radiation has proven to be a dominant driver of SIF yield based on theoretical and experimental analysis [70,99,100]. Fluorescence signals are considered the most direct response to radiation absorbed by the vegetation canopy during photosynthesis, which is coupled with the radiation.

Tramontana et al. [35] concluded that the performance of models using remote sensing data alone is comparable to the best model, implying that climate factors have a minor

effect. However, we found that there is some room for improving model performance when predicting GPP using remote sensing data alone, which is achieved by selecting an optimal combination of features. Previous studies implied that NIR_v, to some extent, captures the influences of climate on canopy development. In other words, the accuracy of NIR_v in predicting GPP is not significantly increased by meteorological data [101]. Nevertheless, our results suggest that incorporating meteorological information increased accuracy significantly for the NIR_v- and LAI-based GPP estimation models (Figure 3). This is likely due to the fact that NIR_v and LAI describe vegetation canopy structure well but cannot fully reflect photosynthetic activity, and the machine learning algorithm can represent the nonlinear interactions between the climate variable and the VIs.

Canopy height, specific leaf area, and leaf nitrogen content were influential explanatory variables in all of the GPP models we developed. Given the globally consistent correlations between Nm, SLA, and rate of photosynthesis [42,43], these photosynthesis-related plant traits can provide substantial constraints on GPP estimates at the global scale [102]. We found that Nm is a crucial explanatory variable in LAI-based models, and its importance in models across biomes even outweighs that of LAI. Our results are consistent with previous studies, which found that combining Nm and LAI could improve the explanation of variability in GPP [103].

Although a universal slope between OCO-2 SIF and flux tower GPP was previously found for diverse biomes by Li et al. [74], two subsequent investigations revealed that the SIF–GPP relationship is not constant throughout time and space but is regulated by environmental conditions [37,38]. The explanation for these findings is thought to be the variances in vegetation canopy structure [37,98], though the mechanism is unknown. We constructed two types of models, PFT-specific and universal, and found that the accuracy difference between them is insignificant. Plant traits associated with photosynthesis may vary considerably among biomes and vegetation types in different geographic zones. The selected plant traits (i.e., Hc, SLA, and Nm) contain a large amount of canopy structure information that is dependent on specific biomes, which aids the predictions of GPP spatial patterns and mechanistically better reveals the photosynthetic capacity of vegetation.

4.3. Sources of Uncertainty

The keys to the success of data-driven models based on machine learning algorithms when upscaling the relationships between the ancillary variables and GPP from site level to global scale are (1) sufficient sample size for training the model, (2) predictors matching the spatial scale of the target variable at the site scale, and (3) gridded inputs with the same spatial and temporal resolution with global coverage. Although the method we used in this study achieved reasonable performance while employing the fewest explanatory variables possible, some aspects still need further improvements.

The original spatial resolution of gridded datasets, including VIs, PFT, plant traits, and climate data, is inconsistent, and resampling introduces uncertainties in global GPP estimates. It is widely assumed that the smaller the pixel corresponding to the location of the flux tower, the more reliable the pixel value is, i.e., the more representative of the environmental conditions around the tower. Limited by the spatial resolution of remote sensing and reanalysis data, the predictors cannot match EC towers' GPP estimates well at the ecosystem scale. The spatial distribution of the flux sites is not even and is notably sparse for some areas and vegetation types (e.g., equatorial regions and EBF). In order to better understand these important yet poorly understood ecosystems' physiological responses to global environmental changes, we call for expansion of eddy covariance sites and ecological research stations in these areas. The plant traits utilized take the form of gridded data that were developed based on the TRY database and which are constant in time. This may be altered as a result of climate change [51]. We expect to clarify the mechanisms that lead to plant trait changes and validate or improve our models using time-varying data.

5. Conclusions

Two widely used machine learning algorithms, i.e., RF and ANN, were employed to estimate global GPP. We dynamically selected variables by removing the ones with the lowest importance and then re-learning, which allows us to achieve the highest possible prediction accuracy using as few explanatory variables as possible, considerably improving prediction efficiency. We systematically compared the performance divergencies of different VIs in predicting GPP and determined the best way to combine each vegetation index with other explanatory variables, reflecting the close relationship between plant traits and ecosystem functions.

Under the data-driven approach, we investigated divergencies in the performance of GPP models developed based on three VIs (i.e., SIF, NIR_v, and LAI). We identified which biophysical properties, environmental factors, and climatic variables paired with individual VIs had the best predictive power for GPP and elucidated the potential mechanisms. We show that each vegetation index is a crucial driver of simulated spatial and temporal variability in GPP, and the SIF-based model performs best when modeled using a single vegetation index. However, NIR_v combined with CO_2, plant traits, and climatic factors can achieve the highest prediction accuracy. Furthermore, we consistently found that preseason temperature was selected as an important factor for estimating GPP in all six models. For the NIR_v- and LAI-based GPP prediction models, solar radiation is the most critical climatic factor. We found that plant traits provide crucial canopy structure information, which effectively enhances the accuracy of all GPP models. Climatic variables are essential factors for improving the accuracy of NIR_v- or LAI-based GPP models, but not for SIF-based models.

Our study provides valuable information on the configuration of data-driven models designed to improve the accuracy of global GPP predictions and provides insights into the underlying physiological and ecological mechanisms involved.

(s.a.sitch@exeter.ac.uk) and P. Friedlingstein (p.friedlingstein@exeter.ac.uk). Foliage economic traits (Nm and SLA) are from Butler et al. [54]. Canopy height data are from Simard et al. [55]. Atmospheric carbon dioxide data can be download at: https://gml.noaa.gov/ccgg/trends (accessed on 24 September 2021).

References

1. Friedlingstein, P.; O'Sullivan, M.; Jones, M.W.; Andrew, R.M.; Hauck, J.; Olsen, A.; Peters, G.P.; Peters, W.; Pongratz, J.; Sitch, S.; et al. Global Carbon Budget 2020. *Earth Syst. Sci. Data* **2020**, *12*, 3269–3340. [CrossRef]
2. He, M.; Ju, W.; Zhou, Y.; Chen, J.; He, H.; Wang, S.; Wang, H.; Guan, D.; Yan, J.; Li, Y.; et al. Development of a two-leaf light use efficiency model for improving the calculation of terrestrial gross primary productivity. *Agric. For. Meteorol.* **2013**, *173*, 28–39. [CrossRef]
3. Beer, C.; Reichstein, M.; Tomelleri, E.; Ciais, P.; Jung, M.; Carvalhais, N.; Rödenbeck, C.; Arain, M.A.; Baldocchi, D.; Bonan, G.B.; et al. Terrestrial Gross Carbon Dioxide Uptake: Global Distribution and Covariation with Climate. *Science* **2010**, *329*, 834–838. [CrossRef]
4. Damm, A.; Elbers, J.A.N.; Erler, A.; Gioli, B.; Hamdi, K.; Hutjes, R.; Kosvancova, M.; Meroni, M.; Miglietta, F.; Moersch, A.; et al. Remote sensing of sun-induced fluorescence to improve modeling of diurnal courses of gross primary production (GPP). *Glob. Chang. Biol.* **2010**, *16*, 171–186. [CrossRef]
5. Garbulsky, M.F.; Filella, I.; Verger, A.; Peñuelas, J. Photosynthetic light use efficiency from satellite sensors: From global to Mediterranean vegetation. *Environ. Exp. Bot.* **2014**, *103*, 3–11. [CrossRef]
6. Ruimy, A.; Saugier, B.; Dedieu, G. Methodology for the estimation of terrestrial net primary production from remotely sensed data. *J. Geophys. Res. Atmos.* **1994**, *99*, 5263–5283. [CrossRef]
7. Anav, A.; Friedlingstein, P.; Beer, C.; Ciais, P.; Harper, A.; Jones, C.; Murray-Tortarolo, G.; Papale, D.; Parazoo, N.C.; Peylin, P.; et al. Spatiotemporal patterns of terrestrial gross primary production: A review. *Rev. Geophys.* **2015**, *53*, 785–818. [CrossRef]
8. Sitch, S.; Friedlingstein, P.; Gruber, N.; Jones, S.D.; Murray-Tortarolo, G.; Ahlström, A.; Doney, S.C.; Graven, H.; Heinze, C.; Huntingford, C.; et al. Recent trends and drivers of regional sources and sinks of carbon dioxide. *Biogeosciences* **2015**, *12*, 653–679. [CrossRef]
9. Chen, J.M.; Mo, G.; Pisek, J.; Liu, J.; Deng, F.; Ishizawa, M.; Chan, D. Effects of foliage clumping on the estimation of global terrestrial gross primary productivity. *Glob. Biogeochem. Cycles* **2012**, *26*. [CrossRef]
10. Coops, N.C.; Ferster, C.J.; Waring, R.H.; Nightingale, J. Comparison of three models for predicting gross primary production across and within forested ecoregions in the contiguous United States. *Remote Sens. Environ.* **2009**, *113*, 680–690. [CrossRef]
11. Abatzoglou, J.T.; Dobrowski, S.Z.; Parks, S.A.; Hegewisch, K.C. TerraClimate, a high-resolution global dataset of monthly climate and climatic water balance from 1958-2015. *Sci. Data* **2018**, *5*, 170191. [CrossRef] [PubMed]
12. Stocker, B.D.; Zscheischler, J.; Keenan, T.F.; Prentice, I.C.; Seneviratne, S.I.; Peñuelas, J. Drought impacts on terrestrial primary production underestimated by satellite monitoring. *Nat. Geosci.* **2019**, *12*, 264–270. [CrossRef]
13. Xiao, X.M.; Hollinger, D.; Aber, J.; Goltz, M.; Davidson, E.A.; Zhang, Q.Y.; Moore, B. Satellite-based modeling of gross primary production in an evergreen needleleaf forest. *Remote Sens. Environ.* **2004**, *89*, 519–534. [CrossRef]
14. 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]
15. He, H.; Liu, M.; Xiao, X.; Ren, X.; Zhang, L.; Sun, X.; Yang, Y.; Li, Y.; Zhao, L.; Shi, P.; et al. Large-scale estimation and uncertainty analysis of gross primary production in Tibetan alpine grasslands. *J. Geophys. Res. Biog.* **2014**, *119*, 466–486. [CrossRef]
16. Yuan, W.; Cai, W.; Nguy-Robertson, A.L.; Fang, H.; Suyker, A.E.; Chen, Y.; Dong, W.; Liu, S.; Zhang, H. Uncertainty in simulating gross primary production of cropland ecosystem from satellite-based models. *Agric. For. Meteorol.* **2015**, *207*, 48–57. [CrossRef]
17. Peng, Y.; Gitelson, A.A.; Keydan, G.; Rundquist, D.C.; Moses, W. Remote estimation of gross primary production in maize and support for a new paradigm based on total crop chlorophyll content. *Remote Sens. Environ.* **2011**, *115*, 978–989. [CrossRef]
18. Wu, C.; Chen, J.M.; Huang, N. Predicting gross primary production from the enhanced vegetation index and photosynthetically active radiation: Evaluation and calibration. *Remote Sens. Environ.* **2011**, *115*, 3424–3435. [CrossRef]
19. Jung, M.; Koirala, S.; Weber, U.; Ichii, K.; Gans, F.; Camps-Valls, G.; Papale, D.; Schwalm, C.; Tramontana, G.; Reichstein, M. The FLUXCOM ensemble of global land-atmosphere energy fluxes. *Sci. Data* **2019**, *6*, 74. [CrossRef]
20. Tramontana, G.; Jung, M.; Schwalm, C.R.; Ichii, K.; Camps-Valls, G.; Ráduly, B.; Reichstein, M.; Arain, M.A.; Cescatti, A.; Kiely, G. Predicting carbon dioxide and energy fluxes across global FLUXNET sites with regression algorithms. *Biogeosciences* **2016**, *13*, 4291–4313. [CrossRef]
21. Jung, M.; Reichstein, M.; Margolis, H.A.; Cescatti, A.; Richardson, A.D.; Arain, M.A.; Arneth, A.; Bernhofer, C.; Bonal, D.; Chen, J.; et al. Global patterns of land-atmosphere fluxes of carbon dioxide, latent heat, and sensible heat derived from eddy covariance, satellite, and meteorological observations. *J. Geophys. Res.* **2011**, *116*, G00J07. [CrossRef]

22. Braghiere, R.K.; Quaife, T.; Black, E.; He, L.; Chen, J.M. Underestimation of Global Photosynthesis in Earth System Models Due to Representation of Vegetation Structure. *Glob. Biogeochem. Cycles* **2019**, *33*, 1358–1369. [CrossRef]

23. Chen, M.; Rafique, R.; Asrar, G.R.; Bond-Lamberty, B.; Ciais, P.; Zhao, F.; Reyer, C.P.O.; Ostberg, S.; Chang, J.; Ito, A.; et al. Regional contribution to variability and trends of global gross primary productivity. *Environ. Res. Lett.* **2017**, *12*, 105005. [CrossRef]

24. Huang, K.; Xia, J.; Wang, Y.; Ahlstrom, A.; Chen, J.; Cook, R.B.; Cui, E.; Fang, Y.; Fisher, J.B.; Huntzinger, D.N.; et al. Enhanced peak growth of global vegetation and its key mechanisms. *Nat. Ecol. Evol.* **2018**, *2*, 1897–1905. [CrossRef] [PubMed]

25. Tucker, C.J. Red and photographic infrared linear combinations for monitoring vegetation. *Remote Sens. Environ.* **1979**, *8*, 127–150. [CrossRef]

26. Jiang, Z.; Huete, A.; Didan, K.; Miura, T. Development of a two-band enhanced vegetation index without a blue band. *Remote Sens. Environ.* **2008**, *112*, 3833–3845. [CrossRef]

27. Xiao, Z.; Liang, S.; Wang, J.; Chen, P.; Yin, X.; Zhang, L.; Song, J. Use of General Regression Neural Networks for Generating the GLASS Leaf Area Index Product From Time-Series MODIS Surface Reflectance. *IEEE Trans. Geosci. Remote Sens.* **2014**, *52*, 209–223. [CrossRef]

28. Zhu, Z.; Bi, J.; Pan, Y.; Ganguly, S.; Anav, A.; Xu, L.; Samanta, A.; Piao, S.; Nemani, R.; Myneni, R. Global Data Sets of Vegetation Leaf Area Index (LAI)3g and Fraction of Photosynthetically Active Radiation (FPAR)3g Derived from Global Inventory Modeling and Mapping Studies (GIMMS) Normalized Difference Vegetation Index (NDVI3g) for the Period 1981 to 2011. *Remote Sens.* **2013**, *5*, 927–948. [CrossRef]

29. Yan, K.; Zou, D.; Yan, G.; Fang, H.; Weiss, M.; Rautiainen, M.; Knyazikhin, Y.; Myneni, R.B. A Bibliometric Visualization Review of the MODIS LAI/FPAR Products from 1995 to 2020. *J. Remote Sens.* **2021**, *2021*, 7410921. [CrossRef]

30. Richardson, A.D.; Braswell, B.H.; Hollinger, D.Y.; Jenkins, J.P.; Ollinger, S.V. Near-surface remote sensing of spatial and temporal variation in canopy phenology. *Ecol. Appl.* **2009**, *19*, 1417–1428. [CrossRef] [PubMed]

31. Verger, A.; Filella, I.; Baret, F.; Peñuelas, J. Vegetation baseline phenology from kilometric global LAI satellite products. *Remote Sens. Environ.* **2016**, *178*, 1–14. [CrossRef]

32. Badgley, G.F.; Field, C.B.; Berry, J.A. Canopy near-infrared reflectance and terrestrial photosynthesis. *Sci. Adv.* **2017**, *3*, e1602244. [CrossRef] [PubMed]

33. Camps-Valls, G.; Campos-Taberner, M.; Moreno-Martínez, Á.; Walther, S.; Duveiller, G.; Cescatti, A.; Mahecha, M.D.; Muñoz-Marí, J.; García-Haro, F.J.; Guanter, L.; et al. A unified vegetation index for quantifying the terrestrial biosphere. *Sci. Adv.* **2021**, *7*, eabc7447. [CrossRef] [PubMed]

34. Pierrat, Z.; Magney, T.; Parazoo, N.C.; Grossmann, K.; Bowling, D.R.; Seibt, U.; Johnson, B.; Helgason, W.; Barr, A.; Bortnik, J.; et al. Diurnal and Seasonal Dynamics of Solar-Induced Chlorophyll Fluorescence, Vegetation Indices, and Gross Primary Productivity in the Boreal Forest. *J. Geophys. Res. Biog.* **2022**, *127*. [CrossRef]

35. Tramontana, G.; Ichii, K.; Camps-Valls, G.; Tomelleri, E.; Papale, D. Uncertainty analysis of gross primary production upscaling using Random Forests, remote sensing and eddy covariance data. *Remote Sens. Environ.* **2015**, *168*, 360–373. [CrossRef]

36. Sun, Y.; Frankenberg, C.; Wood, J.D.; Schimel, D.S.; Jung, M.; Guanter, L.; Drewry, D.T.; Verma, M.; Porcar-Castell, A.; Griffis, T.J.; et al. OCO-2 advances photosynthesis observation from space via solar-induced chlorophyll fluorescence. *Science* **2017**, *358*, eaam5747. [CrossRef]

37. Chen, A.; Mao, J.; Ricciuto, D.; Xiao, J.; Frankenberg, C.; Li, X.; Thornton, P.E.; Gu, L.; Knapp, A.K. Moisture availability mediates the relationship between terrestrial gross primary production and solar-induced chlorophyll fluorescence: Insights from global-scale variations. *Glob. Chang. Biol.* **2021**, *27*, 1144–1156. [CrossRef]

38. Chen, A.; Mao, J.; Ricciuto, D.; Lu, D.; Xiao, J.; Li, X.; Thornton, P.E.; Knapp, A.K. Seasonal changes in GPP/SIF ratios and their climatic determinants across the Northern Hemisphere. *Glob. Chang. Biol.* **2021**, *27*, 5186–5197. [CrossRef] [PubMed]

39. Zhang, Z.; Zhang, Y.; Chen, J.M.; Ju, W.; Migliavacca, M.; El-Madany, T.S. Sensitivity of Estimated Total Canopy SIF Emission to Remotely Sensed LAI and BRDF Products. *J. Remote Sens.* **2021**, *2021*, 9795837. [CrossRef]

40. Diaz, S.; Kattge, J.; Cornelissen, J.H.; Wright, I.J.; Lavorel, S.; Dray, S.; Reu, B.; Kleyer, M.; Wirth, C.; Prentice, I.C.; et al. The global spectrum of plant form and function. *Nature* **2016**, *529*, 167–171. [CrossRef]

41. Migliavacca, M.; Musavi, T.; Mahecha, M.D.; Nelson, J.A.; Knauer, J.; Baldocchi, D.D.; Perez-Priego, O.; Christiansen, R.; Peters, J.; Anderson, K.; et al. The three major axes of terrestrial ecosystem function. *Nature* **2021**, *598*, 468–472. [CrossRef] [PubMed]

42. Reich, P.B.; Walters, M.B.; Ellsworth, D.S. From tropics to tundra: Global convergence in plant functioning. *Proc. Natl. Acad. Sci. USA* **1997**, *94*, 13730–13734. [CrossRef] [PubMed]

43. Wright, I.J.; Reich, P.B.; Westoby, M.; Ackerly, D.D.; Baruch, Z.; Bongers, F.; Cavender-Bares, J.; Chapin, T.; Cornelissen, J.H.C.; Diemer, M.; et al. The worldwide leaf economics spectrum. *Nature* **2004**, *428*, 821–827. [CrossRef] [PubMed]

44. Chang, Q.; Xiao, X.; Doughty, R.; Wu, X.; Jiao, W.; Qin, Y. Assessing variability of optimum air temperature for photosynthesis across site-years, sites and biomes and their effects on photosynthesis estimation. *Agric. For. Meteorol.* **2021**, *298–299*, 108277. [CrossRef]

45. Chen, N.; Song, C.; Xu, X.; Wang, X.; Cong, N.; Jiang, P.; Zu, J.; Sun, L.; Song, Y.; Zuo, Y.; et al. Divergent impacts of atmospheric water demand on gross primary productivity in three typical ecosystems in China. *Agric. For. Meteorol.* **2021**, *307*, 108527. [CrossRef]

46. 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]

47. Li, X.; Xiao, J. A Global, 0.05-Degree Product of Solar-Induced Chlorophyll Fluorescence Derived from OCO-2, MODIS, and Reanalysis Data. *Remote Sens.* **2019**, *11*, 517. [CrossRef]

48. Yuan, H.; Dai, Y.; Xiao, Z.; Ji, D.; Shangguan, W. Reprocessing the MODIS Leaf Area Index products for land surface and climate modelling. *Remote Sens. Environ.* **2011**, *115*, 1171–1187. [CrossRef]

49. 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]

50. Kunstler, G.; Falster, D.; Coomes, D.A.; Hui, F.; Kooyman, R.M.; Laughlin, D.C.; Poorter, L.; Vanderwel, M.; Vieilledent, G.; Wright, S.J.; et al. Plant functional traits have globally consistent effects on competition. *Nature* **2016**, *529*, 204–207. [CrossRef]

51. Madani, N.; Kimball, J.S.; Ballantyne, A.P.; Affleck, D.L.R.; van Bodegom, P.M.; Reich, P.B.; Kattge, J.; Sala, A.; Nazeri, M.; Jones, M.O.; et al. Future global productivity will be affected by plant trait response to climate. *Sci. Rep.* **2018**, *8*, 2870. [CrossRef] [PubMed]

52. Yin, Q.; Wang, L.; Lei, M.; Dang, H.; Quan, J.; Tian, T.; Chai, Y.; Yue, M. The relationships between leaf economics and hydraulic traits of woody plants depend on water availability. *Sci. Total Environ.* **2018**, *621*, 245–252. [CrossRef] [PubMed]

53. Kattge, J.; Bönisch, G.; Díaz, S.; Lavorel, S.; Prentice, I.C.; Leadley, P.; Tautenhahn, S.; Werner, G.D.A.; Aakala, T.; Abedi, M.; et al. TRY plant trait database–enhanced coverage and open access. *Glob. Chang. Biol.* **2019**, *26*, 119–188. [CrossRef] [PubMed]

54. Butler, E.E.; Datta, A.; Flores-Moreno, H.; Chen, M.; Wythers, K.R.; Fazayeli, F.; Banerjee, A.; Atkin, O.K.; Kattge, J.; Amiaud, B.; et al. Mapping local and global variability in plant trait distributions. *Proc. Natl. Acad. Sci. USA* **2017**, *114*, E10937–E10946. [CrossRef]

55. Simard, M.; Pinto, N.; Fisher, J.B.; Baccini, A. Mapping forest canopy height globally with spaceborne lidar. *J. Geophys. Res.* **2011**, *116*. [CrossRef]

56. Buermann, W.; Forkel, M.; O'Sullivan, M.; Sitch, S.; Friedlingstein, P.; Haverd, V.; Jain, A.K.; Kato, E.; Kautz, M.; Lienert, S.; et al. Widespread seasonal compensation effects of spring warming on northern plant productivity. *Nature* **2018**, *562*, 110–114. [CrossRef]

57. Fu, Y.H.; Piao, S.; Zhao, H.; Jeong, S.J.; Wang, X.; Vitasse, Y.; Ciais, P.; Janssens, I.A. Unexpected role of winter precipitation in determining heat requirement for spring vegetation green-up at northern middle and high latitudes. *Glob. Chang. Biol.* **2014**, *20*, 3743–3755. [CrossRef] [PubMed]

58. Fu, Y.H.; Zhao, H.; Piao, S.; Peaucelle, M.; Peng, S.; Zhou, G.; Ciais, P.; Huang, M.; Menzel, A.; Penuelas, J.; et al. Declining global warming effects on the phenology of spring leaf unfolding. *Nature* **2015**, *526*, 104–107. [CrossRef]

59. Wu, D.; Zhao, X.; Liang, S.; Zhou, T.; Huang, K.; Tang, B.; Zhao, W. Time-lag effects of global vegetation responses to climate change. *Glob. Chang. Biol.* **2015**, *21*, 3520–3531. [CrossRef]

60. Loveland, T.R.; Reed, B.C.; Brown, J.F.; Ohlen, D.O.; Zhu, Z.; Yang, L.; Merchant, J.W. Development of a global land cover characteristics database and IGBP DISCover from 1 km AVHRR data. *Int. J. Remote Sens.* **2010**, *21*, 1303–1330. [CrossRef]

61. Keeling, C.; Whorf, T. Atmospheric CO2 records from sites in the SIO air sampling network. *Trends* **1994**, *93*, 16–26.

62. Li, X.; Xiao, J.F. Mapping Photosynthesis Solely from Solar-Induced Chlorophyll Fluorescence: A Global, Fine-Resolution Dataset of Gross Primary Production Derived from OCO-2. *Remote Sens.* **2019**, *11*, 2563. [CrossRef]

63. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]

64. Vincenzi, S.; Zucchetta, M.; Franzoi, P.; Pellizzato, M.; Pranovi, F.; De Leo, G.A.; Torricelli, P. Application of a Random Forest algorithm to predict spatial distribution of the potential yield of Ruditapes philippinarum in the Venice lagoon, Italy. *Ecol. Model.* **2011**, *222*, 1471–1478. [CrossRef]

65. Bishop, C.M. *Neural Networks for Pattern Recognition*; Oxford University Press: Oxford, UK, 1995.

66. Feng, Y.; Cui, N.; Hao, W.; Gao, L.; Gong, D. Estimation of soil temperature from meteorological data using different machine learning models. *Geoderma* **2019**, *338*, 67–77. [CrossRef]

67. Jahan, N.; Gan, T.Y. Modelling the vegetation–climate relationship in a boreal mixedwood forest of Alberta using normalized difference and enhanced vegetation indices. *Int. J. Remote Sens.* **2011**, *32*, 313–335. [CrossRef]

68. Panda, S.S.; Ames, D.P.; Panigrahi, S. Application of Vegetation Indices for Agricultural Crop Yield Prediction Using Neural Network Techniques. *Remote Sens.* **2010**, *2*, 673–696. [CrossRef]

69. Magney, T.S.; Barnes, M.L.; Yang, X. On the Covariation of Chlorophyll Fluorescence and Photosynthesis Across Scales. *Geophys. Res. Lett.* **2020**, *47*, e2020GL091098. [CrossRef]

70. Doughty, R.; Xiao, X.; Köhler, P.; Frankenberg, C.; Qin, Y.; Wu, X.; Ma, S.; Moore, B. Global-Scale Consistency of Spaceborne Vegetation Indices, Chlorophyll Fluorescence, and Photosynthesis. *J. Geophys. Res. Biog.* **2021**, *126*, e2020JG006136. [CrossRef]

71. Walther, S.; Voigt, M.; Thum, T.; Gonsamo, A.; Zhang, Y.; Kohler, P.; Jung, M.; Varlagin, A.; Guanter, L. Satellite chlorophyll fluorescence measurements reveal large-scale decoupling of photosynthesis and greenness dynamics in boreal evergreen forests. *Glob. Chang. Biol.* **2016**, *22*, 2979–2996. [CrossRef]

72. Zhang, J.; Xiao, J.; Tong, X.; Zhang, J.; Meng, P.; Li, J.; Liu, P.; Yu, P. NIRv and SIF better estimate phenology than NDVI and EVI: Effects of spring and autumn phenology on ecosystem production of planted forests. *Agric. For. Meteorol.* **2022**, *315*, 108819. [CrossRef]

73. Wang, C.; Beringer, J.; Hutley, L.B.; Cleverly, J.; Li, J.; Liu, Q.; Sun, Y. Phenology Dynamics of Dryland Ecosystems Along the North Australian Tropical Transect Revealed by Satellite Solar-Induced Chlorophyll Fluorescence. *Geophys. Res. Lett.* **2019**, *46*, 5294–5302. [CrossRef]

74. Li, X.; Xiao, J.; He, B.; Altaf Arain, M.; Beringer, J.; Desai, A.R.; Emmel, C.; Hollinger, D.Y.; Krasnova, A.; Mammarella, I.; et al. Solar-induced chlorophyll fluorescence is strongly correlated with terrestrial photosynthesis for a wide variety of biomes: First global analysis based on OCO-2 and flux tower observations. *Glob. Chang. Biol.* **2018**, *24*, 3990–4008. [CrossRef]

75. Du, S.; Liu, X.; Chen, J.; Liu, L. Prospects for Solar-Induced Chlorophyll Fluorescence Remote Sensing from the SIFIS Payload Onboard the TECIS-1 Satellite. *J. Remote Sens.* **2022**, *2022*, 9845432. [CrossRef]

76. Hinojo-Hinojo, C.; Goulden, M.L. Plant Traits Help Explain the Tight Relationship between Vegetation Indices and Gross Primary Production. *Remote Sens.* **2020**, *12*, 1405. [CrossRef]

77. Dai, Y.; Dickinson, R.E.; Wang, Y.-P. A Two-Big-Leaf Model for Canopy Temperature, Photosynthesis, and Stomatal Conductance. *J. Clim.* **2004**, *17*, 2281–2299. [CrossRef]

78. De Kauwe, M.G.; Keenan, T.F.; Medlyn, B.E.; Prentice, I.C.; Terrer, C. Satellite based estimates underestimate the effect of CO2 fertilization on net primary productivity. *Nat. Clim. Chang.* **2016**, *6*, 892–893. [CrossRef]

79. Jung, M.; Reichstein, M.; Schwalm, C.R.; Huntingford, C.; Sitch, S.; Ahlstrom, A.; Arneth, A.; Camps-Valls, G.; Ciais, P.; Friedlingstein, P.; et al. Compensatory water effects link yearly global land CO2 sink changes to temperature. *Nature* **2017**, *541*, 516–520. [CrossRef]

80. Keenan, T.F.; Prentice, I.C.; Canadell, J.G.; Williams, C.A.; Wang, H.; Raupach, M.; Collatz, G.J. Recent pause in the growth rate of atmospheric CO2 due to enhanced terrestrial carbon uptake. *Nat. Commun.* **2016**, *7*, 13428. [CrossRef] [PubMed]

81. Ryu, Y.; Berry, J.A.; Baldocchi, D.D. What is global photosynthesis? History, uncertainties and opportunities. *Remote Sens. Environ.* **2019**, *223*, 95–114. [CrossRef]

82. Zhu, Z.; Piao, S.; Myneni, R.B.; Huang, M.; Zeng, Z.; Canadell, J.G.; Ciais, P.; Sitch, S.; Friedlingstein, P.; Arneth, A.; et al. Greening of the Earth and its drivers. *Nat. Clim. Chang.* **2016**, *6*, 791–795. [CrossRef]

83. Sage, R.F.; Way, D.A.; Kubien, D.S. Rubisco, Rubisco activase, and global climate change. *J. Exp. Bot.* **2008**, *59*, 1581–1595. [CrossRef]

84. Niu, S.; Li, Z.; Xia, J.; Han, Y.; Wu, M.; Wan, S. Climatic warming changes plant photosynthesis and its temperature dependence in a temperate steppe of northern China. *Environ. Exp. Bot.* **2008**, *63*, 91–101. [CrossRef]

85. Kim, Y.; Kimball, J.S.; Didan, K.; Henebry, G.M. Response of vegetation growth and productivity to spring climate indicators in the conterminous United States derived from satellite remote sensing data fusion. *Agric. For. Meteorol.* **2014**, *194*, 132–143. [CrossRef]

86. Liu, J.; Wennberg, P.O.; Parazoo, N.C.; Yin, Y.; Frankenberg, C. Observational Constraints on the Response of High-Latitude Northern Forests to Warming. *AGU Adv.* **2020**, *1*, e2020AV000228. [CrossRef]

87. Piao, S.; Nan, H.; Huntingford, C.; Ciais, P.; Friedlingstein, P.; Sitch, S.; Peng, S.; Ahlstrom, A.; Canadell, J.G.; Cong, N.; et al. Evidence for a weakening relationship between interannual temperature variability and northern vegetation activity. *Nat. Commun.* **2014**, *5*, 5018. [CrossRef] [PubMed]

88. Xu, L.; Myneni, R.B.; Chapin Iii, F.S.; Callaghan, T.V.; Pinzon, J.E.; Tucker, C.J.; Zhu, Z.; Bi, J.; Ciais, P.; Tømmervik, H.; et al. Temperature and vegetation seasonality diminishment over northern lands. *Nat. Clim. Chang.* **2013**, *3*, 581–586. [CrossRef]

89. Dass, P.; Rawlins, M.A.; Kimball, J.S.; Kim, Y. Environmental controls on the increasing GPP of terrestrial vegetation across northern Eurasia. *Biogeosciences* **2016**, *13*, 45–62. [CrossRef]

90. Kramer, K.; Leinonen, I.; Loustau, D. The importance of phenology for the evaluation of impact of climate change on growth of boreal, temperate and Mediterranean forests ecosystems: An overview. *Int. J. Biometeor.* **2000**, *44*, 67–75. [CrossRef] [PubMed]

91. Piao, S.; Fang, J.; Zhou, L.; Ciais, P.; Zhu, B. Variations in satellite-derived phenology in China's temperate vegetation. *Glob. Chang. Biol.* **2006**, *12*, 672–685. [CrossRef]

92. Ren, P.; Liu, Z.; Zhou, X.; Peng, C.; Xiao, J.; Wang, S.; Li, X.; Li, P. Strong controls of daily minimum temperature on the autumn photosynthetic phenology of subtropical vegetation in China. *For. Ecosyst.* **2021**, *8*, 31. [CrossRef] [PubMed]

93. Zhou, S.; Zhang, Y.; Caylor, K.K.; Luo, Y.; Xiao, X.; Ciais, P.; Huang, Y.; Wang, G. Explaining inter-annual variability of gross primary productivity from plant phenology and physiology. *Agric. For. Meteorol.* **2016**, *226–227*, 246–256. [CrossRef]

94. Buermann, W.; Bikash, P.R.; Jung, M.; Burn, D.H.; Reichstein, M. Earlier springs decrease peak summer productivity in North American boreal forests. *Environ. Res. Lett.* **2013**, *8*, 024027. [CrossRef]

95. Liu, Q.; Fu, Y.H.; Zeng, Z.; Huang, M.; Li, X.; Piao, S. Temperature, precipitation, and insolation effects on autumn vegetation phenology in temperate China. *Glob. Chang. Biol.* **2016**, *22*, 644–655. [CrossRef] [PubMed]

96. Wu, G.; Guan, K.; Jiang, C.; Peng, B.; Kimm, H.; Chen, M.; Yang, X.; Wang, S.; Suyker, A.E.; Bernacchi, C.J.; et al. Radiance-based NIRv as a proxy for GPP of corn and soybean. *Environ. Res. Lett.* **2020**, *15*, 034009. [CrossRef]

97. Zeng, Y.; Badgley, G.; Dechant, B.; Ryu, Y.; Chen, M.; Berry, J.A. A practical approach for estimating the escape ratio of near-infrared solar-induced chlorophyll fluorescence. *Remote Sens. Environ.* **2019**, *232*, 111209. [CrossRef]

98. Ma, Y.; Liu, L.; Chen, R.; Du, S.; Liu, X. Generation of a Global Spatially Continuous TanSat Solar-Induced Chlorophyll Fluorescence Product by Considering the Impact of the Solar Radiation Intensity. *Remote Sens.* **2020**, *12*, 2167. [CrossRef]

99. Zhang, Y.; Joiner, J.; Alemohammad, S.H.; Zhou, S.; Gentine, P. A global spatially contiguous solar-induced fluorescence (CSIF) dataset using neural networks. *Biogeosciences* **2018**, *15*, 5779–5800. [CrossRef]

100. Badgley, G.; Anderegg, L.D.L.; Berry, J.A.; Field, C.B. Terrestrial gross primary production: Using NIRVto scale from site to globe. *Glob. Chang. Biol.* **2019**, *25*, 3731–3740. [CrossRef]

101. Wang, Y.P.; Lu, X.J.; Wright, I.J.; Dai, Y.J.; Rayner, P.J.; Reich, P.B. Correlations among leaf traits provide a significant constraint on the estimate of global gross primary production. *Geophys. Res. Lett.* **2012**, *39*, L19405. [CrossRef]

102. Musavi, T.; Migliavacca, M.; van de Weg, M.J.; Kattge, J.; Wohlfahrt, G.; van Bodegom, P.M.; Reichstein, M.; Bahn, M.; Carrara, A.; Domingues, T.F.; et al. Potential and limitations of inferring ecosystem photosynthetic capacity from leaf functional traits. *Ecol. Evol.* **2016**, *6*, 7352–7366. [CrossRef] [PubMed]

103. Reich, P.B. Key canopy traits drive forest productivity. *Proc. Biol. Sci.* **2012**, *279*, 2128–2134. [CrossRef] [PubMed]