Contents lists available at ScienceDirect

Science of the Total Environment

journal homepage: www.elsevier.com/locate/scitotenv

Improving global gross primary productivity estimation using two-leaf light use efficiency model by considering various environmental factors via machine learning

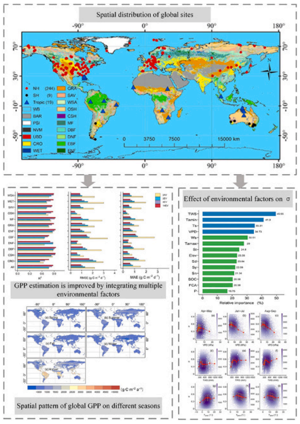
Zhilong Li[a,b], Ziti Jiao[a,b,c,*], Ge Gao[a,b], Jing Guo[a,b], Chenxia Wang[a,b], Sizhe Chen[a,b], Zheyou Tan[a,b], Wenyu Zhao[d]

[a] *State Key Laboratory of Remote Sensing Science, Beijing Normal University, Beijing 100875, China*
[b] *Institute of Remote Sensing Science and Engineering, Faculty of Geographical Science, Beijing Normal University, Beijing 100875, China*
[c] *Beijing Engineering Research Center for Global Land Remote Sensing Products, Beijing Normal University, Beijing 100875, China*
[d] *Key Laboratory of West China's Environment Systems (Ministry of Education), College of Earth and Environmental Sciences, Observation and Research Station on Eco-Environment of Frozen Ground in the Qilian Mountains, Lanzhou University, Lanzhou 730000, China*

HIGHLIGHTS

- A hybrid model is proposed to generate a novel long-term global GPP dataset.
- Random forest has a great advantage in integrating various environmental factors.
- Terrestrial water storage exerts the greatest control over photosynthesis.
- There is an optimal air temperature for photosynthesis.

GRAPHICAL ABSTRACT



ARTICLE INFO

Editor: Alessandra De Marco

Keywords:
GPP
Environmental stress factors
Terrestrial water storage
Hybrid model
Spatiotemporal patterns

ABSTRACT

Distinguishing gross primary productivity (GPP) into sunlit (GPP_{su}) and shaded (GPP_{sh}) components is critical for understanding the carbon exchange between the atmosphere and terrestrial ecosystems under climate change. Recently, the two-leaf light use efficiency (TL-LUE) model has proven effective for simulating global GPP_{su} and GPP_{sh}. However, no known physical method has focused on integrating the overall constraint of intricate environmental factors on photosynthetic capability, and seasonal differences in the foliage clumping index (CI), which most likely influences GPP estimation in LUE models. Here, we propose the TL-CRF model, which uses the random forest technique to integrate various environmental variables, particularly for terrestrial water storage (TWS), into the TL-LUE model. Moreover, we consider seasonal differences in CI at a global scale. Based on 267

[*] Corresponding author at: State Key Laboratory of Remote Sensing Science, Beijing Normal University, Beijing 100875, China.
E-mail address: jiaozt@bnu.edu.cn (Z. Jiao).

https://doi.org/10.1016/j.scitotenv.2024.176673
Received 12 June 2024; Received in revised form 24 August 2024; Accepted 30 September 2024
Available online 2 October 2024

global eddy covariance flux sites, we explored the functional response of vegetation photosynthesis to key environmental factors, and trained and evaluated the TL-CRF model. The TL-CRF model was then used to simulate global eight-day GPP, GPP_{su}, and GPP_{sh} from 2002 to 2020. The results show that the relative prediction error of environmental stress factors on the maximum LUE is reduced by approximately 52 % when these factors are integrated via the RF model. Thus the accuracy of global GPP estimation ($R^2 = 0.87$, RMSE = 0.94 g C m^{-2} d^{-1}, MAE = 0.61 g C m^{-2} d^{-1}) in the TL-CRF model is greater than that ($R^2 = 0.76$, RMSE = 2.18 g C m^{-2} d^{-1}, MAE = 1.50 g C m^{-2} d^{-1}) in the TL-LUE model, although this accuracy awaits further investigation among the released GPP products. TWS exerts the greatest control over ecosystem photosynthesis intensity, making it a suitable water indicator. Furthermore, the results confirm an optimal minimum air temperature for photosynthesis. Overall, these findings indicate a promising method for producing a new global GPP dataset, advancing our understanding of the dynamics and interactions between photosynthesis and environmental factors.

1. Introduction

Gross primary productivity (GPP) is defined as the total amount of carbon dioxide (CO_2) taken up by all leaves through vegetation photosynthesis over a unit of time (Chapin et al., 2002). GPP is a key variable in evaluating the terrestrial carbon cycle and measuring vegetation response to climate change (Beer et al., 2010). Therefore, accurately estimating GPP at different spatial and temporal scales is crucial for understanding the global carbon cycle, ecosystem functions, and climate changes (Anav et al., 2015; Badgley et al., 2019; Zhang et al., 2019).

Recently, numerous light use efficiency (LUE, ε) models have shown significant potential in studying the spatiotemporal dynamics of terrestrial GPP at global or regional scales (Zhao and Running, 2010; Dong et al., 2015; Zhang et al., 2015b; Yuan et al., 2019) because of the few required parameters, such as the maximum ε (ε_{max}), concise frameworks, easy to implementation (Dong et al., 2015; de Mattos et al., 2020), and use of digestible remote sensing data (Wu et al., 2010). GPP is strongly linked to the vegetation canopy structure (Cheng et al., 2015; Xu et al., 2019), which can be effectively depicted by the vegetation clumping index (CI). The two-leaf light use efficiency (TL-LUE) model (He et al., 2013) improves GPP estimation by separating the canopy into sunlit and shaded leaves on the basis of the CI and has been shown to outperform the MOD17 GPP model in various ecosystems (Zhou et al., 2016). In the TL-LUE model, canopy GPP can be decomposed into the GPP of sunlit (GPP_{su}) and shaded leaves (GPP_{sh}). Solar-induced chlorophyll fluorescence (SIF), which is highly related to GPP is emitted mainly from sunlit leaves (Lee et al., 2015; Pinto et al., 2016), so GPP_{su} is further used to refine the relationship between SIF and GPP. Additionally, the GPP_{su} is applied to retrieve several key photosynthetic parameters such as the maximum carboxylation velocity (Xie et al., 2018). Because shaded leaves can effectively absorb diffuse solar radiation, so the GPP_{sh} has a dominant field for vegetation under cloudy skies and dense vegetation (Han et al., 2020). Hence, the GPP_{sh} can be utilized to investigate the response of carbon uptake to changes in the fractions of solar radiation on a large scale. Accurate estimation of GPP_{su} and GPP_{sh} is essential for understanding the intrinsic mechanisms regulating vegetation photosynthesis processes.

Seasonal changes in CI reflect variations in the vegetation canopy (Pisek et al., 2015; He et al., 2016; Wei et al., 2019). With advancements in long-term global CI products, incorporating the temporal variation in CI into remote sensing and land surface simulations has become feasible (Fang, 2021). Our latest study proposed a TL-CLUE model that considers the seasonal differences in CI in the TL-LUE model (Li et al., 2023). As a result, the accuracy of GPP estimation across the North American continent is improved by approximately 9.76 % when TL-CLUE uses three CI estimations (Ω) from different seasons. Sunlit leaves absorb direct and diffuse radiation, whereas shaded leaves absorb only diffuse radiation (Rap et al., 2018). Nevertheless, GPP_{sh} increases with increasing in LAI and diffuse radiation (Chen et al., 2012b). Additionally, electron transport speed and the maximum carboxylation rate limit the photosynthesis rate of shaded and sunlit leaves, respectively (Chen et al., 1999; DE, 1997).

Key differences among LUE models lie in the calculation of canopy absorbed photosynthetically active radiation (APAR) and the scale of environmental stress factors on ε_{max} (σ) (He et al., 2013; Zhang et al., 2015b). Previous studies have reported that differences among five LUE models are likely caused by different calibrations of ε_{max} and descriptions of σ (Xie et al., 2020). Environmental stress factors are a primary source of uncertainty in GPP estimation. First, most LUE models, such as EC-LUE and MOD17, focus on only two or three, or five environmental factors (Wang et al., 2018b). Second, a single indicator or analytic formula is often insufficient to capture the complex responses of vegetation photosynthesis to environmental changes, particularly under extreme conditions. For example, the vapor pressure deficit, soil water content, and plant evaporate fraction explain only 20 %, 6 %, and 36 % of the monthly LUE variation, respectively (Zhang et al., 2015a). Third, most LUE models integrate environmental factors via either the multiplication principle or the law of minimum (LOM). Specifically, the multiple principle states that the total σ is the product of the σ of various environmental stress factors, and is used in most LUE models including the MOD17, VPM, and TL-LUE models. The LOM means that the final σ is the smallest value of σ among all environmental stress factors, which is adopted in the EC-LUE, CCW, and CFLUX models. These two methods usually fail to detect the coupling effects among various environmental factors, although they are relatively easy to manipulate. Moreover, there is spatial heterogeneity in the effects of the same environmental variable on vegetation photosynthesis. For example, air temperature can better explain GPP variation in northern cold sites (Reichstein et al., 2007) but is not a dominant in some tropical biomes, such as savannas and woody savannas (Ma et al., 2014). Finally, differences in the description and calculation of environmental stress factors among various LUE models also lead to errors.

The exploration the synergy of merging machine learning (ML) and process-based models in the modeling of earth system science is promising (Reichstein et al., 2019). To reduce uncertainties in GPP estimation caused by environmental stress factors, a comprehensive scale of diverse environmental stress factors to ε_{max} must be explored by leveraging advanced techniques. With the development of ML methods, various temporal variables of a model can be taken from vast datasets (Geer, 2021), potentially providing an integrated strategy for improving σ by incorporating various environmental stress factors. Combining the eddy covariance (EC) technique with ML methods offers a valuable opportunity to study ecosystem responses to climate change (Tramontana et al., 2020), even though EC has been widely used to measure carbon exchange between the atmosphere and ecosystems. ML methods can detect nonlinear relationships and interactions in complex data, which is used in to improve the understanding of the of environmental stress factors of the TL-LUE model. Recent studies have emphasized the potential of ML in optimizing parameters (Beucler et al., 2020; Han et al., 2020; Yuval and O'Gorman, 2020). Thus, coupling mechanistic and learning models represents a mutual complement between “rationalism” and “empiricism”, which has advanced scientific studies (Bergen et al., 2019). Hybrid models aim to efficiently access Earth observation data by replacing uncertain parameters and processes with ML techniques, while maintaining interpretability and physical consistency (Kraft et al., 2022). These models can improve the predictability of earth systems or components, helping to gradually eliminate errors caused by earth system models and artificial simplifications (Brenowitz and Bretherton, 2018).

In this study, we first update the TL-CLUE model on a global scale, and then employ the random forest (RF) technique to integrate the comprehensive control of multiple environmental factors on vegetation photosynthesis intensity and to further improve global GPP estimation by embedding the RF submodule into the TL-CLUE model (TL-CRF), which could realize complementary natural advantages. Before simulating the global GPP, the TL-CRF model is trained and evaluated on data from 267 global EC flux sites. Additionally, we explored the functional response of vegetation photosynthesis to vital environmental factors. The primary objectives of this study are: a) to reduce the uncertainty of σ estimation by integrating various environmental factors via the RF technique, b) to investigate the functional relationships between vegetation photosynthesis and key environmental factors, and c) to explore the temporal-spatial changes in global GPP, GPP_{su}, and GPP_{sh}. This study provides a reference for assessing global terrestrial productivity and provides our insights into the response of GPP to environmental changes.

2. Data

2.1. Eddy covariance flux data

Flux data derived from FLUXNET2015 (www.fluxdata.org) and AmeriFlux (https://ameriflux.lbl.gov/) have the standard formats of eddy covariance (EC) flux data, advanced quality control, and processing methods, which cover different global regions (Pastorello et al., 2020a). Unlike FLUXNET2015, which only covers from 2001 to 2014, AmeriFlux provides a longer period from 2000 to 2022. The recently published AmeriFlux FLUXNET provides high-quality flux data with an outlier filter, missing value filling, and uncertainty analysis. In the EC flux data, GPP is separated from NEE (net ecosystem exchange) by light-response curves (Reichstein et al., 2005). The data are divided into different groups with multiple quality labels ranging from 0 to 1 on the basis of the proportion of great data (observed data, high quality gap data) to the whole dataset, of which 30-min and eight-day data with quality labels greater than 0.8 are selected. Simultaneously, to reduce measurement errors, the study selected sites that can provide at least 90 % of the meteorological data and flux data as research sites. The consistency between the site descriptions of the vegetation and MODIS land cover (LC) type version 6.1 (MCD12Q1.061) data products was checked, and 267 EC flux sites were selected for the study. The eight-day data is aggregated from 30-minute data that are longer than 5 days to coincide with the temporal resolution of the remote sensing data. The analysis variables are selected from eight-day flux data, including GPP, solar shortwave radiation, air temperature, and vapor pressure deficit (VPD). According to the site description of vegetation and MODIS LC, the distribution of sites in different vegetation types is 52 for ENF (evergreen needleleaf forest), 34 for DBF (deciduous broadleaf forest), 11 for MF (mixed forest), 23 for OSH (open shrub), 6 for WSA (woody savannas), 50 for GRA (grassland), 9 for SAV (savannas), 32 for CRO (crop), 35 for WET (wetland), 7 for EBF (evergreen broadleaf forest), 2 for DNF (deciduous needleleaf forest), and 6 for CSH (close shrub).

2.2. MODIS data

The MOD15A2H.061 product (Myneni et al., 2021) provides an eight-day leaf area index (LAI) with a spatial resolution of 500 m from 2002 to 2020, which generally acts as an important LAI resource for LUE models. The vegetation canopy structure, represented by the LAI, is the key mediating factor for the response of GPP to environmental factors (Zheng et al., 2024). The measuring radius of the EC tower is approximately 20–100 times greater than the height of the EC tower and can range from tens of meters to tens of kilometers (Schmid, 1997; Chen et al., 2012a). Therefore, LAI data (2 × 2 pixels for 1 × 1 km) are used for sites approximately 10 m high for CRO, GRA, WSA, OSH, CSH, WET, and SAV, and LAI data (6 × 6 pixels for 3 × 3 km) are used for those higher than 10 m for ENF, DBF, EBF, DNF, and MF (Zhou et al., 2016). Savitzky-Golay Logistics (Savitzky and Golay, 1964) have been used for smoothing the temporal series of the LAI to weaken noise caused by clouds, fog, ground ice, snow, and atmospheric pollution, and the handling process is implemented via the “prospectr” function of the “savitzkyGolay” package in R (version 4.4.0). Additionally, the LAI is screened by data quality control to reduce the influence of low-quality data on GPP estimation.

The MODIS LC data (Friedl and Sulla-Menashe, 2022), with a spatial resolution of 500 m from 2001 to 2021, originate from the NASA Earth Observing System Data and Information System (EOSDIS), which is generated by an integrated decision tree (Friedl et al., 2010). The LC is classified into 17 types of land cover based on the International Geosphere Biosphere Program (IGBP), which abides by the vegetation of the EC tower site. Hence, the site vegetation is determined by combining the MODIS LC data and vegetation descriptions of the EC site. Additionally, the MODIS LC (MCD12Q1) is also used for simulating the global GPP.

The MODIS CI is retrieved from the MODIS bidirectional reflectance distribution function (BRDF) on the basis of the modified kernel-driven BRDF model (Jiao et al., 2016, 2018). There are significant seasonal differences in CI, which are greater in the leaf-off season than in the leaf-on season, particularly for DBF and MF (Yin et al., 2022), which have the potential to reflect changes in the vegetation canopy (Pisek et al., 2015; He et al., 2016; Wei et al., 2019). Ignoring these seasonal differences in CI would lead to errors in the estimations of GPP (Chen et al., 2012b). With the development of long-term global CI products, temporal differences in CI have been incorporated into studies involving remote sensing and land surface simulations (Fang, 2021).

2.3. Meteorological data

The hour data (GMAO, 2015a, 2015b, 2019) with a spatial resolution of 0.5 × 0.625° from 1980 to 2024 were provided by Modern Era Retrospective analysis for Research and Application, Version 2 (MERRA-2), which was developed by Goddard Earth Observing System Version 5 (GEOS-5) (Rienecker et al., 2011) and utilized by the Global Modeling and Assimilation Office (GMAO). MERRA-2 uses an enhanced assimilation system to reduce the uncertainty and passes the verification of site data, which has been used for the simulation of GPP (Wu et al., 2010; Wang et al., 2017). The increase in GPP is driven mainly by climate change (Xu et al., 2022). The following variables for the study are selected from MERRA-2: surface incoming shortwave flux (SSF), total cloud area fraction (FCA), wind speed (W_s), minimum air temperature (T_{amin}), mean air temperature (T_a), maximum air temperature (T_{amax}), dew point temperature (T_d) at a height of 2 m, soil temperature at layer 1 (T_s), precipitation (P), and terrestrial water storage (TWS). TWS not only contains water content from soil moisture, but also incorporates underground, land surface, snow, and biome water (Humphrey et al., 2018), which is a critical component of the global water and energy balance. These variables are aggregated into an eight-day scale. The VPD is calculated according to the dew points T_d and T_a as follows:

$$SVP = 6.112 \times e^{\frac{17.67 \times Ta}{243.5 + Ta}}$$

$$RH = e^{\frac{17.625 \times Td}{243.04 + Td} - \frac{17.625 \times Ta}{243.04 + Ta}}$$

$$VPD = SVP \times (1 - RH)$$

SVP is the saturated vapor pressure (hPa), and *RH* is the relative humidity (Zheng et al., 2020).

2.4. Soil properties

The SoilGrids (Hengl et al., 2017) system provides various global soil property data with a spatial resolution of 250 × 250 m at seven layers of 0, 5, 15, 30, 60, 100, and 200 cm, including soil organic carbon (SOC), soil bulk density (S_d), cation exchange capacity (CEC), pH, and soil texture (%) such as sand, silt, and clay, of which SOC, S_d, soil nitrogen (S_n), soil silt (S_t), and soil clay (S_y) are selected and aggregated to depths of 0–30 cm.

2.5. USGS digital elevation model

The GMTED2010 elevation dataset (https://earthexplorer.usgs.gov/) with three spatial resolutions of 30, 15, and 7.5″ aggregated the best global elevation (Elev) data, which were derived from a corporation between the United States Geographical Survey (USGS) and the National Geographical-Intelligence Agency (NGA). The GMTED2010 is more accurate than the previous GTOPO30 dataset and is widely used in vegetation growth, hydrological process, cartography, and meteorological analyses.

3. Model and methods

3.1. TL-LUE model

The TL-LUE model, which is based on the algorithm from the MOD17 GPP model, separates the APAR of the vegetation canopy into the APAR of sunlit and shaded leaves via CI. The fundamental equation is as follows:

$$GPP = f(PAR, LAI, \theta, \beta, \alpha, \Omega, \varepsilon_{msu}, \varepsilon_{msh}) \times \sigma \tag{1}$$

The estimation of GPP specifically is as follows:

$$GPP = (\varepsilon_{msu} \times APAR_{su} + \varepsilon_{msh} \times APAR_{sh}) \times \sigma \tag{2}$$

where ε_{msu} and ε_{msh} are the ε_{max} values of sunlit and shaded leaves, respectively, $APAR_{su}$, $APAR_{sh}$ is the $APAR$ (MJ m^{-2} d^{-1}) of sunlit and shaded leaves, respectively, and σ is the scale of the environmental stress factors on ε_{max}. These parameters can be calculated as follows:

$$APAR_{sh} = (1-\alpha) \times \left[\frac{PAR_{dif} - PAR_{difu}}{LAI} + C\right] \times LAI_{sh} \tag{3}$$

$$APAR_{su} = (1-\alpha) \times \left[PAR_{dir} \times \frac{\cos(\beta)}{\cos(\theta)} + \frac{PAR_{dif} - PAR_{difu}}{LAI} + C\right] \times LAI_{su} \tag{4}$$

where α is the canopy albedo, which depends on the vegetation type; β is the sun-leaf angle in the spherical canopy, which is usually set to 60°, θ is the solar zenith angle; PAR_{dif} and PAR_{dir} are the diffuse, and direction partition of PAR, respectively; PAR_{difu} is the scattering PAR of the canopy (Chen et al., 1999); C is the multiple scattering of PAR inside the canopy; LAI_{sh} and LAI_{su} are LAI of shaded and sunlit leaves (Chen et al., 1999), which is expressed as follows:

$$LAI_{su} = 2 \times \cos(\theta) \times \left(1 - e^{\frac{-0.5 \times \Omega \times LAI}{\cos(\theta)}}\right) \tag{5}$$

$$LAI_{sh} = LAI - LAI_{su} \tag{6}$$

where Ω is the estimation of the foliage CI.

The parameter σ in Eq. (1) can be calculated as follows:

$$\sigma = g(T_{amin}) \times h(VPD) \tag{7}$$

$$g(T_{amin}) = \left\{\begin{array}{ll} 0 & T_{amin} \le T_{amin\_min} \\ \dfrac{T_{amin} - T_{amin\_min}}{T_{amin\_max} - T_{amin\_min}} & T_{amin\_min} < T_{amin} < T_{amin\_max} \\ 1 & T_{amin} \ge T_{amin\_max} \end{array}\right\} \tag{8}$$

$$h(VPD) = \left\{\begin{array}{ll} 0 & VPD_{max} \le VPD \\ \dfrac{VPD_{max} - VPD}{VPD_{max} - VPD_{min}} & VPD_{min} < VPD < VPD_{max} \\ 1 & VPD_{min} \ge VPD \end{array}\right\} \tag{9}$$

where T_{amin} is the lowest daily air temperature. T_{amin_min}, VPD_{min}, T_{amin_max}, and VPD_{max} are specific parameters highly correlated with vegetation type (Zhou et al., 2016). These parameters of the TL-LUE model have already been calculated and verified on the basis of site data from global flux networks (Wu et al., 2015; Zhou et al., 2016).

3.2. Seasonal CI estimations on a global scale

The CI is a crucial vegetation structure parameter for distinguishing sunlit and shaded leaves, and thus accurately estimates canopy-scale GPP in the TL-LUE model (Chen et al., 2003). Furthermore, the seasonal differences in CI are potentially useful for mirroring the variation in the vegetation canopy in different leaf growth seasons (He et al., 2016). To investigate the effectiveness of the seasonal variation in CI for integrating various environmental stress factors on the basis of the RF model, the seasonal CI differences on a global scale are estimated on the basis of this latest research (Li et al., 2023), further updating the TL-CLUE model to a global scale by dividing the one-year leaf life cycle into different seasons for three latitudinal zones and averaging the corresponding seasonal Ω. Specifically, in this study, the northern hemisphere (NH) is the region north of 30°N excluding Greenland; the southern hemisphere (SH) is the region south of −30°N excluding Antarctica; and the tropics (Trop) is the region between the NH and SH (−30°N ~ 30°N). In the NH and SH, the one-year cycle of leaf growth is divided into leaf-off (LOS), leaf-scattered (LSS), and leaf-gathered (LGS) seasons on the basis of four phenological thresholds from the MODIS land surface phenology (LSP) product (MCD12Q2 V061). For Trop, one year is viewed as one entire cycle of leaf life because there are no obvious seasonal differences in CI (Yin et al., 2022). Then, the Ω values of the corresponding seasons for the twelve global vegetation types in the three regions are determined according to previous work (Li et al., 2023).

3.3. ML algorithm

ML models, such as RF, can resolve the complex relationships among various variables through nonparametric, nonlinear, and flexible associations, despite the lack of explicit causation in mechanistic models. The ML method is powerful for precisely recognizing complex relationships and patterns in multiple variable datasets (Kraft et al., 2020), which provides a way to reduce the uncertainties caused by various environmental factors in LUE models. Among them, RF (Breiman, 2001; Cutler and Zhao, 2001) combines CART decision tree construction and the bagging integration algorithm, randomly selects features of the nodes on the basis of the parallel combination of decision trees and uses voting to determine the best model, which is a complex and strong classifier (Surhone et al., 2010) consisting of multiple weak classifiers (decision trees). The RF model not only is easy to run, has a computation cost, is robust to outliers, and effectively avoids overfitting, but also simulates the complex relationships among numerous driving variables fewer parameters than mechanism models (Breiman, 2001), which have been successfully used in the remote sensing community (Tramontana et al., 2016). The RF model can characterize the temporal changes and magnitude of GPP well, particularly for MF and ENF (Chang et al., 2023). The study also considered other ML models such as artificial neural networks (ANNs), support vector machines (SVMs), and gradient-boosted regression trees (GBRTs). In this study, on the basis of the global EC flux sites, these ML models for each vegetation type were trained and verified via tenfold cross-validation by the coefficient of determination (R^2), root mean square error (RMSE), and mean absolute error (MAE). As a result, the RF model outperforms the other models; thus, it is selected for the study.

3.4. Design of a hybrid model

The objective of this study is to reduce the uncertainty of global GPP estimation by integrating diverse environmental stress factors via RF technique after considering the seasonal differences in CI, which mainly includes five parts: 1) the TL-CLUE model is updated to a global scale according to our latest published work (Li et al., 2023); 2) analysis of temporal variation and uncertainty in the σ for different vegetation types; 3) improving the σ estimation by integrating multiple environmental stress factors via the RF technique, and further improving global GPP estimation by embedding the RF submodule into the TL-CLUE; 4) the ecological response of vegetation photosynthesis to key environmental factors; and 5) the spatial and temporal patterns of global GPP, GPP_{su}, and GPP_{sh} from 2002 to 2020 based on the TL-CRF model.

The relationships between numerous environmental factors and GPP are generally nonlinear (Bao et al., 2022). Multiple environmental stress factors constrain the ε_{max}, but the TL-LUE model considers only the VPD and daily minimum air temperature, which undoubtedly leads to uncertainty in GPP estimation. Climate variables play a critical role in global GPP prediction (Lu et al., 2024). The estimation of σ is expected to be improved by integrating various environmental factors, including meteorology, vegetation, topography, and soil properties via the RF model. On the basis of 70 % random selection of the global 267 EC flux site, in theory, the real σ value can be derived via the TL-CLUE model, and the calculation process is implemented in R platform (version 4.3.0). The RF model (Eq. 10) is trained to describe the quantitative relationship between σ and many environmental factors including TWS, VPD, T_{amin}, T_s, W_s, T_{amax}, FCA, S_t, S_n, Elev, S_d, S_y, SOC, and P. As a result, this study intends to improve global GPP estimation (Eq. (11)). Embedding this RF submodule into the TL-CLUE framework (TL-CRF) can also be performed on R platform (version 4.3.0). Moreover, the relative importance of each explanatory variable to σ is qualified by the increase in the percentage of mean squared error (%IncMSE) based on the RF algorithm (Breiman, 2001; Chipman et al., 1998).

$$\sigma = r\left(TWS, T_s, VPD, T_{amin}, T_{amax}, W_s, P, FCA, Elev, S_t, S_n, S_d, S_y, SOC\right) \quad (10)$$

$$GPP = f(\Omega_i, PAR, LAI, \theta, \beta, \alpha, \varepsilon_{msu}, \varepsilon_{msh}) \times \sigma \quad (11)$$

where Ω_i is Ω in the ith season for each vegetation type.

3.5. Model evaluation and statistical analysis

On the basis of a test dataset (30 % of the 267 global EC flux sites), global σ and GPP estimations for different vegetation types were assessed via R^2 (Eq. (12)), RMSE (Eq. (13)), MAE (Eq. (14)), and RPE (relative predictive errors) (Eq. (15)). The effects of seasonal differences in CI on global GPP estimations were evaluated against 267 global EC sites via R^2 (Eq. (12)), RMSE (Eq. (13)), Bias (Eq. (16)), and relative to the mean predictive bias (R_{bias}). The consistency between the estimated and the observed values was interpreted via linear regression at the eight-day, monthly, and annual scales. Eight-day GPP anomalies were calculated by subtracting the monthly mean from the eight-day value. To ensure that the RF model is more effective than the analytic formula is, the scales of VPD and T_{amin} to ε_{max} are calculated via the RF algorithm and the analytic formula from the TL-LUE model, respectively. To balance running speed and accuracy, the TL-CRF model is driven by four main explanatory variables (4EV) and fourteen explanatory variables (14EV).

To explore the ecological significance of the driving variables used in the RF model, the functional relationships between key environmental factors and vegetation photosynthesis were analyzed via the piecewise averaging method (Liu et al., 2020) for each two consecutive months during the growing season (Tramontana et al., 2020). Specifically, the VPD, TWS, T_{amin}, and σ are all divided into ten equal subsets according to ten percentile thresholds, such as the 10th, 10th, ……, and 90th percentile, respectively. For each subset of TWS, T_{amin}, and VPD are decoupled so that their effects on σ can be separated. The functional relationships between the σ and VPD for the two months are then explored on the condition that the TWS value is limited to its 50th ~ 60th percentile. Analogously, the functional relationships between σ and TWS and T_{amin} for each double month are investigated when the VPD is restricted to the 50th ~ 60th percentiles. The above statistical processes are implemented in R the platform (version 4.3.0).

$$R^2 = \frac{\sum_{i=1}^{n}\left(v_{mi} - \overline{v_o}\right)^2}{\sum_{i=1}^{n}\left(v_{oi} - \overline{v_o}\right)^2} \quad (12)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}\left(v_{oi} - v_{mi}\right)^2}{\mathrm{n}}} \quad (13)$$

$$MAE = \frac{\sum_{i=1}^{n}\left|v_{oi} - v_{mi}\right|}{\mathrm{n}} \quad (14)$$

$$RPE = \frac{RMSE}{\overline{v_o}} \quad (15)$$

$$Bias = \frac{\sum_{i=1}^{n} v_{mi} - v_{oi}}{\mathrm{n}} \quad (16)$$

where v_o and v_m are the observed values and simulated values, respectively. Where *i* and *n* represent the ordinal value, and size of the resample, respectively.

3.6. Global GPP simulations from 2002 to 2020

Currently, a spatial resolution of 0.05 × 0.05° is widely used in numerous global or regional carbon cycle studies (Bi et al., 2022; Chen et al., 2021; Li and Xiao, 2019). Additionally, this spatial resolution is at a medium level among all the inputs in the study, realizing a perfect balance between the estimated accuracy and computational efficiency. Hence, this study simulated global GPP, GPP_{su}, and GPP_{sh} with a spatial resolution of 0.05 × 0.05° from 2002 to 2020 on the TL-CRF model driven by various environmental variables including SSF, LAI, LC, TWS, VPD, T_{amin}, T_s, W_s, $T_{amax,}$ FCA, S_t, S_n, Elev, S_d, S_y, SOC, and P, which have different data sources leading to different spatial resolutions and geographic coordinate systems. Therefore, the spatial data of these variables must be unified to the same geographic projection (WGS84) and resampled to the same spatial resolution of 0.05 × 0.05°. The LC was resampled to 0.05° × 0.05° via the nearest neighbor technique in ArcGIS (version 10.6.0). Additionally, we used the modified cosine interpolation function (Zhao et al., 2006) to spatially interpolate Elev, LAI, meteorological variables (temperature, precipitation, etc.), and soil properties (soil organic carbon, soil silt, etc.) into 0.05° × 0.05°, which was implemented in MATLAB (R2020a). The modified cosine function improves the accuracy of the meteorological inputs. In theory, meteorological variables from one edge to adjacent edges are impossible to abruptly change. These global spatial data are aggregated on an eight-day scale. This study explores the seasonal changes and spatial patterns of global GPP, GPP_{su}, and GPP_{sh}.

4. Results

4.1. Contribution of the seasonal CI to global GPP estimation

The uncertainties in global GPP estimation are reduced by the TL-CLUE model, with lower RMSE, bias, and R_{bias} values (Table 1). Compared with the TL-LUE model, the R_{bias} in the TL-CLUE model is reduced by approximately 9.2 % for the NH with three Ωs, 4.6 % for Trop with one Ω, and 11.1 % for SH with three Ωs (Table 1), which can be attributed to the fact that three Ωs from different seasons more accurately reflect the seasonal changes in the vegetation canopy during different leaf growth stages. The bias of GPP estimation in the TL-CLUE model is lower than that in the TL-LUE model in the NH, with the predicted values aligning more closely with the observed values. The RMSE of the TL-CLUE model is approximately 1.2–1.6 g C m^{-2} d^{-1} for DNF, OSH, SAV, and CSH, 2.3–3.1 g C m^{-2} d^{-1} for WSA, ENF, GRA, and WET, and 3.8–4.5 g C m^{-2} d^{-1} for CRO, MF, EBF, and DBF in the NH (Table 1), with the maximum occurring in DBF and the minimum occurring in DNF. There are differences in the reduction in uncertainties in GPP estimation among various vegetation types. Overall, the global GPP estimation is improved by approximately 8.3 % (R_{bias}) in the TL-CLUE model when three Ωs from three leaf growth stages are incorporated.

Compared with that of the TL-LUE model, the regression line in the TL-CLUE model is closer to the 1:1 theoretical line in the Northern Hemisphere, Tropics, and Southern Hemisphere (Fig. 1). Compared with that of the TL-LUE model, the RMSE of the TL-CLUE model is reduced by approximately 8.6 % for NH, 6 % for Trop, and 10.1 % for SH (Fig. 1), indicating that the fluctuation in GPP estimation can be mitigated by accounting for seasonal differences in CI. The bias in the TL-CLUE model is approximately 0.28 g C m^{-2} d^{-1} lower than that in the TL-LUE model, and the slope is also closer to 1 (Fig. 1). The uncertainties of the global eight-day GPP estimation are reduced in the TL-CLUE model by incorporating the seasonal differences in the CI.

4.2. Seasonal cycles of σ and its estimation evaluations

4.2.1. Mean seasonal cycles of eight-day σ

σ exhibited obvious seasonal changes that were greater in the leaf-on season than the leaf-off season, particularly for arbor vegetation such as DNF, ENF, MF, and DBF (Fig. 2). Notably, there is great uncertainty in the σ value measured by the SD for various vegetation types (Fig. 2). Compared with ENF, DBF, MF, and DNF, there is a frequent partial fluctuation in the entire seasonal cycle of σ in OSH, CSH, WSA, and SAV, although the range of σ is narrower. Similar seasonal magnitudes and cycles of σ are observed for WSA and SAV (Fig, 2). In the leaf-off season, the σ with uncertainty is approximately 0.18–0.23 (0.25) for MF and ENF; 0.28–0.32 (0.15–0.26) for CRO, DBF, and SAV; 0.34 (0.28–0.43) for CSH, GRA, and WET; and 0.43–0.68 (0.19–0.28) for EBF, OSH, CSH and WSA (Fig. 2). The differences in the σ between the leaf-off season and leaf-on season are approximately 0.55 for DBF, 0.56 for ENF, 0.60 for MF, 0.13 for OSH, 0.45 for CRO, 0.31 for GRA, 0.01 for WSA, 0.34 for SAV, 0.33 for WET, 0.21 for EBF, and 0.36 for CSH (Fig. 2), which indicates that there is a significant seasonal difference in the σ for most vegetation types. Additionally, in the leaf-on season, the σ with the corresponding uncertainty is approximately 0.84 (0.13) for DBF, 0.74 (0.17) for ENF, 0.82 (0.11) for MF, 0.62 (0.25) for OSH, 0.77 (0.28) for CRO, 0.66 (0.23) for GRA, 0.57 (0.22) for WSA, 0.62 (0.14) for SAV, 0.67 (0.29) for WET, 0.89 (0.09) for EBF, 0.80 (0.18) for CSH and 0.67 (0.11) for DNF (Fig. 2). The results show great uncertainty in σ, which likely leads to a large error in the GPP estimation in the TL-CLUE model.

4.2.2. Accuracy of σ estimated via the RF algorithm

The RF model notably improves the precision of the σ estimation, which is particularly evident under the 14EV setting. Comparative

Table 1
Comparison of eight-day GPP estimation between the TL-CLUE and TL-LUE models for different vegetation types in the NH, Trop, and SH, respectively. The TL-CLUE model improves the overall accuracy of the GPP estimation by approximately 9.18 for NH, 4.61 for Trop, and 11.07 % (R_{bias}) for SH by incorporating the seasonal differences in CI.

Region	IGBP	TL-CLUE			TL-LUE			Difference		
		RMSE	Bias	R_{bias} (%)	RMSE	Bias	R_{bias}	RMSE	Bias	R_{bias}
NH	CRO	3.79	1.13	29.67	4.03	1.48	38.78	−0.24	−0.35	−9.11
	DBF	4.49	2.52	58.88	4.87	2.85	66.71	−0.38	−0.33	−7.83
	ENF	2.86	0.83	21.97	3.16	1.07	28.45	−0.30	−0.24	−6.48
	MF	3.82	1.56	39.05	4.12	1.81	45.40	−0.30	−0.25	−6.34
	GRA	3.12	1.21	43.95	3.48	1.50	54.31	−0.36	−0.29	−10.35
	OSH	1.26	0.38	40.70	1.37	0.43	46.21	−0.11	−0.05	−5.50
	WSA	2.27	1.47	85.92	2.64	1.76	103.21	−0.38	−0.29	−17.29
	SAV	1.35	0.84	56.06	1.56	1.00	66.27	−0.20	−0.15	−10.21
	CSH	1.61	0.06	2.43	1.73	0.22	9.03	−0.12	−0.16	−6.59
	WET	3.57	1.75	61.80	3.96	2.08	73.61	−0.39	−0.33	−11.82
	EBF	4.17	2.53	59.76	4.96	3.21	75.72	−0.79	−0.68	−15.95
	DNF	1.21	0.72	30.83	1.26	0.78	33.47	−0.05	−0.06	−2.64
	All	2.79	1.25	44.25	3.10	1.52	53.43	−0.30	−0.27	−9.18
Trop	DBF	1.74	0.96	21.40	1.93	1.19	26.77	−0.19	−0.24	−5.36
	ENF	1.96	0.73	21.20	2.21	1.07	30.87	−0.25	−0.33	−9.68
	GRA	2.40	1.75	56.51	2.58	1.91	61.70	−0.18	−0.16	−5.19
	WSA	2.99	2.34	42.75	3.26	2.66	48.49	−0.27	−0.31	−5.73
	SAV	2.44	1.55	51.73	2.60	1.72	57.32	−0.16	−0.17	−5.59
	CSH	1.84	0.33	6.53	1.86	0.39	7.53	−0.02	−0.05	−1.00
	WET	6.61	0.76	10.17	6.74	0.96	12.93	−0.12	−0.21	−2.76
	EBF	5.87	5.20	63.10	6.23	5.61	68.09	−0.36	−0.41	−4.99
	DNF	1.27	−0.66	−25.16	1.25	−0.63	−23.98	0.02	−0.03	−1.18
	All	3.01	1.44	27.58	3.18	1.65	32.19	−0.17	−0.21	−4.61
SH	MF	5.53	−4.99	−63.63	5.37	−4.82	−61.45	0.16	−0.17	−2.18
	GRA	2.51	1.74	71.39	2.75	1.96	80.08	−0.24	−0.21	−8.68
	WSA	3.06	2.56	82.12	3.41	2.78	89.34	−0.35	−0.22	−7.22
	SAV	1.99	1.71	128.31	2.38	2.05	154.15	−0.39	−0.34	−25.84
	WET	2.93	2.16	111.80	3.29	2.45	126.95	−0.36	−0.29	−15.15
	EBF	5.03	2.80	37.80	5.65	3.34	45.17	−0.62	−0.54	−7.37
	All	3.51	1.00	61.30	3.81	1.29	72.37	−0.30	−0.30	−11.07

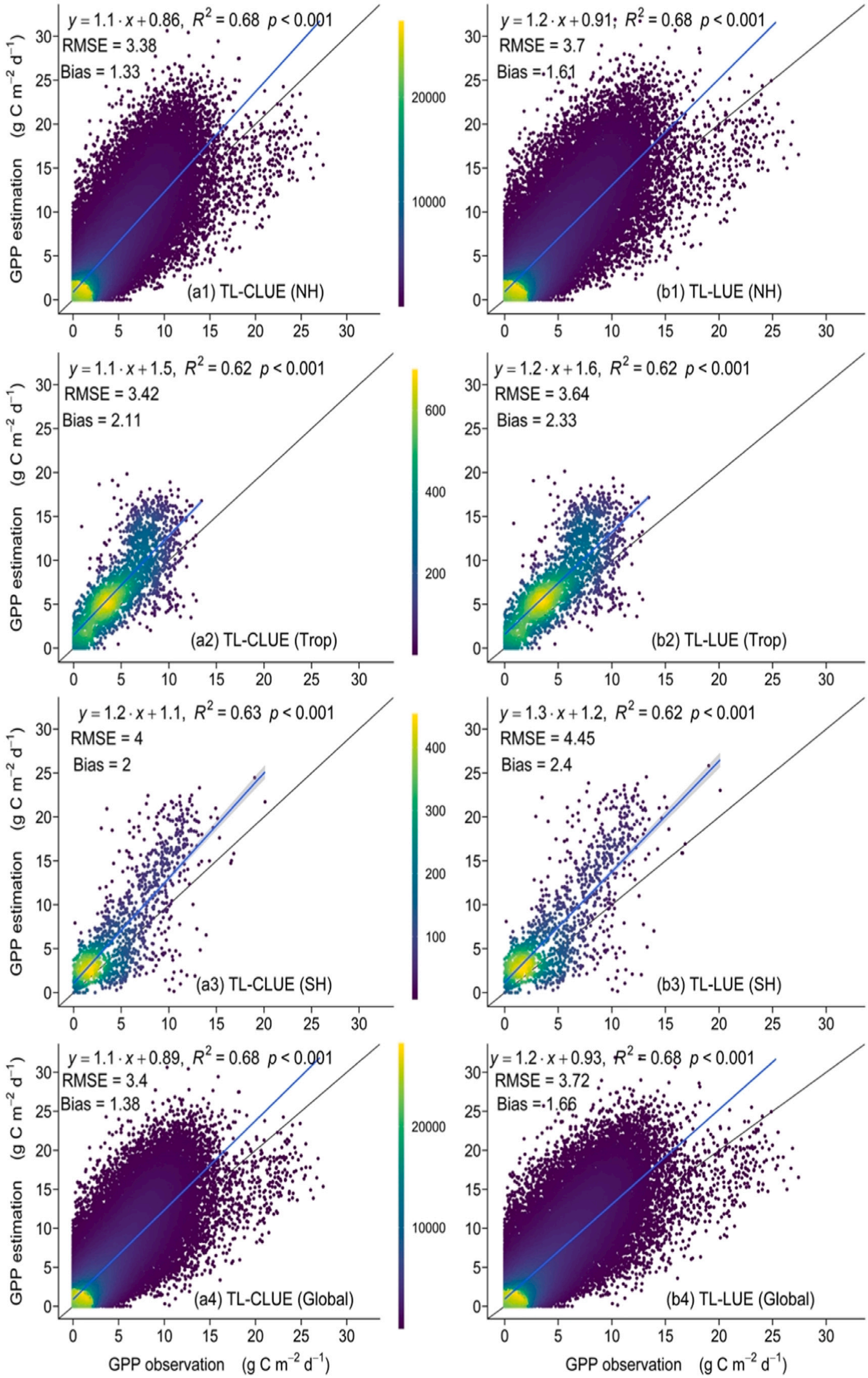


Fig. 1. Consistency between the GPP observations and GPP estimations in the TL-CLUE (a1–a4) and TL-LUE (b1–b4) models in the NH (a1, b1), Trop (a2, b2), SH (a3, b3), and global (a4, b4)regions. The blue line is a linear regression line and the black line represents the 1:1 theoretical line.

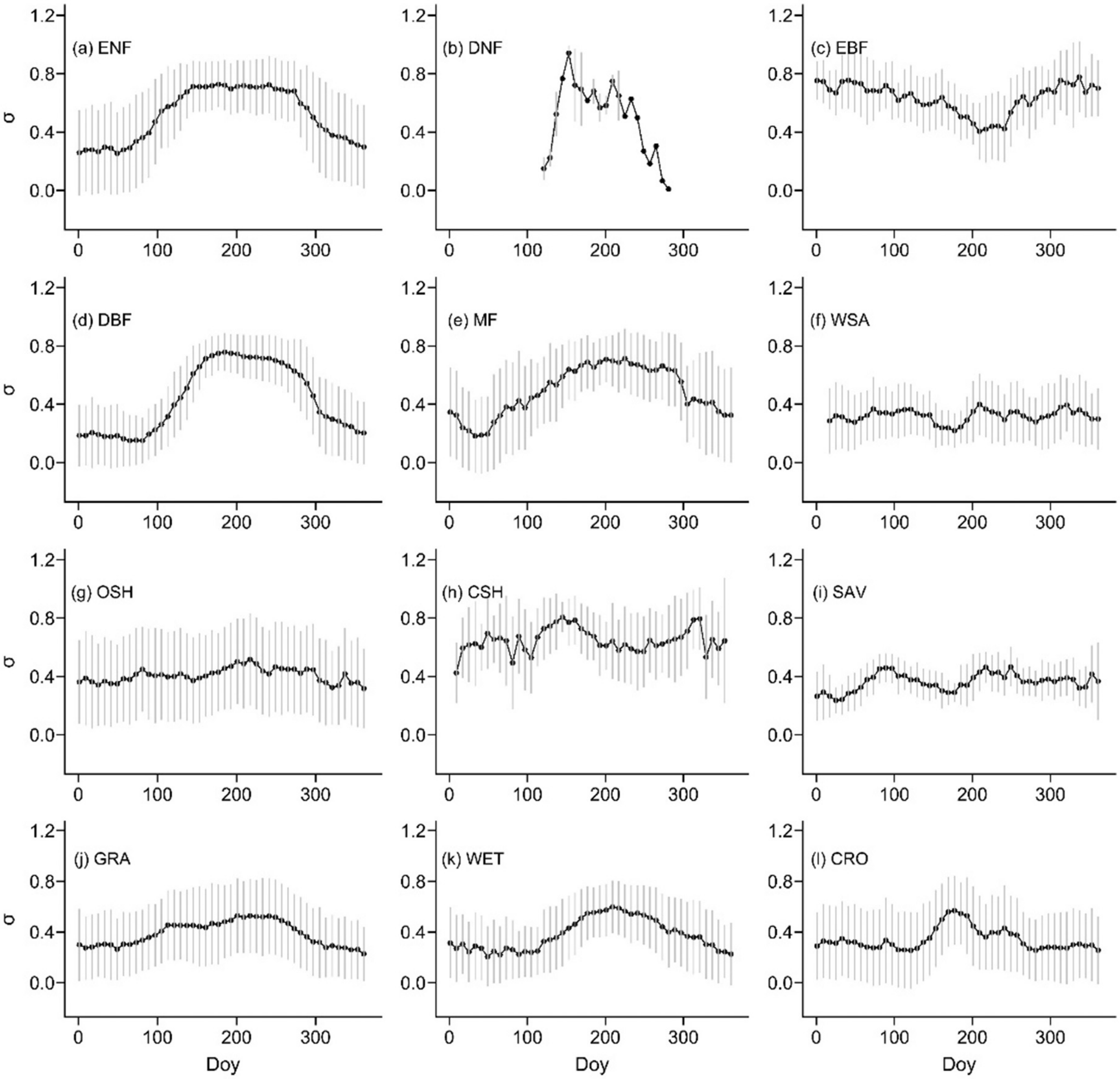


Fig. 2. Global changes (black points) in the eight-day σ for different vegetation types over 2002–2020 based on analysis data from study sites located in the Northern Hemisphere; the grey bar represents the σ uncertainty measured by the standard deviation (SD). There is great seasonal variation in σ, particularly for ENF, DNF, EBF, and MF. However, its uncertainty cannot be ignored, which most likely threatens the global GPP estimation in the TL-LUE model.

analysis reveals that the overall RMSE and RPE for the 2EV scenario are approximately 32.4 and 33.8 % lower than those for the 2AV scenario (Fig. 3), indicating that the RF model can capture nonlinear mutual relationships among environmental factors, surpassing the simple analytic formula employed in the LUE models. Furthermore, the overall value of R^2 demonstrates a hierarchical trend across the different explanatory variable scenarios, with the highest value observed for 14EV, followed by 4EV, and the lowest for 2EV. Correspondingly, the overall RMSE and RPE exhibit an inverse rank, with the lowest values recorded for 14EV, followed by 4EV, and the highest for 2EV (Fig. 3). These findings affirm that incorporating more environmental stress factors refines σ estimations, facilitated by the RF algorithm. Comparing 14EV to 4EV, a notable improvement is observed, with the overall R^2 being approximately 16.4 % higher and the overall RMSE and RPE being approximately 11.8 % and 14.8 % lower, respectively. However, the computational efficiency of the 4EV outperforms that of the 14EV. Additionally, the RPE varies across vegetation types, ranging from approximately 1.5 to 27 % for different vegetation types. Notably, the lowest RPE was observed in DNF, whereas the highest RPE occured in CRO (Fig. 3).

4.3. Validation of GPP estimation against sites

4.3.1. Accuracy of GPP estimation in different models

The performance of the TL-CRF model surpasses that of the TL-CLUE model, as evidenced by significantly lower RMSE and MAE values, along with a higher R^2, indicating that global GPP estimation is improved by integrating multiple environmental factors via the RF technique. Specifically, the RMSE and MAE of 2EV are approximately 40.4 and 43.4 % lower than those of 2AV (Fig. 4), respectively, underscoring the advantages of the RF algorithm in capturing the complex interactions among various driving variables. Furthermore, compared with the 2EV scenario, the inclusion of additional explanatory variables (4EV and 14EV) leads to remarkable improvements. The R^2 values increase by approximately 9 % and 13 % for 4EV and 14EV, respectively, whereas the RMSE decreases by approximately 18 % and 28 % (Fig. 4), indicating improved global GPP estimations with the increasing explanatory variables. Notably, the 14EV scenario results in a 0.03 higher R^2 than the 4EV scenario does, with approximately 12 % lower RMSE and MAE values (Fig. 4), further confirming the efficacy of integrating a broader range of explanatory variables. Moreover, variations in GPP estimation

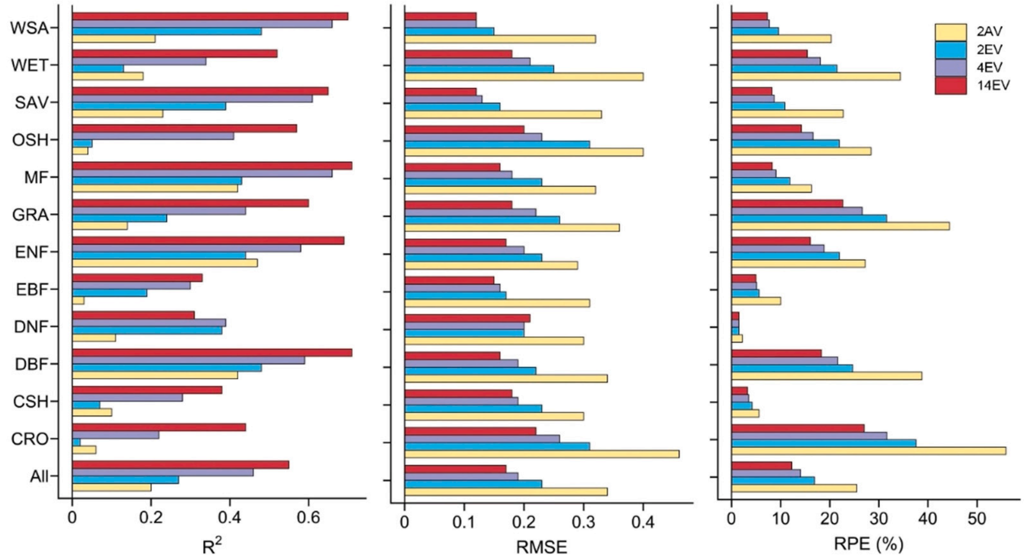


Fig. 3. Comparison of the σ estimation in the TL-CLUE model driven by two environmental stress factors (VPD and T_{amin}) via the analytic formula (2AV), two explanatory factors (VPD and T_{amin}) (2EV), four main explanatory factors (TWS, T_{amin}, T_s, and VPD) (4EV), and fourteen explanatory factors (14EV) via the RF model for different vegetation types. The σ estimation based on 2EV is significantly more precise than that based on 2AV, which might be attributed to the fact that the RF technique can capture the complex nonlinear relationships between VPD and T_{amin} as opposed to the analytic formula employed in the TL-LUE model. Moreover, increasing the number of explanatory variables in the RF model leads to the high accuracy of σ estimation.

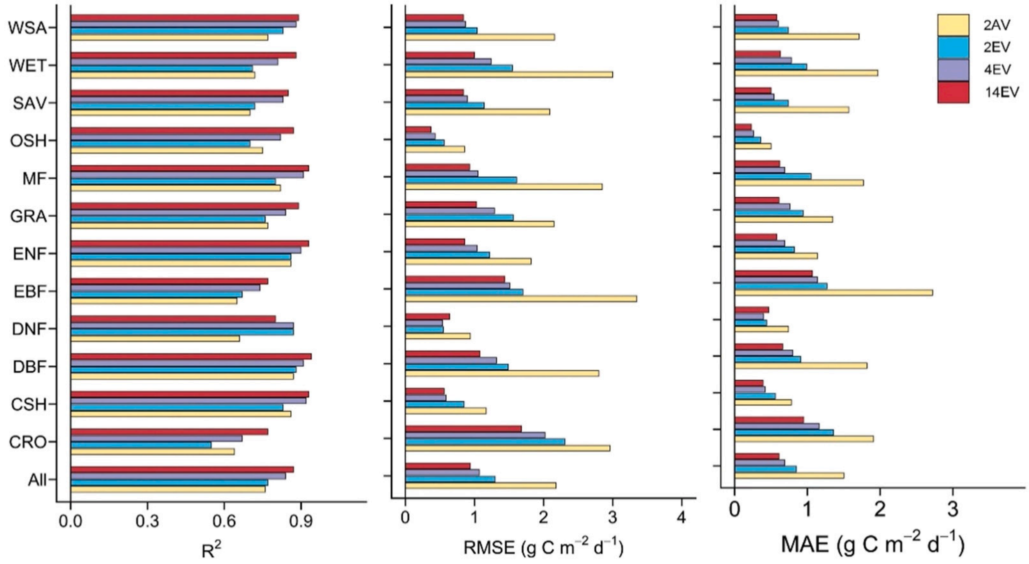


Fig. 4. Comparison of the eight-day GPP estimation among the four scenarios of integrations of environmental stress factors for different vegetation types; its legend is the same as that in Fig. 3.

accuracy across different vegetation types can be observed (Fig. 4). In the TL-CRF model driven by 14EV, the R^2 values range from approximately 0.77 to 0.94, with the highest R^2 observed in DBF and the lowest in CRO (Fig. 4). Correspondingly, the MAEs vary across vegetation types, with the lowest values observed in OSH and the highest in CRO (Fig. 4). Compared with those of the TL-CRF model driven by 14EV and the TL-CLUE model, the R^2 values are increased by approximately 20 %∼47 %, whereas both the RMSE and MAE are reduced by approximately 43 %∼88 % across different vegetation types (Fig. 4). These findings highlight the significant reduction in uncertainties in GPP estimation through the integration of numerous environmental factors the RF model.

4.3.2. *Consistency of GPP estimations with GPP observations*

The GPP estimated by the TL-CRF model is greater than that estimated by the TL-CLUE model, which is consistent with the observed GPP for eight-day, monthly, and yearly values and for the eight-day average values with respect to the monthly values (Fig. 5). The R^2 of the TL-CRF model is approximately 0.9 for eight-day scales, 0.91 for month scales, and 0.92 for year scales, demonstrating a gradual increase with increasing time scale for the correlation between the GPP estimation in the TL-CRF model and the observed GPP (Fig. 5). In terms of bias, the RMSE of the TL-CRF model (0.43 g C m^{-2} d^{-1}) is lower than that of the TL-CLUE model (0.59 g C m^{-2} d^{-1}) (Fig. 5). Additionally, the correlation of the eight-day GPP estimation anomaly with the GPP observation anomaly is notably greater in the TL-CRF model ($R^2 = 0.45$) than that in the TL-CLUE model ($R^2 = 0.34$) (Fig. 5).

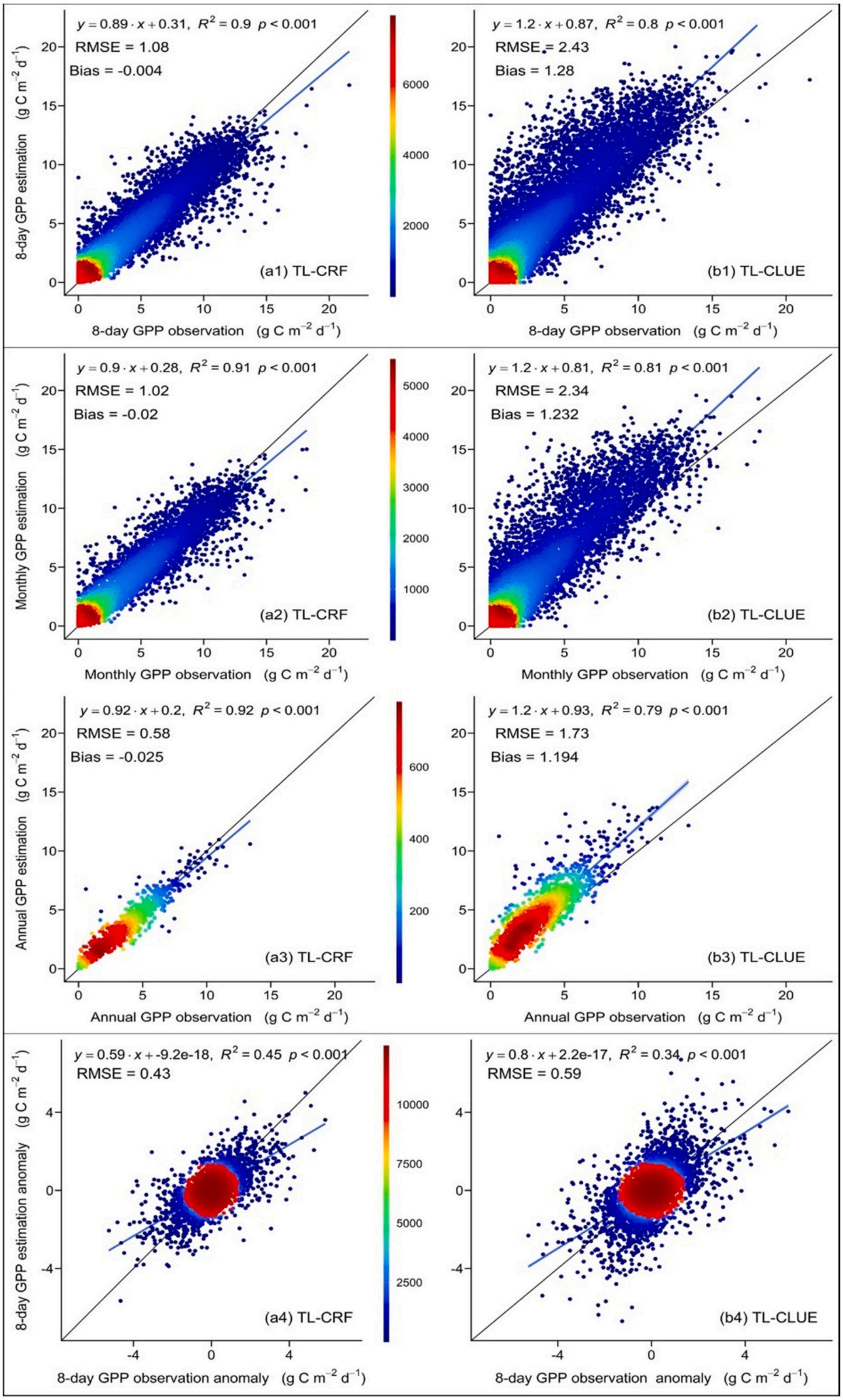


Fig. 5. Comparison of the consistency of the observed GPP and GPP estimated by the TL-CRF (a1−a4), and TL-CLUE (b1−b4) models on the eight-day (a1, b1), monthly (a2, b2), and annual (a3, b3) scales, and eight-day anomalies (a4, b4). Compared with the regression line (blue solid line) of the TL-CLUE model, that of the TL-CRF model is closer to the 1:1 theoretical line represented by the black solid line, regardless of the eight-day, monthly, and annual scales.

4.4. *Effects of the environmental variables on σ*

4.4.1. *The relative importance of the explanatory variables to σ*

The relative importance of environmental stress factors to the σ can be categorized into four ties: ranging from approximately 35 to 50 % for TWS, T_{amin}, T_s, and VPD, 29 % for T_{amax} and W_s, ranging from 20 to 24 % for S_t, Elev, S_d, S_y, S_n, SOC and FCA, and 17 % for P (Fig. 6). TWS with a relative importance of 49.55 %, was the most influential environmental factor (Fig. 6), implying its significant impact on vegetation photosynthesis.. Notably, T_{amin} with a relative importance of 41 %, stands out as the most essential temperature stress factor for ε_{max} relative to T_s and T_{amax} (Fig. 6). In contrast, soil properties exhibit lower significance than do meteorological factors (Fig. 6), with soil silt emerging as the most influential soil stress factor. These results underscore the considerable necessity of incorporating TWS and T_{amin} into LUE models as they serve as better indicators of water and temperature, respectively, thereby improving σ estimation.

4.4.2. *Functional relationships between σ and meteorological factors*

To distinctly illustrate the functional connections between GPP and three meteorological factors, VPD, TWS, and T_{amin}, the study used point density to weaken the interference of some abnormal points to accentuate the dominant curve shape. Fig. 7 shows that the respective functional response of σ to VPD, TWS, and T_{amin} is clearly and slightly differed across the three analyzed growth periods (Fig. 7), despite the possible changes caused by the intrinsic ecosystem properties. As shown in the data with relatively high point density in Fig. 7a, b, c, the functional relationships between the σ and VPD show a strong and negative correlation, particularly at the growth peak (June–July). The functional response of σ to TWS follows the expected increase in σ with TWS but a sudden turning point might be due to occasional extreme weather events. The functional response of σ to T_{amin} can be expressed by the bell curve (Fig. 7), which is in line with the current knowledge. There is an optimal value for the minimum air temperature.

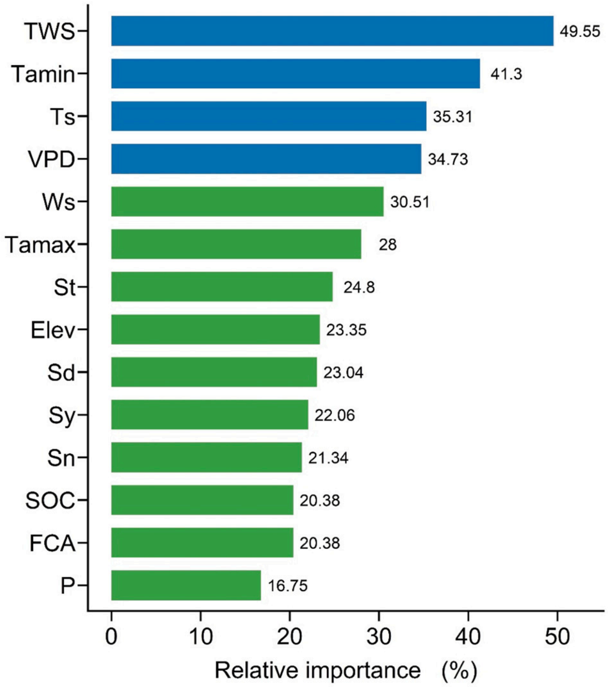


Fig. 6. The relative importance of the explanatory variables to σ is determined via the RF algorithm. TWS is the most important environmental factor and has a strong influence on vegetation photosynthesis.

4.5. *Spatial and temporal patterns of global GPP, GPP_{su}, and GPP_{sh}*

The annual global GPP exhibited considerable spatial heterogeneity from 2002 to 2020 (Fig. 8). Regions with low annual GPP values, less than 500 g C m^{-2} a^{-1}, are predominantly observed in most Australia, southern South America, central Asia, and northern China at 65°N. Intermediate annual GPP values (1000–2000 g C m^{-2} a^{-1}) are found in belt regions between 45 and 60°N (Fig. 8). Conversely, high annual GPP, ranging from 2500 to 4000 g C m^{-2} a^{-1}, is primarily concentrated in the Amazon, Congo Basin, northern regions of South America, Middle Africa, Southeast Asia, the U.S. corn belt, and western Europe (Fig. 8). During summer, spatial GPP patterns mirror those of the annual GPP, despite substantially lower magnitudes, ranging from 1500 to 2000 g C m^{-2} a^{-1} (Fig. 8). The spatial patterns and magnitudes of GPP in spring are analogous to those in fall. However, during winter, the GPP is near 0 across much of the northern hemisphere. Additionally, these results also revealed significant seasonal differences in global GPP.

Irrespective of season, the spatial patterns of both GPP_{sh} and GPP_{su} (Fig. 9) closely resemble those of the GPP (Fig. 8). The GPP_{sh} tends to surpass the GPP_{su} near the equator and in the regions between 45 and 60°N (Fig. 9). The annual GPP_{sh} exceeds the GPP_{su} by approximately 500 g C m^{-2} a^{-1}. Both GPP_{sh} and GPP_{su} exhibit the highest values in summer, followed by those in fall and spring, and the lowest values occur in winter (Fig. 9). Regardless of the season, the contribution of GPP_{sh} to the total GPP was greater than that of GPP_{su} to total GPP. An increase in the LAI and diffuse solar radiation would lead to high GPP_{sh}, thereby increasing GPP. Therefore, the GPP, GPP_{su}, and GPP_{sh} datasets provided by this study could be employed to explore the differences and connections in GPP between sunlit and shaded leaves, and further study the internal mechanisms of different carbon assimilation processes.

5. Discussions

Our study is novel in that it compiles a fresh global GPP dataset of sunlit and shaded leaves via various environmental stress factors and seasonal changes in canopy structure and investigates the respective photosynthetic roles of these environmental variables. The results demonstrate that the TL-CRF model (R^2 = 0.87, RMSE = 0.94, MAE = 0.61 g C m^{-2} d^{-1}) significantly outperforms the TL-LUE model (R^2 = 0.76, RMSE = 2.18, MAE = 1.5 g C m^{-2} d^{-1}) (Fig. 4). This hybrid model uses qualitative hypotheses and respects physical principles, allowing the TL-CRF model to harness the advantages of both process-based and data-driven approaches. To further evaluate the accuracy of GPP estimation in the TL-CRF model, it is crucial to compare it with other published GPP products, which is essential for assessing global or regional vegetation productivity.

Unlike the linear analytic formula in the TL-LUE model, the RF model improves the global GPP estimation by capturing nonlinear relationships between VPD and Tamin (Fig. 4), which may be linked to the various parameterization strategies and structures of the model and environmental gradients (Pastorello et al., 2020b). The relationship between the target variable and driven variables in the ML model is determined mainly by data statistics rather than by imperfect assumptions. The TL-CRF model rarely assumes that the GPP is positive and that σ ranges from 0 to 1. In contrast to the analytic formula of the TL-LUE model, the RF model has few established connections between the aim variable and driven variables, which is robust to this incomplete or incorrect formula. These traits of the ML are suitable forresearch because the response patterns of vegetation physiological processes to environmental factors vary with vegetation type and element. A typical example is the stress effect of complex environmental factors on photosynthetic capacity.

Although the RF model is known for its robustness and excellent ability to address complex relationships, the computational complexity of evaluating predictor combinations likely becoming a limiting factor, particularly when working with large ecological datasets. In the

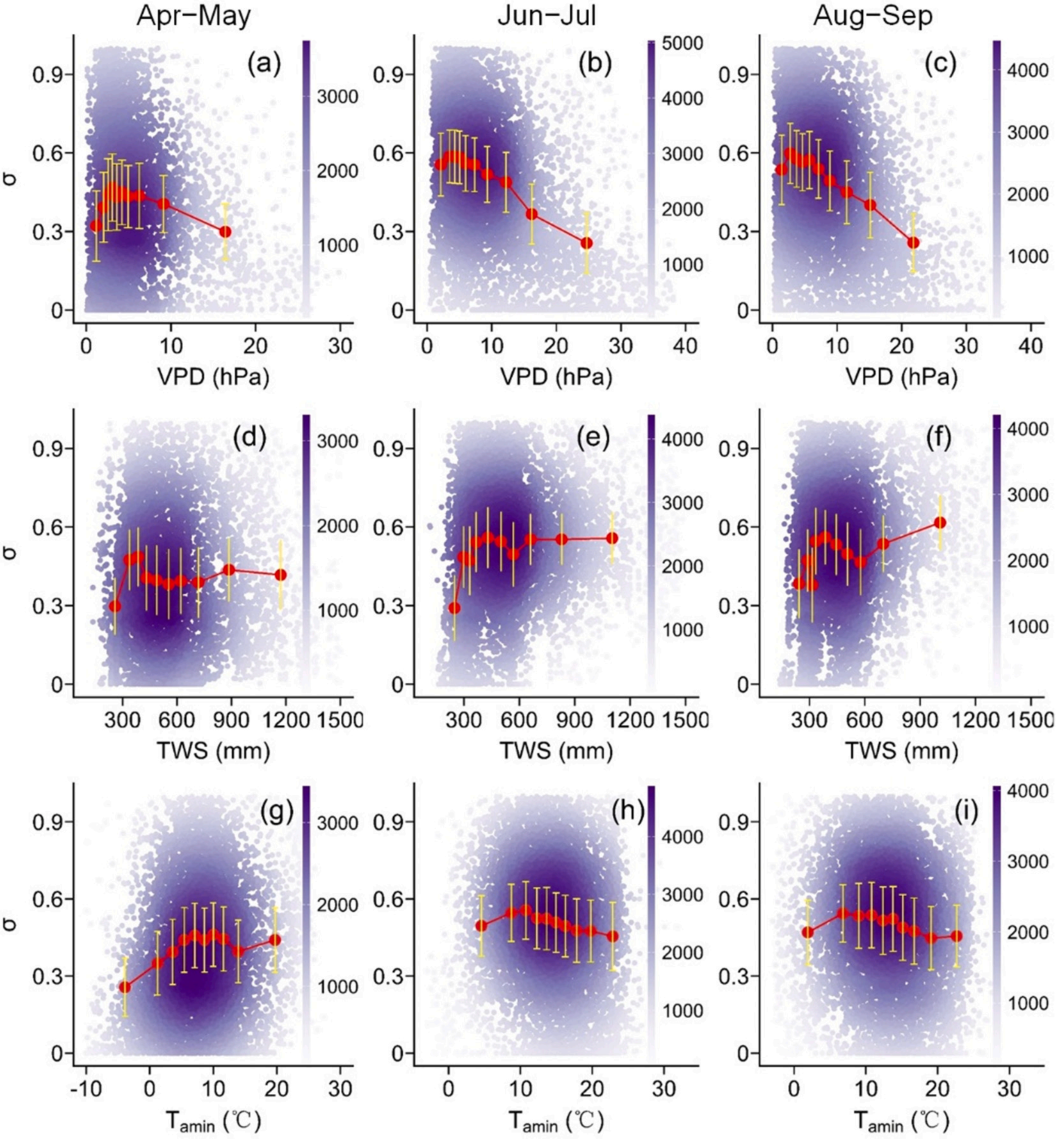


Fig. 7. Univariate ecosystem functional relationships between σ and VPD (hPa, a–c), TWS (mm, d–f), and T_{amin} (°C, g–i) based on the global EC site data for each two months of the growing season. The analysis data are classified into ten bins on the basis of σ and three meteorological factors. The red dots represent the means of each corresponding bin, the yellow lines represent the mean fluctuations measured by the standard deviation (SD), the gradient purple dots represent the distributions of site-specific raw data by point density, and the red lines represent the functional relationships between the σ and three meteorological factors.

ecological community, to capture complex environmental gradients and temporal changes, datasets are generally large and contain many environmental variables (Cutler et al., 2007), which leads to high computational costs. Additionally, an increase in the number of decision trees generally improves the accuracy of the model but also increases the computational time. The RF model can utilize parallel processing to significantly reduce the training time of the model. Each tree in the RF model can be individually constructed. Moreover, the computational efficiency can also be improved by some optimizing tools and library functions such as the "randomFroest" package in the R platform.

The incorporating of more driving variables leads to more accurate global GPP estimation in the TL-CRF model. Regardless of magnitude or trend, our results are more consistent with the GPP observations, which can be attributed to the fact that the ML algorithm can detect the explicit or implicit features related to photosynthesis from many driven variables, which can hardly be modeled physically. The physical formula generally requires the clear role of a single variable in an ecosystem and its quantitative relationship with the aim variable. Additionally, GPP is closely related to vegetation type (Xu et al., 2019), climate factors (Wang et al., 2018b, 2020; Baldocchi, 2020), and soil moisture (Stocker et al., 2019). These are incorporated in the study, which provides the TL-CRF model with rich helpful information to improve global GPP estimation. Different vegetation types and characteristics influence the proportional contributions of biological and abiotic factors to water use efficiency (Zhao et al., 2021). For example, precipitation during the growing season is a significant factor for GPP in different regions of the Mongolian Plateau (Ding et al., 2024), even though TWS was found to be more influential on a global scale.

Most carbon cycle models based on traditional water indices such as precipitation, drought indices, and soil moisture most likely underestimate the response of vegetation to drought (Humphrey et al., 2018). This study reveals that TWS is the most important environmental factor (Fig. 6). TWS integrates all water fluxes, regulating exchanges among different Earth system components (Tapley et al., 2019), and offers a thorough assessment of drought conditions (Zhao et al., 2017). Furthermore, changes in TWS are closely related to variations in droughts, floods, and global sea levels (Scanlon et al., 2018; Wang et al., 2018a). Therefore, TWS is a more adaptive water indicator than soil

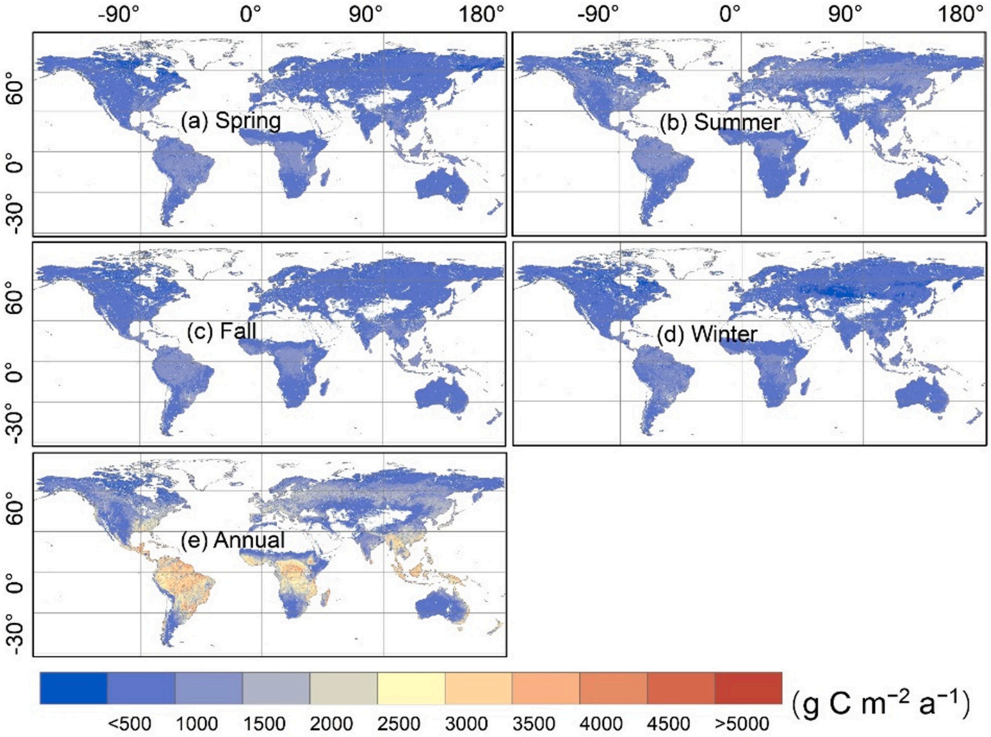


Fig. 8. Spatial patterns of global GPP in spring (a), summer (b), fall (c), winter (d), and year (e) from 2002 to 2020.

moisture in reflecting the ecological response of vegetation photosynthesis to changes in water. The MERRA-2 product used in the study provides a global gridded TWS estimation by combining multiple observations from the ground, airplane, and satellite with models (Gelaro et al., 2017). The TL-CRF model could detect the feedback of the water cycle on changes in the ecosystem since it also uses the MERRA-2 TWS as input in this study.

Indeed, the quality of site data may be one of the main sources of uncertainty in the hybrid model that needs to be addressed. The observation radius of an EC tower generally ranges from hundreds of meters to kilometers, leading to variability in the carbon footprint. EC towers measure gas flux between the land and atmosphere by detecting turbulence generally driven by wind speed. This study indirectly considers the effect of footprint variability on flux observations by using wind speed as an input. However, the uncertainty caused by the carbon footprint remains a significant concern. Additionally, inaccuracies in flux data processing, such as imperfect GPP separation methods and incomplete gap-filling algorithms (Moffat et al., 2007), can affect the training and evaluation of the ML model (Jung et al., 2020). The EC flux data used in this study were processed through a standard and comprehensive data processing pipeline, ensuring their reliability for the training model. Moreover, the distribution and density of sites pose a challenge to the accuracy of the hybrid model. In this study, 267 EC sites from the FLUXNET2015 and Ameriflux datasets were used to construct the TL-CRF model, reducing the uncertainty caused by insufficient site numbers, particularly in Europe and North America where site observations are intensive. Therefore, increasing the number of EC sites and improving data quality in ecologically sensitive regions, such as the Qinghai–Tibet Plateau and the Amazon rainforest, are fundamental for further studies of the carbon cycle and its response to environmental changes.

To estimate the global seasonal differences in the CI, the one-year leaf life cycle is divided into three periods, namely, the leaf-off, leaf-scattered, and leaf-gathered seasons, according to the four phenological thresholds of the MODIS LSP product, which are calculated via the nnhanced vegetation index 2 (EVI2). EVI2 is widely used to evaluate and monitor vegetation growth and health, particularly for phenological occurrences. However, variations in vegetation structure are not immediately detected by relative vegetation indices. Generally, the increase in leaf greenness, represented by the chlorophyll concentration, lags behind the changes in canopy structure (Li et al., 2022). Therefore, MidGreenup was selected as the start date of the leaf-gathering season in this study instead of Maturity from the MODIS LSP product.

Although we selected high-quality MODIS CI data inverted by the hotspot-adjusted model with a backup algorithm (Jiao et al., 2018), CI data retrieved from remote sensing directional observations tend to remain uncertain. First, insufficient BRDF information, particularly for the reconstructed BRDF information, can affect the inversion of key land surface parameters such as CI because of viewing angles. Second, CI retrieval from remote sensing data is influenced by reflectance underestimation in hotspot regions (Dong et al., 2015), whereas the hotspot effect has already been more accurately reconstructed by the optimized kernel-driven RossThick-LiSparseReciprocal BRDF model (RTLSRJ) (Jiao et al., 2018), which improves the geometric–optical kernel and combines it with hotspot adjustment of the volumetric kernel (Jiao et al., 2016). Additionally, CI product uncertainty could also be caused by heterogeneous pixels and short-term fluctuations, which are not elaborated in this study due to space limitations. Hence, future work should be undertaken to develop an advanced CI product to promote further understanding of the structure and temporal changes in global vegetation (Wei et al., 2019). Considering the seasonal differences in CI reduces the overestimation of the eight-day global GPP estimation by approximately 8.3 %. However, the potential information on the seasonal CI likely does not play a remarkable role in the integration process of the seasonal variation in multiple environmental factors based on the RF algorithm.

This study improves global GPP estimation by incorporating multiple environmental stress factors. Nevertheless, vegetation photosynthesis is also disturbed by accidental processes such as wildfires and abrupt permafrost thaw. These emergencies are not currently included in most land surface process models. Additionally, a hybrid model generally requires enormous datasets as inputs to drive. Recently, although

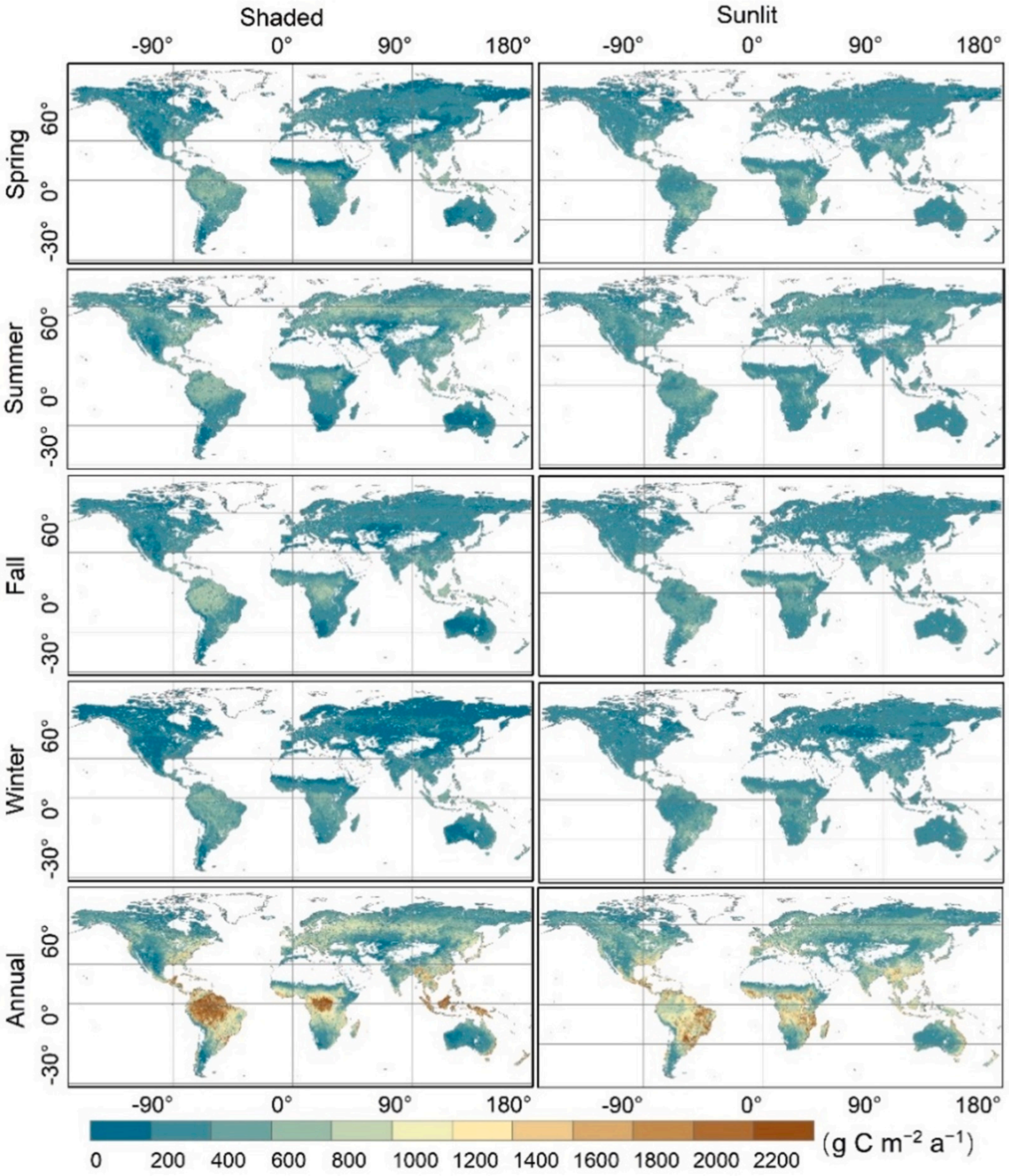


Fig. 9. Comparison of spatial characteristics of global GPP_{sh} and GPP_{su} based on the TL-CRF model for different seasons from 2002 to 2020.

remote sensing observations have accumulated abundant continuous large-scale datasets, it is necessary to coordinate these data from different resources before inputting them into the model. Hence, there is an urgent need to improve the temporal and spatial resolutions of datasets, which are essential for estimating the carbon cycle and capturing irregular variations in transition areas, particularly for mixed vegetation. The development of data assimilation techniques (Khaki et al., 2020; Huang et al., 2024) and spatial interpolation methods (Zakeri and Mariethoz, 2021; Hu et al., 2019) could help overcome this challenge. A big data frame algorithm also facilitates the processing of plentiful remote sensing data in cloud-computing platforms such as the Google Earth Engine (GEE) (Sun et al., 2019). Moreover, computing power is another limiting factor for hybrid models. With improvements in computing power and algorithm optimization such as parallel processing and cloud-computing, significant progress has been made in terms of computing power.

The TL-CRF model provides a new pathway for estimating global or regional GPP, even other key ecological indicators. From a methodological perspective, the TL-CRF model is an example of basic hybrid modeling that couples a machine learning technique with a process-based model (Reichstein et al., 2019). The TL-CRF model enhances the flexibility in modeling diverse ecosystems and climate conditions, which improves the applicability of hybrid models in diverse environmental scenarios. It is also essential for ML from only diagnostic applications to geographical process predictions, to support research on climate change and the development of mitigation strategies (Reichstein et al., 2019). The GPP products generated in this study would be highly interesting for the precise management of agriculture and forests and for climate change mitigation. Some new vegetation indices that have emerged in recent decades are expected to advance the study of vegetation productivity. For example, the long-term GPP dataset based on near-infrared reflectance of terrestrial vegetation (NIR_v) better captured seasonal and interannual changes in terrestrial GPP (Wang et al., 2021). Recently, the increase in data flow from Earth observation systems and the accessibility and availability of powerful artificial intelligence tools, particularly within the ML community, have promoted numerous innovations and developments to overcome current shortcomings in Earth system models (Irrgang et al., 2021).

6. Conclusions

The goal of this study was to improve global GPP estimation by embedding the RF submodule into the TL-LUE model. Before various environmental stress factors are integrated, the seasonal variations in canopy structure are considered by averaging the CIs. The main results are as follows:

(1) The TL-CRF model outperforms the TL-CLUE model, showing superior performance ($\Delta_{R}^{2} = 0.11$, $\Delta_{RMSE} = -1.24$, $\Delta_{MAE} = -0.89$ g C m^{-2} d^{-1}) and greater consistency with GPP observations at eight-day ($R^2 = 0.9$), monthly ($R^2 = 0.91$), and annual ($R^2 = 0.92$) scales.

(2) TWS was found to be the most important water stress factor, suggesting that TWS has greater potential to reflect the constraints of water content on vegetation photosynthesis. This highlights the necessity of incorporating TWS into LUE models.

(3) Analysis of the functional relationships confirmed that there is an optimal value of air temperature for vegetation photosynthesis, particularly at peak growth.

(4) This study simulated the global GPP, GPP_{su}, and GPP_{sh} from 2002 to 2020 via the TL-CRF model, which exhibited significant spatial heterogeneity and seasonal variation. The annual GPP_{su} was approximately 500 g C m^{-2} a^{-1} lower than the annual GPP_{sh}.

This study provides potential as an ecological and environmental basis for mitigating global warming and managing ecosystems. There are few constant patterns in the functional response of vegetation photosynthesis to environmental factors in terrestrial ecosystems. The comprehensive control of photosynthesis intensity by environmental stress factors is crucial for the assessment of large-scale carbon exchange. Despite these encouraging results, future work is needed to further examine the accuracy of GPP estimation by comparing it with other published GPP products such as MODIS GPP. With intelligent remote sensing coming, future work should be undertaken to intelligently retrieve carbon cycle parameters by automatically selecting the main driven data from multisource remote sensing big data on the basis of multiscale synthesis observations of satellites, airplanes, and fields.

CRediT authorship contribution statement

Zhilong Li: Writing – original draft, Methodology, Formal analysis, Data curation, Conceptualization. **Ziti Jiao:** Writing – review & editing, Resources, Project administration, Funding acquisition, Conceptualization. **Ge Gao:** Writing – original draft, Methodology, Data curation. **Jing Guo:** Validation, Software, Formal analysis. **Chenxia Wang:** Visualization, Methodology, Data curation. **Sizhe Chen:** Visualization, Validation. **Zheyou Tan:** Visualization, Formal analysis. **Wenyu Zhao:** Writing – review & editing, Validation.

Declaration of competing interest

The authors declare that no conflict of interest appeared in the paper.

Data availability

All data used in this study can be publicly accessed. The AmeriFlux and FLUXNET2015 datasets can be downloaded from https://ameriflux.lbl.gov/login/?redirect_to=/data/download-data/, accessed on 22 October 2022 and https://fluxnet.org/login/?redirect_to=/data/download-data/, accessed on 23 October 2023, respectively. The MODIS vegetation indices with 500 m of spatial resolution including eight-day LAI, year Land Surface Phenology, and Land Cover can be downloaded from https://search.earthdata.nasa.gov/search, accessed on 10 March 2023. Meteorological data from NASA MERRA-2 can be downloaded from https://disc.gsfc.nasa.gov/, accessed on 20 March 2023. Soil properties from the ISRIC dataset can be downloaded from https://www.soilgrids.org/, accessed on 24 March 2023, and the global digital elevation model from USGS can be downloaded from https://earthexplorer.usgs.gov/, accessed on 25 March 2023.

Acknowledgment

This work was supported by the General Program of the Natural Science Foundation of China (41971288) and the Major Program of the Natural Science Foundation of China (42090013). We appreciate the AmeriFlux and FLUXNET2015 communities for sharing flux data on global eddy covariance sites. We are grateful to the data centers of MODIS, NASA MERRA-2, USGS, and ISRIC for providing global spatial data products.

References

Anav, A., Friedlingstein, P., Beer, C., Ciais, P., Harper, A., Jones, C., et al., 2015. Spatiotemporal patterns of terrestrial gross primary production: a review. Rev. Geophys. 53, 785–818. https://doi.org/10.1002/2015RG000483.

Badgley, G., Anderegg, L.D., Berry, J.A., Field, C.B., 2019. Terrestrial gross primary production: using NIRV to scale from site to globe. Glob. Chang. Biol. 25, 3731–3740. https://doi.org/10.1111/gcb.14729.

Baldocchi, D.D., 2020. How eddy covariance flux measurements have contributed to our understanding of Global Change Biology. Glob. Chang. Biol. 26, 242–260. https://doi.org/10.1111/gcb.14807.

Bao, S., Wutzler, T., Koirala, S., Cuntz, M., Ibrom, A., Besnard, S., et al., 2022. Environment-sensitivity functions for gross primary productivity in light use efficiency models. Agric. For. Meteorol. 312, 108708. https://doi.org/10.1016/j.agrformet.2021.108708.

Beer, C., Reichstein, M., Tomelleri, E., Ciais, P., Jung, M., Carvalhais, N., et al., 2010. Terrestrial gross carbon dioxide uptake: global distribution and covariation with climate. Science 329, 834–838. https://doi.org/10.1126/science.1184984.

Bergen, K.J., Johnson, P.A., de Hoop, M.V., Beroza, G.C., 2019. Machine learning for data-driven discovery in solid earth geoscience. Science 363, eaau0323. https://doi.org/10.1126/science.aau03.

Beucler, T., Pritchard, M., Gentine, P., Rasp, S., 2020. Towards physically-consistent, data-driven models of convection. In: IGARSS 2020–2020 IEEE International Geoscience and Remote Sensing Symposium, Waikoloa, HI, USA, 2020. IEEE, pp. 3987–3990. https://doi.org/10.1109/IGARSS39084.2020.9324569.

Bi, W., He, W., Zhou, Y., Ju, W., Liu, Y., Liu, Y., et al., 2022. A global 0.05 dataset for gross primary production of sunlit and shaded vegetation canopies from 1992 to 2020. Sci. Data 9, 213. https://doi.org/10.1038/s41597-022-01309-2.

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32. https://doi.org/10.1023/A:1010933404324.

Brenowitz, N.D., Bretherton, C.S., 2018. Prognostic validation of a neural network unified physics parameterization. Geophys. Res. Lett. 45, 6289–6298. https://doi.org/10.1029/2018GL078510.

Chang, X., Xing, Y., Gong, W., Yang, C., Guo, Z., Wang, D., et al., 2023. Evaluating gross primary productivity over 9 ChinaFlux sites based on random forest regression models, remote sensing, and eddy covariance data. Sci. Total Environ. 875, 162601. https://doi.org/10.1016/j.scitotenv.2023.162601.

Chapin, F.S., Matson, P.A., Mooney, H.A., Vitousek, P.M., 2002. Principles of Terrestrial Ecosystem Ecology. Springer, New York, NY, USA, pp. 3–22.

Chen, J., Liu, J., Cihlar, J., Goulden, M., 1999. Daily canopy photosynthesis model through temporal and spatial scaling for remote sensing applications. Ecol. Model. 124, 99–119. https://doi.org/10.1016/S0304-3800(99)00156-8.

Chen, Y., Feng, X., Fu, B., Wu, X., Gao, Z., 2021. Improved global maps of the optimum growth temperature, maximum light use efficiency, and gross primary production for vegetation. J. Geophys. Res.: Biogeosci. 126, e2020JG005651. https://doi.org/10.1029/2020JG005651.

Chen, J.M., Ju, W., Cihlar, J., Price, D., Liu, J., Chen, W., et al., 2003. Spatial distribution of carbon sources and sinks in Canada's forests. Tellus Ser. B Chem. Phys. Meteorol. 55, 622–641. https://doi.org/10.3402/tellusb.v55i2.16711.

Chen, B., Coops, N.C., Fu, D., Margolis, H.A., Amiro, B.D., Black, T.A., et al., 2012a. Characterizing spatial representativeness of flux tower eddy-covariance measurements across the Canadian Carbon Program Network using remote sensing and footprint analysis. Remote Sens. Environ. 124, 742–755. https://doi.org/10.1016/j.rse.2012.06.007.

Chen, J.M., Mo, G., Pisek, J., Liu, J., Deng, F., Ishizawa, M., et al., 2012b. Effects of foliage clumping on the estimation of global terrestrial gross primary productivity. Glob. Biogeochem. Cycles 26, GB1019. https://doi.org/10.1029/2010GB003996.

Cheng, S.J., Bohrer, G., Steiner, A.L., Hollinger, D.Y., Suyker, A., Phillips, R.P., et al., 2015. Variations in the influence of diffuse light on gross primary productivity in temperate ecosystems. Agric. For. Meteorol. 201, 98–110. https://doi.org/10.1016/j.agrformet.2014.11.002.

Chipman, H.A., George, E.I., McCulloch, R.E., 1998. Bayesian CART model search. J. Am. Stat. Assoc. 93, 935–948. https://doi.org/10.1080/01621459.1998.10473750.

Cutler, D.R., Edwards, Beard, K.H., Cutler, A., Hess, K.T., Gibson, J., et al., 2007. Random forests for classification in ecology. Ecology 88, 2783–2792. https://doi.org/10.1890/07-0539.1.

Cutler, A., Zhao, G., 2001. Pert-perfect random tree ensembles. Computing Science and Statistics 33, 90–94.

DE, P., 1997. Simple scaling of photosynthesis from leaves to canopies without the errors of big-leaf models. Plant Cell Environ. 20, 537–557. https://doi.org/10.1111/j.1365-3040.1997.00094.x.

Ding, L., Li, Z., Wang, X., Shen, B., Xiao, L., Dong, G., et al., 2024. Spatiotemporal patterns and driving factors of gross primary productivity over the Mongolian Plateau steppe in the past 20 years. Sci. Total Environ. 170886. https://doi.org/10.1016/j.scitotenv.2024.170886.

Dong, J., Xiao, X., Wagle, P., Zhang, G., Zhou, Y., Jin, C., et al., 2015. Comparison of four EVI-based models for estimating gross primary production of maize and soybean croplands and tallgrass prairie under severe drought. Remote Sens. Environ. 162, 154–168. https://doi.org/10.1016/j.rse.2015.02.022.

Fang, H., 2021. Canopy clumping index (CI): a review of methods, characteristics, and applications. Agric. For. Meteorol. 303, 108374. https://doi.org/10.1016/j.agrformet.2021.108374.

Friedl, M.A., Sulla-Menashe, D., Tan, B., Schneider, A., Ramankutty, N., Sibley, A., et al., 2010. MODIS Collection 5 global land cover: algorithm refinements and characterization of new datasets. Remote Sens. Environ. 114, 168–182. https://doi.org/10.1016/j.rse.2009.08.016.

Friedl, M., Sulla-Menashe, D., 2022. MODIS/Terra+Aqua Land Cover Type Yearly L3 Global 500m SIN Grid V061 [Data set]. NASA EOSDIS Land Processes Distributed Active Archive Center. https://doi.org/10.5067/MODIS/MCD12Q1.061. Accessed 2023-08-14.

Geer, A.J., 2021. Learning earth system models from observations: machine learning or data assimilation? Phil. Trans. R. Soc. A 379, 20200089. https://doi.org/10.1098/rsta.2020.0089.

Gelaro, R., McCarty, W., Suárez, M.J., Todling, R., Molod, A., Takacs, L., et al., 2017. The modern-era retrospective analysis for research and applications, version 2 (MERRA-2). J. Clim. 30, 5419–5454. https://doi.org/10.1175/JCLI-D-16-0758.1.

GMAO, 2015a. MERRA-2 tavg1_2d_lnd_Nx: 2d,1-Hourly,Time-Averaged,Single-Level, Assimilation,Land Surface Diagnostics V5.12.4. Greenbelt, MD, USA, Goddard Earth Sciences Data and Information Services Center (GES DISC). https://doi.org/10.5067/RKPHT8KC1Y1T. Accessed 2023-08-14.

GMAO, 2015b. MERRA-2 tavg1_2d_slv_Nx: 2d,1-Hourly,Time-Averaged,Single-Level, Assimilation,Single-Level Diagnostics V5.12.4. Greenbelt, MD, USA, Goddard Earth Sciences Data and Information Services Center (GES DISC). https://doi.org/10.5067/VJAFPLI1CSIV. Accessed 2023-08-14.

GMAO, 2019. MERRA-2 tavg1_2d_rad_Nx: 2d,1-Hourly,Time-Averaged,Single-Level, Assimilation,Radiation Diagnostics V5.12.4. Greenbelt, MD, USA, Goddard Earth Sciences Data and Information Services Center (GES DISC). https://doi.org/10.5067/Q9QMY5PBNV1T. Accessed 2023-08-14.

Han, Y., Zhang, G.J., Huang, X., Wang, Y., 2020. A moist physics parameterization based on deep learning. J. Adv. Model. Earth Syst. 12, e2020MS002076. https://doi.org/10.1029/2020MS002076.

He, M., Ju, W., Zhou, Y., Chen, J., He, H., Wang, S., et al., 2013. Development of a two-leaf light use efficiency model for improving the calculation of terrestrial gross primary productivity. Agric. For. Meteorol. 173, 28–39. https://doi.org/10.1016/j.agrformet.2013.01.003.

He, L., Liu, J., Chen, J.M., Croft, H., Wang, R., Sprintsin, M., et al., 2016. Inter-and intra-annual variations of clumping index derived from the MODIS BRDF product. Int. J. Appl. Earth Obs. Geoinf. 44, 53–60. https://doi.org/10.1016/j.jag.2015.07.007.

Hengl, T., Mendes de Jesus, J., Heuvelink, G.B., Ruiperez Gonzalez, M., Kilibarda, M., Blagotić, A., et al., 2017. SoilGrids250m: global gridded soil information based on machine learning. PLoS One 12, e0169748. https://doi.org/10.1371/journal.pone.0169748.

Hu, Q., Li, Z., Wang, L., Huang, Y., Wang, Y., Li, L., 2019. Rainfall spatial estimations: A review from spatial interpolation to multi-source data merging. Water 11, 579. https://doi.org/10.3390/w11030579.

Huang, J., Song, J., Huang, H., Zhuo, W., Niu, Q., Wu, S., et al., 2024. Progress and perspectives in data assimilation algorithms for remote sensing and crop growth model. Sci. Remote. Sens. 10, 100146. https://doi.org/10.1016/j.srs.2024.100146.

Humphrey, V., Zscheischler, J., Ciais, P., Gudmundsson, L., Sitch, S., Seneviratne, S.I., 2018. Sensitivity of atmospheric CO2 growth rate to observed changes in terrestrial water storage. Nature 560, 628–631. https://doi.org/10.1038/s41586-018-0424-4.

Irrgang, C., Boers, N., Sonnewald, M., Barnes, E.A., Kadow, C., Staneva, J., et al., 2021. Towards neural Earth system modelling by integrating artificial intelligence in Earth system science. Nat. Mach. Intell. 3, 667–674. https://doi.org/10.1038/s42256-021-00374-3.

Jiao, Z., Schaaf, C.B., Dong, Y., Román, M., Hill, M.J., Chen, J.M., et al., 2016. A method for improving hotspot directional signatures in BRDF models used for MODIS. Remote Sens. Environ. 186, 135–151. https://doi.org/10.1016/j.rse.2016.08.007.

Jiao, Z., Dong, Y., Schaaf, C.B., Chen, J.M., Román, M., Wang, Z., et al., 2018. An algorithm for the retrieval of the clumping index (CI) from the MODIS BRDF product using an adjusted version of the kernel-driven BRDF model. Remote Sens. Environ. 209, 594–611. https://doi.org/10.1016/j.rse.2018.02.041.

Jung, M., Schwalm, C., Migliavacca, M., Walther, S., Camps-Valls, G., Koirala, S., et al., 2020. Scaling carbon fluxes from eddy covariance sites to globe: synthesis and evaluation of the FLUXCOM approach. Biogeosciences. https://doi.org/10.5194/bg-17-1343-2020.

Kraft, B., Jung, M., Körner, M., Reichstein, M., 2020. Hybrid modeling: fusion of a deep approach and physics-based model for global hydrological modeling. Int. Arch. Photogramm. Remote. Sens. Spat. Inf. Sci. 43, 1537–1544. https://doi.org/10.5194/isprs-archives-XLIII-B2-2020-1537-2020.

Khaki, M., Hendricks Franssen, H.-J., Han, S., 2020. Multi-mission satellite remote sensing data for improving land hydrological models via data assimilation. Sci. Rep. 10, 18791. https://doi.org/10.1038/s41598-020-75710-5.

Kraft, B., Jung, M., Körner, M., Koirala, S., Reichstein, M., 2022. Towards hybrid modeling of the global hydrological cycle. Hydrol. Earth Syst. Sci. 26, 1579–1614. https://doi.org/10.5194/hess-26-1579-2022.

Lee, J.E., Berry, J.A., van der Tol, C., Yang, X., Guanter, L., Damm, A., et al., 2015. Simulations of chlorophyll fluorescence incorporated into the Community Land Model version 4. Glob. Chang. Biol. 21, 3469–3477. https://doi.org/10.1111/gcb.12948.

Li, J., Lu, X., Ju, W., Li, J., Zhu, S., Zhou, Y., 2022. Seasonal changes of leaf chlorophyll content as a proxy of photosynthetic capacity in winter wheat and paddy rice. Ecol. Indic. 140, 109018. https://doi.org/10.1016/j.ecolind.2022.109018.

Li, Z., Jiao, Z., Wang, C., Yin, S., Guo, J., Tong, Y., et al., 2023. Seasonal effect of the vegetation clumping index on gross primary productivity estimated by a two-leaf light use efficiency model. Remote Sens. 15, 5537. https://doi.org/10.3390/rs15235537.

Li, X., Xiao, J., 2019. A global, 0.05-degree product of solar-induced chlorophyll fluorescence derived from OCO-2, MODIS, and reanalysis data. Remote Sens. 11, 517. https://doi.org/10.3390/rs11050517.

Liu, L., Gudmundsson, L., Hauser, M., Qin, D., Li, S., Seneviratne, S.I., 2020. Soil moisture dominates dryness stress on ecosystem production globally. Nat. Commun. 11, 4892. https://doi.org/10.1038/s41467-020-18631-1.

Lu, Q., Liu, H., Wei, L., Zhong, Y., Zhou, Z., 2024. Global prediction of gross primary productivity under future climate change. Sci. Total Environ. 912, 169239. https://doi.org/10.1016/j.scitotenv.2023.169239.

Ma, X., Huete, A., Yu, Q., Restrepo-Coupe, N., Beringer, J., Hutley, L.B., et al., 2014. Parameterization of an ecosystem light-use-efficiency model for predicting savanna GPP using MODIS EVI. Remote Sens. Environ. 154, 253–271. https://doi.org/10.1016/j.rse.2014.08.025.

Myneni, R., Knyazikhin, Y., Park, T., 2021. MODIS/Terra Leaf Area Index/FPAR 8-Day L4 Global 500m SIN Grid V061 [Data set]. NASA EOSDIS Land Processes Distributed Active Archive Center. https://doi.org/10.5067/MODIS/MOD15A2H.061. Accessed 2023-08-14.

de Mattos, E.M., Binkley, D., Campoe, O.C., Alvares, C.A., Stape, J.L., 2020. Variation in canopy structure, leaf area, light interception and light use efficiency among Eucalyptus clones. For. Ecol. Manag. 463, 118038. https://doi.org/10.1016/j.foreco.2020.118038.

Moffat, A.M., Papale, D., Reichstein, M., Hollinger, D.Y., Richardson, A.D., Barr, A.G., et al., 2007. Comprehensive comparison of gap-filling techniques for eddy covariance net carbon fluxes. Agric. For. Meteorol. 147, 209–232. https://doi.org/10.1016/j.agrformet.2007.08.011.

Pastorello, G., Trotta, C., Canfora, E., Chu, H., Christianson, D., Cheah, Y.-W., et al., 2020a. The FLUXNET2015 dataset and the ONEFlux processing pipeline for eddy covariance data. Sci. Data 7, 1–27. https://doi.org/10.1038/s41597-020-0534-3.

Pastorello, G., Trotta, C., Canfora, E., Chu, H., Christianson, D., Cheah, Y.-W., et al., 2020b. The FLUXNET2015 dataset and the ONEFlux processing pipeline for eddy covariance data. Sci. Data 7, 225. https://doi.org/10.1038/s41597-020-0534-3.

Pinto, F., Damm, A., Schickling, A., Panigada, C., Cogliati, S., Müller-Linow, M., et al., 2016. Sun-induced chlorophyll fluorescence from high-resolution imaging spectroscopy data to quantify spatio-temporal patterns of photosynthetic function in crop canopies. Plant Cell Environ. 39, 1500–1512. https://doi.org/10.1111/pce.12710.

Pisek, J., Govind, A., Arndt, S.K., Hocking, D., Wardlaw, T.J., Fang, H., et al., 2015. Intercomparison of clumping index estimates from POLDER, MODIS, and MISR satellite data over reference sites. ISPRS J. Photogramm. Remote Sens. 101, 47–56. https://doi.org/10.1016/j.isprsjprs.2014.11.004.

Rap, A., Scott, C., Reddington, C., Mercado, L., Ellis, R., Garraway, S., et al., 2018. Enhanced global primary production by biogenic aerosol via diffuse radiation fertilization. Nat. Geosci. 11, 640–644. https://doi.org/10.1038/s41561-018-0208-3.

Reichstein, M., Falge, E., Baldocchi, D., Papale, D., Aubinet, M., Berbigier, P., et al., 2005. On the separation of net ecosystem exchange into assimilation and ecosystem respiration: review and improved algorithm. Glob. Chang. Biol. 11, 1424–1439. https://doi.org/10.1111/j.1365-2486.2005.001002.x.

Reichstein, M., Papale, D., Valentini, R., Aubinet, M., Bernhofer, C., Knohl, A., et al., 2007. Determinants of terrestrial ecosystem carbon balance inferred from European eddy covariance flux sites. Geophys. Res. Lett. 34. https://doi.org/10.1029/2006GL027880.

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., et al., 2019. Deep learning and process understanding for data-driven earth system science. Nature 566, 195–204. https://doi.org/10.1038/s41586-019-0912-1.

Rienecker, M.M., Suarez, M.J., Gelaro, R., Todling, R., Bacmeister, J., Liu, E., et al., 2011. MERRA: NASA's modern-era retrospective analysis for research and applications. J. Clim. 24, 3624–3648. https://doi.org/10.1175/JCLI-D-11-00015.1.

Savitzky, A., Golay, M.J., 1964. Smoothing and differentiation of data by simplified least squares procedures. Anal. Chem. 36, 1627–1639. https://doi.org/10.1021/ac60214a047.

Scanlon, B.R., Zhang, Z., Save, H., Sun, A.Y., Müller Schmied, H., Van Beek, L.P., et al., 2018. Global models underestimate large decadal declining and rising water storage trends relative to GRACE satellite data. Proc. Natl. Acad. Sci. U. S. A. 115, E1080–E1089. https://doi.org/10.1073/pnas.1704665115.

Schmid, H., 1997. Experimental design for flux measurements: matching scales of observations and fluxes. Agric. For. Meteorol. 87, 179–200. https://doi.org/10.1016/S0168-1923(97)00011-7.

Stocker, B.D., Zscheischler, J., Keenan, T.F., Prentice, I.C., Seneviratne, S.I., Peñuelas, J., 2019. Drought impacts on terrestrial primary production underestimated by satellite monitoring. Nat. Geosci. 12, 264–270. https://doi.org/10.1038/s41561-019-0318-6.

Sun, J., Zhang, Y., Wu, Z., Zhu, Y., Yin, X., Ding, Z., et al., 2019. An efficient and scalable framework for processing remotely sensed big data in cloud computing environments. IEEE Trans. Geosci. Electron. 57, 4294–4308. https://doi.org/10.1109/TGRS.2018.2890513.

Surhone, L.M., Tennoe, M.T., Henssonow, S.F., Breiman, L., 2010. Random forest. Mach. Learn. 45, 5–32.

Tapley, B.D., Watkins, M.M., Flechtner, F., Reigber, C., Bettadpur, S., Rodell, M., et al., 2019. Contributions of GRACE to understanding climate change. Nat. Clim. Chang. 9, 358–369. https://doi.org/10.1038/s41558-019-0456-2.

Tramontana, G., Jung, M., Schwalm, C.R., Ichii, K., Camps-Valls, G., Ráduly, B., et al., 2016. Predicting carbon dioxide and energy fluxes across global FLUXNET sites with regression algorithms. Biogeosciences 13, 4291–4313. https://doi.org/10.5194/bg-13-4291-2016.

Tramontana, G., Migliavacca, M., Jung, M., Reichstein, M., Keenan, T.F., Camps-Valls, G., et al., 2020. Partitioning net carbon dioxide fluxes into photosynthesis and respiration using neural networks. Glob. Chang. Biol. 26, 5235–5253. https://doi.org/10.1111/gcb.15203.

Wang, L., Zhu, H., Lin, A., Zou, L., Qin, W., Du, Q., 2017. Evaluation of the latest MODIS GPP products across multiple biomes using global eddy covariance flux data. Remote Sens. 9, 418. https://doi.org/10.3390/rs9050418.

Wang, J., Song, C., Reager, J.T., Yao, F., Famiglietti, J.S., Sheng, Y., et al., 2018a. Recent global decline in endorheic basin water storages. Nat. Geosci. 11, 926–932. https://doi.org/10.1038/s41561-018-0265-7.

Wang, S., Ibrom, A., Bauer-Gottwein, P., Garcia, M., 2018b. Incorporating diffuse radiation into a light use efficiency and evapotranspiration model: an 11-year study in a high latitude deciduous forest. Agric. For. Meteorol. 248, 479–493. https://doi.org/10.1016/j.agrformet.2017.10.023.

Wang, S., Zhang, Y., Ju, W., Chen, J.M., Ciais, P., Cescatti, A., et al., 2020. Recent global decline of CO2 fertilization effects on vegetation photosynthesis. Science 370, 1295–1300. https://doi.org/10.1126/science.abb7772.

Wang, S., Zhang, Y., Ju, W., Qiu, B., Zhang, Z., 2021. Tracking the seasonal and inter-annual variations of global gross primary production during last four decades using satellite near-infrared reflectance data. Sci. Total Environ. 755, 142569. https://doi.org/10.1016/j.scitotenv.2020.142569.

Wei, S., Fang, H., Schaaf, C.B., He, L., Chen, J.M., 2019. Global 500 m clumping index product derived from MODIS BRDF data (2001–2017). Remote Sens. Environ. 232, 111296. https://doi.org/10.1016/j.rse.2019.111296.

Wu, C., Munger, J.W., Niu, Z., Kuang, D., 2010. Comparison of multiple models for estimating gross primary production using MODIS and eddy covariance data in Harvard Forest. Remote Sens. Environ. 114, 2925–2939. https://doi.org/10.1016/j.rse.2010.07.012.

Wu, X., Ju, W., Zhou, Y., He, M., Law, B.E., Black, T.A., et al., 2015. Performance of linear and nonlinear two-leaf light use efficiency models at different temporal scales. Remote Sens. 7, 2238–2278. https://doi.org/10.3390/rs70302238.

Xie, X., Li, A., Jin, H., Yin, G., Nan, X., 2018. Derivation of temporally continuous leaf maximum carboxylation rate (Vcmax) from the sunlit leaf gross photosynthesis productivity through combining BEPS model with light response curve at tower flux sites. Agric. For. Meteorol. 259, 82–94. https://doi.org/10.1016/j.agrformet.2018.04.017.

Xie, X., Li, A., Tan, J., Lei, G., Jin, H., Zhang, Z., 2020. Uncertainty analysis of multiple global GPP datasets in characterizing the lagged effect of drought on photosynthesis. Ecol. Indic. 113, 106224. https://doi.org/10.1016/j.ecolind.2020.106224.

Xu, X., Du, H., Fan, W., Hu, J., Mao, F., Dong, H., 2019. Long-term trend in vegetation gross primary production, phenology and their relationships inferred from the FLUXNET data. J. Environ. Manag. 246, 605–616. https://doi.org/10.1016/j.jenvman.2019.06.023.

Xu, X., Jiao, F., Liu, H., Gong, H., Zou, C., Lin, N., et al., 2022. Persistence of increasing vegetation gross primary production under the interactions of climate change and land use changes in Northwest China. Sci. Total Environ. 834, 155086. https://doi.org/10.1016/j.scitotenv.2022.155086.

Yin, S., Jiao, Z., Dong, Y., Zhang, X., Cui, L., Xie, R., et al., 2022. Evaluation of the consistency of the vegetation clumping index retrieved from updated MODIS BRDF data. Remote Sens. 14, 3997. https://doi.org/10.3390/rs14163997.

Yuan, W., Zheng, Y., Piao, S., Ciais, P., Lombardozzi, D., Wang, Y., et al., 2019. Increased atmospheric vapor pressure deficit reduces global vegetation growth. Sci. Adv. 5, eaax1396. https://doi.org/10.1126/sciadv.aax1396.

Yuval, J., O`Gorman, P.A., 2020. Stable machine-learning parameterization of subgrid processes for climate modeling at a range of resolutions. Nat. Commun. 11, 3295. https://doi.org/10.1038/s41467-020-17142-3.

Zakeri, F., Mariethoz, G., 2021. A review of geostatistical simulation models applied to satellite remote sensing: Methods and applications. Remote Sens. Environ. 259, 112381. https://doi.org/10.1016/j.rse.2021.112381.

Zhang, Y., Song, C., Sun, G., Band, L.E., Noormets, A., Zhang, Q., 2015a. Understanding moisture stress on light use efficiency across terrestrial ecosystems based on global flux and remote-sensing data. Eur. J. Vasc. Endovasc. Surg. 120, 2053–2066. https://doi.org/10.1002/2015JG003023.

Zhang, Y., Song, C., Sun, G., Band, L.E., Noormets, A., Zhang, Q., 2015b. Understanding moisture stress on light use efficiency across terrestrial ecosystems based on global flux and remote-sensing data. Eur. J. Vasc. Endovasc. Surg. 120, 2053–2066. https://doi.org/10.1002/2015JG003023.

Zhang, Y., Kong, D., Gan, R., Chiew, F.H., McVicar, T.R., Zhang, Q., et al., 2019. Coupled estimation of 500 m and 8-day resolution global evapotranspiration and gross primary production in 2002–2017. Remote Sens. Environ. 222, 165–182. https://doi.org/10.1016/j.rse.2018.12.031.

Zhao, M., Running, S.W., 2010. Drought-induced reduction in global terrestrial net primary production from 2000 through 2009. Science 329, 940–943. https://doi.org/10.1126/science.1192666.

Zhao, M., Running, S.W., Nemani, R.R., 2006. Sensitivity of Moderate Resolution Imaging Spectroradiometer (MODIS) terrestrial primary production to the accuracy of meteorological reanalyses. Eur. J. Vasc. Endovasc. Surg. 111, G01002. https://doi.org/10.1029/2004JG000004.

Zhao, M., Geruo, A., Velicogna, I., Kimball, J.S., 2017. Satellite observations of regional drought severity in the continental United States using GRACE-based terrestrial water storage changes. J. Clim. 30, 6297–6308. https://doi.org/10.1175/JCLI-D-16-0458.1.

Zhao, J., Feng, H., Xu, T., Xiao, J., Guerrieri, R., Liu, S., et al., 2021. Physiological and environmental control on ecosystem water use efficiency in response to drought across the northern hemisphere. Sci. Total Environ. 758, 143599. https://doi.org/10.1016/j.scitotenv.2020.143599.

Zheng, Y., Shen, R., Wang, Y., Li, X., Liu, S., Liang, S., et al., 2020. Improved estimate of global gross primary production for reproducing its long-term variation, 1982–2017. Earth Syst. Sci. Data 12, 2725–2746. https://doi.org/10.5194/essd-12-2725-2020.

Zheng, Y., Zhao, W., Chen, A., Chen, Y., Chen, J., Zhu, Z., 2024. Vegetation canopy structure mediates the response of gross primary production to environmental drivers across multiple temporal scales. Sci. Total Environ. 917, 170439. https://doi.org/10.1016/j.scitotenv.2024.170439.

Zhou, Y., Wu, X., Ju, W., Chen, J.M., Wang, S., Wang, H., et al., 2016. Global parameterization and validation of a two-leaf light use efficiency model for predicting gross primary production across FLUXNET sites. Geophys. Res.: Biogeosci. 121, 1045–1072. https://doi.org/10.1002/2014JG002876.