Ecological Informatics 66 (2021) 101479

Contents lists available at ScienceDirect

Ecological Informatics

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

Optimization of environmental variable functions of GPP quantitative model based on SCE-UA and minimum loss screening method

Lin Zhang [a], Tianwei Ren [a], Yaoqi Yu [a], Yuan Yao [a], Cheng Li [b], Yuanyuan Zhao [a,c], Qianlai Zhuang [d,e,f], Zhe Liu [a,c,*], Xiaodong Zhang [a,c], Shaoming Li [a,c]

[a] College of Land Science and Technology, China Agricultural University, Beijing 100083, China
[b] College of Information and Electrical Engineering, China Agricultural University, Beijing 100083, China
[c] Key Laboratory of Remote Sensing for Agri-Hazards, Ministry of Agriculture and Rural Affairs, Beijing 100083, China
[d] Department of Earth, Atmospheric, Planetary Sciences, Purdue University, West Lafayette, IN 47907, USA
[e] Department of Agronomy, Purdue University, West Lafayette, IN 47907, USA
[f] Purdue Climate Change Research Center, USA

ARTICLE INFO

Keywords:
Gross primary productivity (GPP)
Model discovery
SCE-UA algorithm
Minimum loss screening method
Parameter optimization

ABSTRACT

Environmental variable functions are of key importance for gross primary productivity (GPP) modeling. This study proposes a method about the optimization of environmental variable function to obtain a more robust and accurate GPP quantitative model. The key idea is to explore the impact of environmental factors on the accuracy of the GPP quantitative model from the following three aspects: the first is using tensor as the alternative environmental factor equation to construct the basis-function set of photosynthetically active radiation (PAR), atmospheric temperature and atmospheric carbon dioxide concentration for the given environmental conditions, and soil moisture functions of new environmental conditions. The second is building 144 candidate model based on a tensor product. The third is finding the best model from the candidates according to the Shuffled Complex Evolution (SCE-UA) algorithm and the Minimum Loss Screening Method. Through the above experiments, we have the following conclusions: First, this paper obtains two new best models from 144 candidate models, and their accuracy is higher than that of the initial model, indicating that this paper proposes a more robust and accurate GPP quantitative model. Then, the model proposed in this paper has common characteristics, that is, PAR and atmospheric temperature can be replaced by more appropriate quantitative functions, named Sigmoid-like function and Q10 equation, and the carbon dioxide equation can use half-saturated equation or Sigmoid function. Finally, the method in this paper can provide new ideas for simulating the fluxes of other ecosystems, including soil carbon decomposition and plant respiration.

1. Introduction

Terrestrial gross primary productivity (GPP) is the total photosynthetic uptake or carbon assimilation by plant and is a key component in terrestrial carbon cycle (Schaefer et al., 2012). The carbon absorption by vegetation depends on climate variability, historical climate disturbance, the utilization of water and nutrient, soil type, species composition and community structure. GPP is a measure of the carbon intake by vegetation, normal observed by eddy covariance flux tower. If the simulated GPP is too low or too high, then the predicted crop yield, leaf area index, wood biomass, and soil biomass may also be inaccurate (Schaefer et al., 2008). Therefore, improving the accuracy of GPP simulation is an urgent problem to be solved in the fields of terrestrial ecosystems, ecological environment, and agriculture.

Since the large-area GPP estimation is limited by the spatial and temporal location of the flux tower, many scholars have developed GPP quantitative models for long-term estimation of regional and global GPP. The GPP quantitative model is divided into process model (PM), enzyme kinetics (EK) model and light utilization efficiency (LUE) model (Schaefer et al., 2012). The PM model simulates the carbon cycle process of the entire life cycle of the vegetation, uses the statistical relationship between the observed environmental conditions and the GPP eddy covariance flux data, and then uses various reanalysis weather products to extend it to the regional or global scale (Beer et al., 2010). The EK

* Corresponding author at: College of Land Science and Technology, China Agricultural University, Beijing 100083, China.
E-mail address: liuz@cau.edu.cn (Z. Liu).

https://doi.org/10.1016/j.ecoinf.2021.101479
Received 26 July 2021; Received in revised form 27 October 2021; Accepted 1 November 2021
Available online 10 November 2021

model quantifies GPP through stomatal water loss under enzyme kinetics and stomatal conductance balance on the leaf scale (Collatz et al., 1991; Collatz et al., 1992). Most of these stomatal conductance models are based on the empirical correlation between conductance, photosynthesis, and relative humidity (Ball et al., 1987) or insufficient vapor pressure (VPD) (Wang et al., 1998). The LUE model uses photosynthetically active radiation (PAR), the remote-sensing PAR fraction absorbed by vegetation (fPAR), and biomass conversion factors (commonly referred to as light utilization efficiency) to estimate GPP (Field et al., 1995; Goetz et al., 1999; Landsberg and Waring, 1997; Monteith, 1972; Prince and Goward, 1995; Running et al., 2000; Running et al., 2004; Heinsch et al., 2003). Schaefer et al. (2012) believe that there is a difference between the GPP estimated by these three models and the GPP observed in the field. Through experiments, they found that among the EK models, the three best-performing models are DLEM (Tian et al., 2010), SIB (Baker et al., 2008) and ISOLSM (Riley et al., 2002), and RMSE is about 2.0 umol C m^{-2} s^{-1}. However, the worst-performing model is CNCLASS (Arain et al., 2006), with an estimated RMSE of 4.5 umol C m^{-2} s^{-1}. When the estimation scale of the EK model is set to the monthly or annual scale, the error will be magnified to a greater extent. The error of LUE model in GPP estimation is higher than that of EK model. The RMSE of the two best-performing models ISAM (Yang et al., 2009) and MODIS_5.1 (Heinsch et al., 2003) in GPP estimation is about 2.5 umol C m^{-2} s^{-1}, while the RMSE of the worst-performing model DNDC (Li et al., 2010) can even reach more than 5.5 umol C m^{-2} s^{-1}. On average, these two models overestimate GPP in winter, spring, and autumn, and underestimate it in summer.

Any errors in the simulation of GPP by the above two models will be propagated in other models, thereby introducing errors in simulating other biomass and fluxes. Since PM models tend to be part of large-scale ecosystem models, such as CESM model (Hurrell et al., 2013), CLM model (Bonan et al., 2002), and TEM model (McGuire et al., 1992; Raich et al., 1991). The error in the GPP estimation is easily brought into the net ecosystem exchange (NEE) and the total ecosystem respiration (Schaefer et al., 2012), so the propagation of GPP error is more pronounced in the PM. Therefore, this study focuses on minimizing the error of PM through optimizing the environment variable functions.

PM modeling method belongs to mechanism modeling method in process modeling (Beer et al., 2010). According to the mechanism process of the ecosystem, reasonable assumptions are put forward for the equations in the model, and then the rationality of the model is evaluated through speculation, deduction, statistical analysis and verification (Hoyle, 1995). In the process of modeling, assumptions are put forward mainly based on personal experience. Therefore, the environmental variable equation of PM is usually obtained through trial and error based on experience and experimental data (Beer et al., 2010; Sun et al., 2017). However, this trial-and-error method usually takes a lot of time to verify the hypothesis step by step, which is inefficient, and the results often have great uncertainties.

As an example, Terrestrial Ecosystem Model (TEM) is a process-based biogeochemical model, which has been widely used to quantify GPP (Hayes et al., 2014; McGuire et al., 1992; Raich and Schlesinger, 1992; Zhuang et al., 2001; Zhuang et al., 2002; Zhuang et al., 2010; Zhuang et al., 2011; Zhuang et al., 2013). GPP is originally defined as a function of the irradiance of PAR, atmospheric CO2 concentrations, moisture availability, air average temperature, the relative photosynthetic capacity of the vegetation and nitrogen availability (Raich and Schlesinger, 1992). Zhuang et al. (2002) has found through hypothesis, derivation and verification that the ratio of vegetation canopy leaf biomass to the maximum canopy leaf biomass is also an important factor in quantifying GPP. Subsequently, Zhuang et al. (2011) has found that adding the freeze-thaw index of the following month can better quantify the influence of the freeze-thaw dynamics on GPP. The case shows that a reasonable model optimization is necessary, at meanwhile, manual check of the model is not efficient enough. In addition, due to the mutual constraints of various environmental conditions, the functions expressed by various environmental variables are likely to have the best combination when quantifying GPP. This indicates that the environmental variable equations of other processes in the ecosystem model can be used as a reference for the equations of the same environmental variables in the GPP model, thereby providing a possibility to improve or develop a new GPP model.

In recent years, the growth in computing power of multi-core processors has made high-efficient and automatic physical discovery possible (Zhang and Lin, 2018). This makes it possible for researchers to develop new GPP models with the help of high performance computing. The matrixization of models and the introduction of tensor calculations can solve the problem of establishing a high-dimensional matrixed candidate model set under high performance computing. Therefore, this paper applied high performance computing and machine learning in building large-scale GPP candidate model set and model optimizing.

With the help of high performance computing cluster, this paper proposed a method to automatically optimize GPP quantitative model using flux data and various environmental data. First, a basis function set is established to collect a large number of environmental curve functions that may affect GPP quantification, namely, environmental variable functions. Then, tensor product is used for the basis-function set to obtain 144 candidate models, and optimize the parameters of all candidate models based on the SCE-UA algorithm (Kan et al., 2016). After obtaining the optimal parameters of all candidate models, by setting thresholds, using the Minimum Loss Screening Method, candidate models that perform well in training and testing are screened. Finally, through cross-validation experiments, the model with high stability is obtained and regarded as the best model.

2. Material and methods

2.1. Study area and materials

Harvard Forest is located on the outskirts of Boston, Massachusetts, USA, it is one of the oldest forest-atmosphere carbon exchange study areas in North America. This paper selects the EMS vortex flux tower of Harvard Forest (longitude: −72.171478, latitude: +42.537755, altitude: 340 m), which was installed in 1989 and its eddy current measurement results constitute the longest record of net ecosystem exchange (NEE) in North American forest. Based on the long-term records of NEE, the long-term effects of climate disturbances on carbon flux can be further determined. Climate, soil moisture, and vegetation data are also used to assess the monthly GPP of deciduous broad-leaved forests (Munger and Wofsy, 1999). This tower has also been used to compare a variety of GPP quantitative models (Wu et al., 2010), and the schematic diagram of the study area is shown in Fig. 1.

In the study area, we obtain the PAR (unit: μmol m^{-2} s^{-1}) at a height of 28 m above the ground from the EMS vortex flux tower, and add the data to a monthly scale value (unit: μmol m^{-2} $month^{-1}$), the monthly average atmospheric temperature of the plant canopy is also obtained at a height of 27.9 m above the ground (unit: °C). In addition, the atmospheric carbon dioxide concentration data (unit: ppm) is obtained at 29 m off the ground, and the default value of carbon dioxide data is filled with the global measurement data observed by the US NOAA (https://www.esrl.noaa.gov/gmd/ccgg/trends/), and finally data accumulated into monthly average carbon dioxide. In this paper, through the EMS tower, the net radiation of vegetation canopy, daily minimum temperature, daily maximum temperature, hourly average temperature, air relative humidity, daily sunshine hours, wind speed of canopy, as well as longitude, latitude and altitude data of the station are summarized. Then the average monthly evapotranspiration EET (unit: mm) is calculated by Penman equation (Allen et al., 1998). Because it is impossible to obtain continuous soil moisture content data from the EMS tower, this paper uses the soil moisture data of Barre Woods soil warming experiment in Harvard Forest, which is the relative soil moisture value measured by the No. 6 TDR probe in the experimental

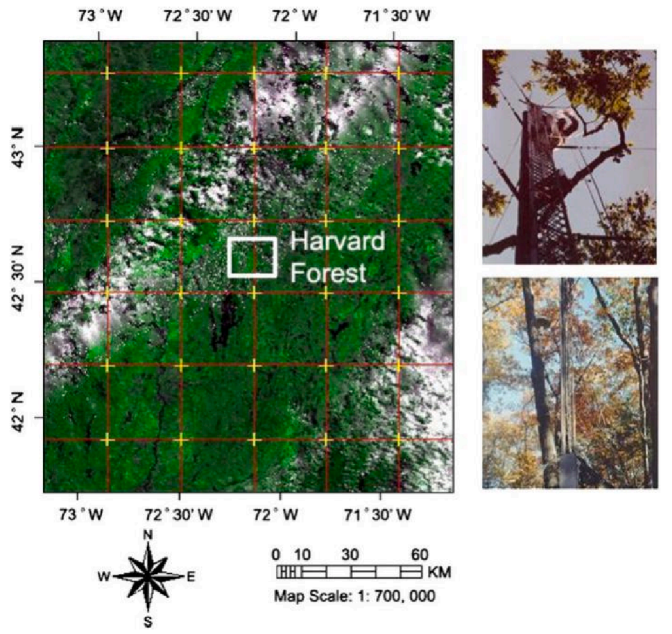

Fig. 1. Study area, where the left shows the geographical location (longitude: −72.171478, latitude: +42.537755) of EMS vortex flux tower, and the right is the real picture.

control area at a depth of 50 cm from the ground. Finally, we summarized the GPP observations per second measured by the EMS tower to the monthly scale. Under the 3*3 grid window around the EMS tower, the 8-day GPP measured by the MOD17A2H algorithm in the MODIS Terra and MODIS Aqua satellites are summarized to the monthly scale.

2.2. Initial GPP model

The Terrestrial Ecosystem Model (TEM) (Hayes et al., 2014; McGuire et al., 1992; Raich and Schlesinger, 1992; Zhuang et al., 2001; Zhuang et al., 2002; Zhuang et al., 2010; Zhuang et al., 2011; Zhuang et al., 2013) has been widely used to quantify GPP. As a biogeochemical model, this model is also widely used in the study of terrestrial ecosystems in China (Hao, 2015; Li et al., 2016). In this model, GPP was first defined in detail as the model of PAR, atmospheric carbon dioxide (CO_2) concentration, water availability, average temperature, relative photosynthetic capacity of vegetation and nitrogen availability (Raich and Schlesinger, 1992). In this paper, the latest GPP quantization method of TEM is adopted:

$$GPP = Cmax*f(PAR)*f(P)*f(F)*f(T)*f(C)*f(NA) \tag{1}$$

Where *Cmax* represents the maximum monthly carbon uptake by vegetation under ideal conditions (McGuire et al., 1992), *f*(*PAR*) is the influence equation of PAR on vegetation carbon absorption (McGuire et al., 1992; Raich and Schlesinger, 1992). *f*(*P*) represents the monthly leaf area relative to the leaf area during the maximum leaf area month and depends on the estimated monthly evapotranspiration (Raich et al., 1991). *f*(*F*) is a scalar function ranging from 0.0 to 1.0, representing the ratio of canopy leaf biomass to the maximum leaf biomass (Zhuang et al., 2002). *f*(*T*) represents the influence equation of monthly mean atmospheric temperature on vegetation carbon absorption (Zhuang et al., 2004). *f*(*C*) represents the effect of elevated atmospheric CO_2 concentration on the absorption of CO_2 by plant canopy cells (McGuire et al., 1997; Pan et al., 1998). *NA* is nitrogen availability, and *f*(*NA*) simulates the limiting effect of plant nitrogen status on vegetation carbon uptake (Pan et al., 1998). All functions are as follows:

$$f(PAR) = \frac{PAR}{k_i + PAR} \tag{2}$$

Where k_i is the irradiance parameter at half of the maximum carbon absorption rate.

$$f(P) = KLEAF_j \tag{3}$$

$$KLEAF_j = \mathrm{a}^*\left(\frac{EET_j}{EET_{max}}\right) + b^*\left(KLEAF_{j-1}\right) + c \tag{4}$$

$$KLEAF_j = 1.0 \text{ if } KLEAF_j > 1.0 \tag{5}$$

$$KLEAF_j = \frac{KLEAF_t}{KLEAF_{max}} \text{ if } KLEAF_{max} < 1.0 \tag{6}$$

$$KLEAF_j = min \text{ if } KLEAF_j < min \tag{7}$$

Where $KLEAF_j$ indicates the relative change of the light conversion capacity of mature vegetation based on the estimated evapotranspiration (EET) and the light conversion capacity of the previous month. The time step *j* represents the month. EET_j is the largest EET generated in any month during the historical period. The a, *b* and *c* are regression parameters. The *min* represents a preset minimum value of *KLEAF*, and $KLEAF_{max}$ is the maximum *KLEAF* in the historical period.

$$f(F) = \frac{1.0}{1.0 + m_1{}^*e^{m_2{}^*\sqrt{f(C_v)}}}, f(C_v) = \frac{m_3{}^*C_v}{1.0 + m_4{}^*C_v} \tag{8}$$

Where m_1, m_2, m_3, and m_4 are the parameters of *f*(*F*) similar to the Sigmoid function. $f(C_v)$ is the hyperbolic function of the state variable for vegetation carbon (C_v).

$$f(T) = \frac{(T - T_{min})^*(T - T_{max})}{[(T - T_{min})^*(T - T_{max})] - (T - T_{opt})^2} \tag{9}$$

Where *T* represents the monthly average temperature of the vegetation canopy, T_{min}, T_{max} and T_{opt} represents the minimum, maximum and optimal temperature at which the vegetation absorbs carbon.

$$f(C) = \frac{\mathrm{C}_i}{kc + \mathrm{C}_i} \tag{10}$$

$$\mathrm{C}_i = \mathrm{C}_a{}^*\mathrm{G}_v \tag{11}$$

$$\mathrm{G}_v = 0.10 + \left(0.9^*\frac{EET}{PET}\right) \tag{12}$$

Where *kc* is the parameter at half of the maximum rate of carbon absorption. C_i is the CO_2 concentration in the leaf. C_a is the concentration of CO_2 in the atmosphere. G_v is a unitless multiplier, which explains the change in the electrical conductivity of the leaf to CO_2 due to the change in water using efficiency. *EET* is the actual evapotranspiration and *PET* is the potential evapotranspiration. Here it is assumed that actual evapotranspiration is equal to potential evapotranspiration.

$$f(NA) = 1 \tag{13}$$

In this paper, it is assumed that the availability of vegetation nitrogen fully satisfies the carbon uptake by vegetation.

2.3. Constructing GPP candidate models with different environment variable functions

In this paper, the required candidate models are increase by constructing the basis-function set. The advantage of this method is that it can carry out a wide range of automated trial-and-error experiments. Even if there are unreasonable basis-functions, it can be eliminated by the Minimum Loss Screening Method, and at least the initial model after global parameter optimization can be obtained. The construction of the basis-function set mainly includes three steps: one is to construct the existing environmental variable basis-function set; the other is to construct a new environmental variable basis-function set; the third is to construct a candidate model set through the tensor product.

2.3.1. Building a basis-function set of existing environment conditions

This paper assumes that in the GPP model, the existing environmental variable functions are not the best, and there are some environmental variable functions that can better fit the relationship between environmental variables and vegetation carbon uptake. The formula constructed is as follows:

$$F(x) = [f(x)\, f_1(x)\, f_2(x) \ldots\ldots, f_n(x)\,]^T \tag{14}$$

Where x is an environment variable that already exists in the initial model. The more rational the base functions are created in the set, the more likely it is to find environmental variable functions suitable for quantifying GPP.

When constructing the basis-function set of PAR. Firstly, it is assumed that the solar radiation intensity is optimal when the photosynthetic efficiency of vegetation is the highest, and the decrease or increase of the solar radiation intensity on this basis will inhibit the photosynthesis of vegetation to some extent. According to this assumption, the basis-function is constructed as shown in $f_1(PAR)$. Then, we continue to assume that when the amount of PAR is too low or too high, its change will have a smaller impact on vegetation photosynthesis than when the amount of radiation is appropriate. Sigmoid function is a common S-shaped function in biology (Han and Moraga, 1995), which can perform data smoothing, compression and normalization. The slow monotone increasing characteristic of this function has been verified to have a high similarity with the change of ecological environment mechanism (Zhuang et al., 2002). So we can try to use the Sigmoid function as the basis-function of PAR, as shown in $f_2(PAR)$. Therefore, we can construct the basis function set of PAR: $F(PAR) = [f(PAR)\, f_1(PAR)\, f_2(PAR)]^T$.

$$f_1(PAR) = \frac{(PAR - PAR_{min})^*(PAR - PAR_{max})}{[(PAR - PAR_{min})^*(PAR - PAR_{max})\,] - (PAR - PAR_{opt})^2} \tag{15}$$

$$f_2(PAR) = \frac{1.0}{1.0 + p1^* e^{p2^*\sqrt{f(p)}}}, f(p) = \frac{p3^*PAR}{1.0 + p4^*PAR} \tag{16}$$

Where PAR represents the photosynthetically active radiation in the vegetation canopy, PAR_{min} and PAR_{max} represent the lowest and the highest of photosynthetic radiation required for the vegetation to absorb carbon respectively, and PAR_{opt} represents the optimal photosynthetically effective radiation.

When constructing the basis-function set for the atmospheric temperature. First of all, it can be expressed by $f_1(T)$. Then, when the temperature is too low or too high, the effect of its change on vegetation photosynthesis is less than that at the appropriate temperature, so we try to use the Sigmoid-like function as the basis of T, as shown in $f_2(T)$. In addition, we also try to use the half-saturated function as the basis-function of T, as shown in formula $f_3(T)$. Therefore, we can construct the basis-function set of T: $F(T) = [f(T)\, f_1(T)\, f_2(T)\, f_3(T)]^T$.

$$f_1(T) = Q_{10}^{\frac{T(canopy) - Tr}{10}} \tag{17}$$

$$f_2(T) = \frac{1.0}{1.0 + t1^* e^{t2^*\sqrt{f(t)}}}, f(t) = \frac{t3^*T}{1.0 + t4^*T} \tag{18}$$

$$f_3(T) = \frac{T}{kt + T} \tag{19}$$

Where Q_{10} (Walter and Heimann, 2000; Zhuang et al., 2004) is similar to formula (9) and is the sensitivity of ecosystem respiration to temperature, that is, the multiple of the increase in respiration rate for every 10 °C increase in temperature. $T(canopy)$ represents the temperature of the vegetation canopy, Tr is the parameter of Q_{10}.

Similar to PAR and T, we can try the following function $f_1(C)$ and $f_2(C)$ as the basis-function of CO_2. Therefore, we can construct the basis-function set of CO_2: $F(C) = [f(C)\, f_1(C)\, f_2(C)]^T$.

$$f_1(C) = \frac{1.0}{1.0 + c1^* e^{c2^*\sqrt{f(c)}}}, f(c) = \frac{c3^*CO_2}{1.0 + c4^*CO_2} \tag{20}$$

$$f_2(C) = \frac{(CO_2 - CO_2min)^*(CO_2 - CO_2max)}{[(CO_2 - CO_2min)^*(CO_2 - CO_2max)\,] - (CO_2 - CO_2opt)^2} \tag{21}$$

2.3.2. Building a basis-function set of new environment conditions

The search for new models can be derived from known physical laws or based on empirical observations of physical behaviour (Zhang and Lin, 2019). In the previous section, only the existing environmental conditions were considered. Therefore, next we try to explore new environmental variables to find a more suitable quantitative model. The formula is as follows:

$$F(y) = [1\, f(y)\, f_1(y)\, f_2(y) \ldots\ldots, f_n(y)\,]^T \tag{22}$$

Where y is the new environment variable. Here, vector 1 needs to be added to the basis function set to preserve the original state of the model. Therefore, each set of basic functions used for new environment variables contains at least two elements. This paper uses this method to construct a new set of functions.

The $f(C)$ in formula (10) indicates that the leaf conductivity can represent the limit of CO_2 absorption of water (McGuire et al., 1992; Raich et al., 1991; Wang et al., 2018). However, the direct influence of water on GPP has not been reflected in formula (1), so we establish the function of the influence of soil moisture on vegetation carbon absorption to better represent this phenomenon. We assume that the effect of soil moisture on GPP is similar to the effect of soil moisture on decomposition of soil organic carbon (Tian et al., 1999) and methane oxidation (Zhuang et al., 2004), and then the basis function $f(SM)$ of soil moisture is obtained. In addition, try half-saturated function and Sigmoid-like function to get$f_1(SM)$ and $f_2(SM)$. Finally, this study constructs the basis-function set of SM: $F(SM) = [1\, f(SM)\, f_1(SM)\, f_2(SM)]^T$.

$$f(SM) = \frac{(SM - SMmin)^*(SM - SMmax)}{[(SM - SMmin)^*(SM - SMmax)\,] - (SM - SMopt)^2} \tag{23}$$

$$f_1(SM) = \frac{SM}{ks + SM} \tag{24}$$

$$f_2(SM) = \frac{1.0}{1.0 + m1^* e^{m2^*\sqrt{f(S)}}}, f(S) = \frac{m3^*SM}{1.0 + m4^*SM} \tag{25}$$

Where $SMmin$, $SMmax$ and $SMopt$ are the minimum, maximum and optimal soil moisture for carbon absorption by vegetation.

2.3.3. Building all GPP model candidates

All candidate GPP models are constructed by tensor product in this paper, as shown in Formula (26) and (27):

$$D = F(PAR) \otimes F(T) \otimes F(CO_2) \otimes F(SM) \otimes F(R)^T \tag{26}$$

$$F(R) = Cmax^*[f(P)\,] \otimes [f(F)\,] \otimes [f(NA)\,] \tag{27}$$

We use H to represent the number of functions in the basis-function set, and use M to represent the number of candidate models. Then, the number of candidate models of GPP can be obtained is represented by formula (28):

$$\mathrm{M} = H_{PAR}{}^*H_T{}^*H_{CO_2}{}^*H_{SM}{}^*H_R \tag{28}$$

According to 2.3.1 and 2.3.2, it can be found that the number of functions of the five basis function sets PAR, T, CO_2, SM and R are 3, 4, 3, 4 and 1 respectively, so the number of candidate models constituted is 144.

2.4. Optimizing environment variable functions

Screening optimal environment variable functions from many can-

didates often requires multiple assessment criteria. Machine learning divides datasets into training and testing sets in modeling and knowledge discovery (Liu et al., 2016a, 2016b, 2017). Here, the Minimum Loss Screening Method is used to judge the performance of each candidate by observing and comparing the training and testing loss value. The candidates with good performance based on the loss thresholds are then identified. The loss function of the minimization is:

$$loss = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(GPP_{obs,i} - GPP_{candidate,i}\right)^2} \quad (29)$$

Where $GPP_{obs,\ i}$ and $GPP_{candidate,\ i}$ are the observations and simulations from every GPP candidate model. In the loss function, N is the number of data pairs for comparison. So, loss is one kind of root mean square error (RMSE). The steps of the Minimum Loss Screening Method are as follows:

Step 0. Initialize. Select j = 0, $\boldsymbol{O} = \{\}$, $\boldsymbol{U} = \{\}$. Input all candidates $\boldsymbol{D}$, initial values of all parameters $\boldsymbol{C}$ for each candidate, training the loss threshold L_t and testing the loss threshold L_e, where $\boldsymbol{D} = \{D_i, i = 1, \ldots, S\}$, $\boldsymbol{C} = \{\boldsymbol{C}_i, i = 1, \ldots, S\}$, S represents the number of candidates, D_i represents the i-th candidate, and $\boldsymbol{C}_i$ represents the set of initial values of all the parameters of the i-th candidate.

Step 1. Model training. j++, select D_j and $\boldsymbol{C}_j$, then update $\boldsymbol{C}_j$ with SCE-UA until satisfying termination condition.

Step 2. Model primary screening. Run D_j with $\boldsymbol{C}_j$ from Step 1. Check the loss value after model running. If the loss value $< L_t$, next; otherwise, return to Step 1.

Step 3. Model secondary screening. Test D_j with $\boldsymbol{C}_j$. Check the loss value after testing. If the loss value $< L_e$, add D_j into the $\boldsymbol{O}$ and add $\boldsymbol{C}_j$ into the $\boldsymbol{U}$, then next; otherwise, return to Step 1.

Step 4. Update $\boldsymbol{O}$ and $\boldsymbol{U}$. If $j < S$, return to Step 1; otherwise, output $\boldsymbol{O}$ and $\boldsymbol{U}$.

Dividing the dataset into different proportions of the training and evaluation sets can prevent overfitting in the experiment; it can also be used for observing the method's generalization ability for new datasets (Kohavi, 1995; Liu and Cocea, 2017). Therefore, different training and test data proportion sets are used for each experiment, the results of each experiment are then summarized. The optimal state is that some candidates appear simultaneously in each group of experimental results, indicating that these candidates have high stability and will not change with the change of dataset. We take the intersection of the sets of all the experimental results as the final GPP quantification models:

$$\boldsymbol{Final} = \boldsymbol{O}_1 \cap \boldsymbol{O}_2 \cap \ldots \cap \boldsymbol{O}_I \quad (30)$$

Where $\boldsymbol{O}_i$ represents the result of i-th experiment. I is the total number of experiments.

When no candidates appear simultaneously in each set of experimental results, the size of thresholds L_t and L_e should be constantly adjusted to obtain good candidates. If L_t is much higher, you cannot rule out candidates for poor performance in training; if L_e is too low, you tend to ignore candidates that have very good predictive performance.

3. Results

According to 2.3.3, 144 candidates are built for GPP models. We conducted a total of four experiments. In the first experiment, we split the dataset into a training set and an evaluation set by 90.9% (from 2004 to 2013) and 9.1% (in 2014), respectively. In the second experiment, we used 81.8% (from 2004 to 2012) of the data as the training set and 18.2% (from 2013 to 2014) as the evaluation set. In the third experiment, 72.7% (from 2004 to 2011) of the data is used for training and 27.3% for evaluation (from 2012 to 2014). In the fourth experiment, we split the data into 63.6% for training (from 2004 to 2010) and 36.4% for evaluation (from 2011 to 2014). Table 1 shows the performance of the initial model in the four experimental groups.

Table 1
Training and evaluation performance list of the initial GPP model.

Experiments	Training set/test set (%)	RMSE for training	RMSE for evaluation
1	90.9 / 9.1	42.08	27.96
2	81.8 / 18.2	37.04	30.32
3	72.7 / 27.3	42.2	29.2
4	63.6 / 36.4	42.02	30.66

3.1. Parameter optimization results of GPP model

According to Minimum Loss Screening Method in section 2.4, we adjusted the thresholds L_t and L_e to obtain the six candidates with the smallest evaluation RMSE value from 144 candidates. Table 2 shows the 6 candidate models selected in each experiment. As we can see in this table, the loss value during training is often greater than that of the evaluation. This difference is caused by over-fitting due to the complexity of the data, or to the existence of multiple optimal local solutions in the process of parameter optimization. But the error caused by this phenomenon can be reduced by optimizing the global parameters of each candidate several times. Finally, after multiple optimizations, each parameter value falls within a specific interval, and this paper uses the interval average value as the final parameter value to reduce the impact of the local optimal solution. Through several experiments, it is found that some models appear stably in each result, so it further explains that the four cross-validation experiments performed in this article are necessary.

Based on the results in Table 2, we can see that most of the selected candidates (22 times selected) are centered between the Model-109 and Model-144 ranges. Only two candidates (Model-7 and Model-35) are not in this range. One common feature of the candidates in this range is that the temperature equation is replaced by the Q10 equation. This indicates that the carbon absorption capacity of vegetation increases with the increase of temperature, and the Q10 equation can better fit the relationship between temperature and carbon absorption of vegetation than the hyperbolic equation. Another common feature is that among the 22 selected candidate models, the PAR equation of 15 models is a Sigmoid-like function. It shows that with the increase of light capacity, the carbon absorption capacity of vegetation also increases. However, when the amount of light radiation is too low or too high, the effect of its change

Table 2
Results of four experiments using the Minimum Loss Screening Method.

		Candidates	Training RMSE	Evaluation RMSE
Experiment 1	$Lt = 39$ $Le = 23$	Model-115	35.89	22.96
		Model-133	38.46	21.79
		Model-137	37.19	22.81
		Model-139	37.23	22.52
		Model-140	37.99	21.48
		Model-141	37.46	22.55
Experiment 2	$Lt = 37$ $Le = 28$	Model-119	35.5	26.23
		Model-133	36.9	24.4
		Model-136	36.45	27.21
		Model-137	36.98	24.33
		Model-140	37.54	22.5
		Model-142	35.21	25.18
Experiment 3	$Lt = 41$ $Le = 26$	Model-133	39.55	24.99
		Model-137	38.68	24.97
		Model-140	38.19	24.41
		Model-141	38.56	24.62
		Model-112	40.11	22.97
		Model-138	36.27	25.65
Experiment 4	$Lt = 41$ $Le = 26$	Model-7	38.7	25.47
		Model-35	40.42	25.42
		Model-133	39.24	25.31
		Model-137	40.52	22.83
		Model-139	37.67	25.62
		Model-143	37.79	24.75

on the photosynthesis of vegetation is less than that of the appropriate amount of light radiation. Therefore, the Sigmoid-like function can better fit the relationship between photosynthetic radiation and vegetation carbon absorption.

Fig. 2-1 shows scatter plots of the training results among the initial model and the six better candidates in four experiments. Except for Model-140 in experiment 2, the loss values of the other selected candidates in four training experiments are all lower than the initial training model. Then, Fig. 2-2 shows scatter plots of the evaluation capacity among the initial model and the six better candidates in the four experiments. In the four experiments, the loss values of the selected candidates in the evaluation are all lower than those in the initial model. According to the trend line, except for experiment 1, we also found that these candidates are closer to the observations than the initial model. This illustrates that the Minimum Loss Screening Method can help find candidates with a better evaluation than the initial model.

3.2. Optimized environment variable function

In each experiment, this paper uses different proportions of training and test set data. According to 3.1, it is found that both Model-133 and Model-137 appear in the results of the four cross-validation experiments, then we compared these two candidates and the performance of the initial model in the four experiments. We found that Model-133 and Model-137 have a high stability. As shown in Fig. 3-1, in the four experiments, the loss values of Model-133 and Model-137 are smaller than

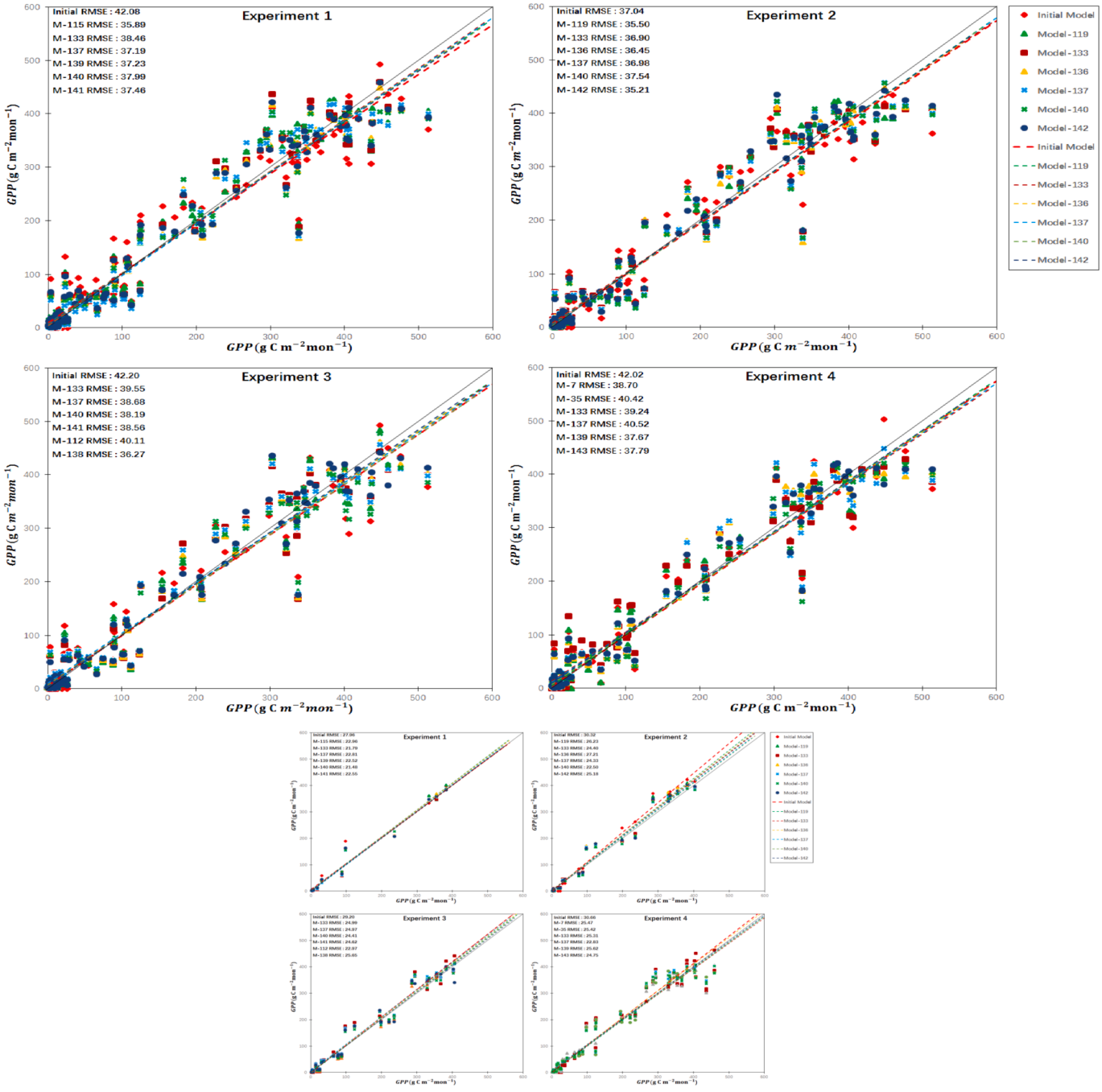

Fig. 2. 2-1. Scatter plot of training results between the six candidate models selected by four experiments and the initial model (units: g C m^{-2} $month^{-1}$). 2-2. Scatter plot of evaluation results between the six candidate models selected by four experiments and the initial model (units: g C m^{-2} $month^{-1}$).

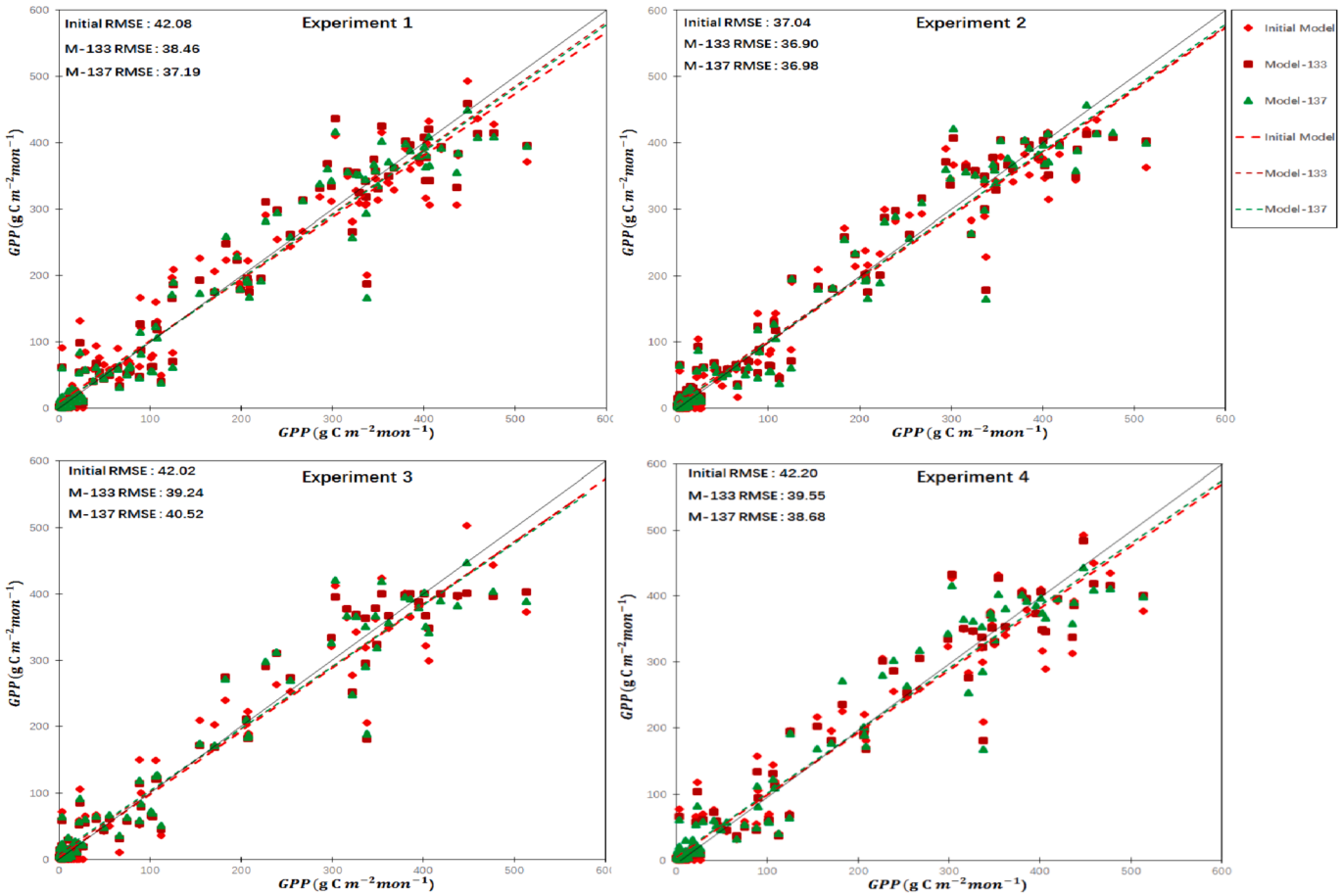

Fig. 3. 3-1. Scatter plot of training results of the initial model and the new models (Model-133 and Model-137) in four experiments (units: g C m^{-2} $month^{-1}$). 3-2. Scatter plot of evaluation results of the initial model and the new models (Model-133 and Model-137) in four experiments (units: g C m^{-2} $month^{-1}$).

the that of the initial model in the training experiments. As shown in Fig. 3-2, the Model-133 and Model-137 loss values are smaller than the initial Model in the evaluation experiments. Therefore, there is a higher accuracy and stability in Model-133 and Model-137 than the initial model during training and test, and they can be used as the best model for GPP quantification. Table 3 (units: g C m^{-2} $month^{-1}$) shows the formulas for the two best candidate models (Model-133 and Model-137). It can be found that the PAR function of these two new models is replaced by Sigmoid-like function, and the temperature function is replaced by Q10. In addition, the CO_2 function of Model-137 is replaced by the Sigmoid-like function, while other functions remain unchanged.

When the optimal model is output, the calibration of global parameters is completed simultaneously. Global parameters include not only the parameters of the GPP quantitative model, but also other parameters of the forest ecosystem model. In the process of vegetation carbon absorption, part of the carbon source is consumed as autotrophic respiration of vegetation, such as maintenance respiration and growth respiration of vegetation. Another part of the carbon source falls in the form of fallen leaves and twigs during the growth of vegetation. The parameters of the best models (Model-133 and Model-137) obtained through global parameter optimization of all candidate models are shown in Table 4 and Table 5. The parameters of the two new models can be used directly in the Harvard Forest.

3.3. Sensitivity analysis

Sensitivity analysis is used to measure the contribution of each object, factors or changes of parameters to the results (Saltelli et al., 2000). For complex models, sensitivity analysis is conducive to mining the influence of some key parameters on the model (Collins and Avissar, 1994). Since a new model is proposed in this paper, it is necessary to conduct sensitivity analysis on environmental variables and global parameters, so as to find out which environmental variables in the new model have a key impact on the carbon absorption of vegetation.

The degree of influence in the environmental variables for Model-133 and Model-137 on the GPP estimate is shown in Table 6. We can see that for the Model-133, the estimated value remains unchanged as the PAR increases or decreases by 10%. When EET and T increase by 10% and 5 °C, respectively, the estimated value of GPP also remains unchanged. However, when EET decreases by 10%, the estimated value decreases by 5.8%. With T decreases by 5 °C, the estimated value decreases from 413.12 g C m^{-2} $month^{-1}$ to 353.30 g C m^{-2} $month^{-1}$, a decrease of 14.48%, which has the greatest impact on the accuracy of estimating GPP. When CO_2 increases or decreases by 10%, the estimated value also increases and decreases by 3.74% and 4.22%, respectively. Therefore, when using Model-133, it is necessary to improve the positive errors of monthly average temperature, evapotranspiration and CO_2 concentration, among which the influence of monthly average temperature is the most obvious.

For Model-137, the estimates remain the same when CO_2 is increased or decreased by 10%. When PAR is increased or decreased by 10%, the estimated change is only 0.01%, with little effect. When EET increases or decreases by 10%, the estimated value also increases and decreases by 7.24%. However, when T decreases by 5 °C, the estimated value decreases from 413.12 g C m^{-2} $month^{-1}$ to 356.91 g C m^{-2} $month^{-1}$, decreasing by 13.77%, which has the greatest impact on the accuracy of

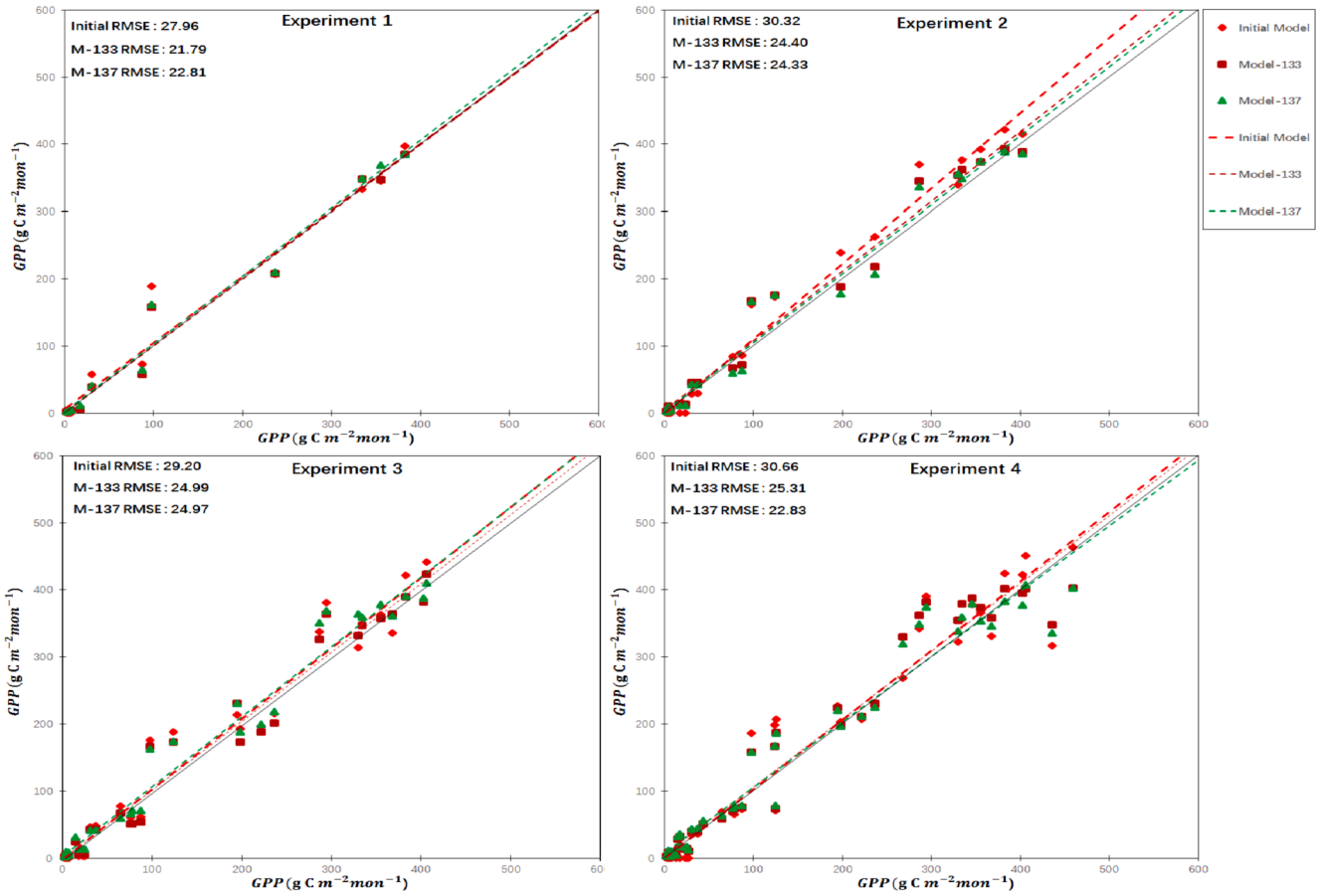

Fig. 3. (*continued*).

Table 3
Model equation for the optimal GPP quantitative model(Model-133 and Model-137).

Candidates	Model equation
Model-133	$GPP = Cmax * f_2(PAR) * f(P) * f(F) * f_1(T) * f(C) * f(NA)$
Model-137	$GPP = Cmax * f_2(PAR) * f(P) * f(F) * f_1(T) * f_1(C) * f(NA)$

estimating GPP. Therefore, when using Model-137, it is necessary to improve the measurement accuracy of monthly average temperature and evapotranspiration to reduce the influence of measurement errors on the estimation results. In summary, the two models are most sensitive to changes in air temperature, followed by evapotranspiration.

The influence of global parameters of Model-133 and Model-137 on GPP estimation is shown in Fig. 4. We can find that when the parameter *Cmax* increases or decreases by 10%, the result of model estimation also increase or decrease, and it has the greatest impact on the accuracy of estimating GPP. For Model-133, the parameters a and b in formula (4), *kc* in formula (10), and m_1 in formula (8) all affect the estimation result, and in descending order of influence is a > *kc* > b > m_1. For Model-137, parameter a in formula (4), parameters *p*1, *p*2, *p*3, *p*4 in formula (16), parameters m_1, m_2, m_3, m_4 in formula (8), and parameters *c*2, *c*3 and *c*4 in formula (20) all affect the estimation results. In summary, the two new models have the highest sensitivity to the parameter *Cmax*. Therefore, in order to reduce the impact on the estimation accuracy of GPP, it is necessary to perform multiple parameter optimizations on *Cmax* to find the global optimal solution.

4. Discussions

Aiming at the problem of GPP modeling, this study proposes a method to develop a new GPP model. The result shows that this method is feasible in the Harvard Forest. And it shall be also applicable for learning algorithms for modeling other ecosystem fluxes including soil carbon decomposition and plant respiration. In addition, our method can be easily used to find other important factors for ecological process modeling, like spatial texture information (Guo et al., 2020; Hoyle, 1995). As long as the formula containing the new factors is added to the candidate, our model can automatically calculate whether the new factors can be introduced to improve the estimation accuracy.

The advantage of this study is that as long as flux data is available, the method can be applied to any region to find the optimal model. Therefore, our method is not only applicable to the Harvard Forest deciduous broad-leaved forest ecosystem, but also applicable to other vegetation types, such as temperate grasslands, polar tundra, temperate mixed forests, tropical evergreen forests, etc. The best GPP quantitative model may be different with different regions' flux data. In conclusion, our developed method can be used for model selection and optimization when observational data are available for any terrestrial ecosystems. The method can also be extended to study other processes for terrestrial ecosystems at regional scales.

4.1. Comparison of the new model with other models in the Harvard Forest

We summarize the GPP quantification experiments conducted in the Harvard Forest area, Table 7 shows the comparison of the new models with other models related to Harvard Forest flux data. According to Fig. 3-2, the maximum RMSE of our new models is reduced to below

Table 4
Global parameters for Model-133.

Parameters	Value	Units	Description	Reference
Cmax	832.3987	g m^{-2} $month^{-1}$	Monthly maximum rate of Photosynthesis C	(Raich et al., 1991)
Q_{10}	4.270871		Ecosystem specific Q10 coefficient indicating the air temperature dependency of photosynthesis	(Zhuang et al., 2004)
Tr	17.42568	°C	Ecosystem-specific reference air temperature used in the Q10 function for simulating the effects of air temperature on photosynthesis	(Zhuang et al., 2004)
a	0.7809454		Regression-derived parameter for phenological processes	(Raich et al., 1991)
b	0.2857267		Regression-derived parameter for phenological processes	(Raich et al., 1991)
c	0.0596473		Regression-derived parameter for phenological processes	(Raich et al., 1991)
Min	0.03691715		Parameter for phenological processes	(Raich et al., 1991)
m1	0.2328779		Parameter for canopy leaf biomass equation	(Zhuang et al., 2002)
m2	−0.2258299		Parameter for canopy leaf biomass equation	(Zhuang et al., 2002)
m3	0.119329		Parameter for logistic function of Cv	(Zhuang et al., 2002)
m4	0.53885		Parameter for logistic function of Cv	(Zhuang et al., 2002)
Cv0	17,516.97	g m^{-2}	Initial C in vegetation (in 2004)	(Zhuang et al., 2002)
Kr	0.0003903169	g g^{-1} $month^{-1}$	Plant respiration rate* at 0 °C	(Raich et al., 1991)
KFALL	0.0001907366	g g^{-1} $month^{-1}$	Proportion of Cv lost as Lc monthly	(Raich et al., 1991)
P1	2.39273		Parameter for canopy photosynthetically active radiation equation	
P2	−3.220593		Parameter for canopy photosynthetically active radiation equation	
P3	7.67201		Parameter for logistic function of f(radiation)	
P4	2.180516		Parameter for logistic function of f(radiation)	
KC	251.7685	μL/L	Half-saturation constant for CO_2-C uptake by plants	(Raich et al., 1991)

Table 5
Global parameters for Model-137.

Parameters	Value	Units	Description	Reference
Cmax	710.6023	g m^{-2} $month^{-1}$	Monthly maximum rate of Photosynthesis C	(Raich et al., 1991)
Q_{10}	4.926471		Ecosystem specific Q10 coefficient indicating the air temperature dependency of photosynthesis	(Zhuang et al., 2004)
Tr	17.27769	°C	Ecosystem-specific reference air temperature used in the Q10 function for simulating the effects of air temperature on photosynthesis	(Zhuang et al., 2004)
a	0.7202978		Regression-derived parameter for phenological processes	(Raich et al., 1991)
b	0.1733052		Regression-derived parameter for phenological processes	(Raich et al., 1991)
c	0.1007003		Regression-derived parameter for phenological processes	(Raich et al., 1991)
Min	0.04930058		Parameter for phenological processes	(Raich et al., 1991)
m1	0.3365532		Parameter for canopy leaf biomass equation	(Zhuang et al., 2002)
m2	−0.5732397		Parameter for canopy leaf biomass equation	(Zhuang et al., 2002)
m3	0.7387806		Parameter for logistic function of Cv	(Zhuang et al., 2002)
m4	0.6666389		Parameter for logistic function of Cv	(Zhuang et al., 2002)
CV0	16,407.54	g m^{-2}	Initial C in vegetation (in 2004)	(Zhuang et al., 2002)
Kr	0.000306816	g g^{-1} $month^{-1}$	Plant respiration rate* at 0 °C	(Raich et al., 1991)
KFALL	0.00053412	g g^{-1} $month^{-1}$	Proportion of Cv lost as Lc monthly	(Raich et al., 1991)
p1	1.415015		Parameter for canopy photosynthetically active radiation equation	
p2	−2.186898		Parameter for canopy photosynthetically active radiation equation	
p3	3.335645		Parameter for logistic function of f(radiation)	
p4	6.177337		Parameter for logistic function of f(radiation)	
c1	8.900292		Parameter for canopy moisture limitation on CO_2 assimilation	
c2	−6.000216		Parameter for canopy moisture limitation on CO_2 assimilation	
c3	16.49978		Parameter for logistic function of Gv	
c4	16.40041		Parameter for logistic function of Gv	

25.50 g C m^{-2} $month^{-1}$ when estimated from 2011 to 2014. Chen et al. (2011) used SAT-TEM and TEM models to simulate the Harvard Forest GPP from 2002 to 2006, and the RMSE is 45.62 g C m^{-2} $month^{-1}$ and 58.63 g C m^{-2} $month^{-1}$, respectively. Schaefer et al. (2012) used 23 EK and LUE models to simulate the Harvard Forest comprehensive GPP value from 1991 to 2006, with RMSE of about 80.35 g C m^{-2} $month^{-1}$. Yebra et al. (2015) estimated GPP using 891 days between 2000 and 2011 in Harvard Forest with satellite-derived light-use efficiency and canopy conductance, they first used Cross-site optimization with a RMSE

Table 6
Sensitivity analysis of Model-133 and Model-137 to the changes in EET, PAR, T and CO_2.

		Baseline	EET	EET	PAR	PAR	T	T	CO_2	CO_2
			10%	−10%	10%	−10%	+5 °C	−5 °C	10%	−10%
Model-133	Consumption (units:g C m^{-2} $month^{-1}$)	413.12	413.12	389.17	413.12	413.12	413.12	353.3	428.56	395.7
	Change (%)	0	0	−5.8	0	0	0	−14.48	3.74	−4.22
Model-137	Consumption (units:g C m^{-2} $month^{-1}$)	413.92	443.9	383.95	413.95	413.89	413.92	356.91	413.92	413.92
	Change (%)	0	7.24	−7.24	0.01	−0.01	0	−13.77	0	0

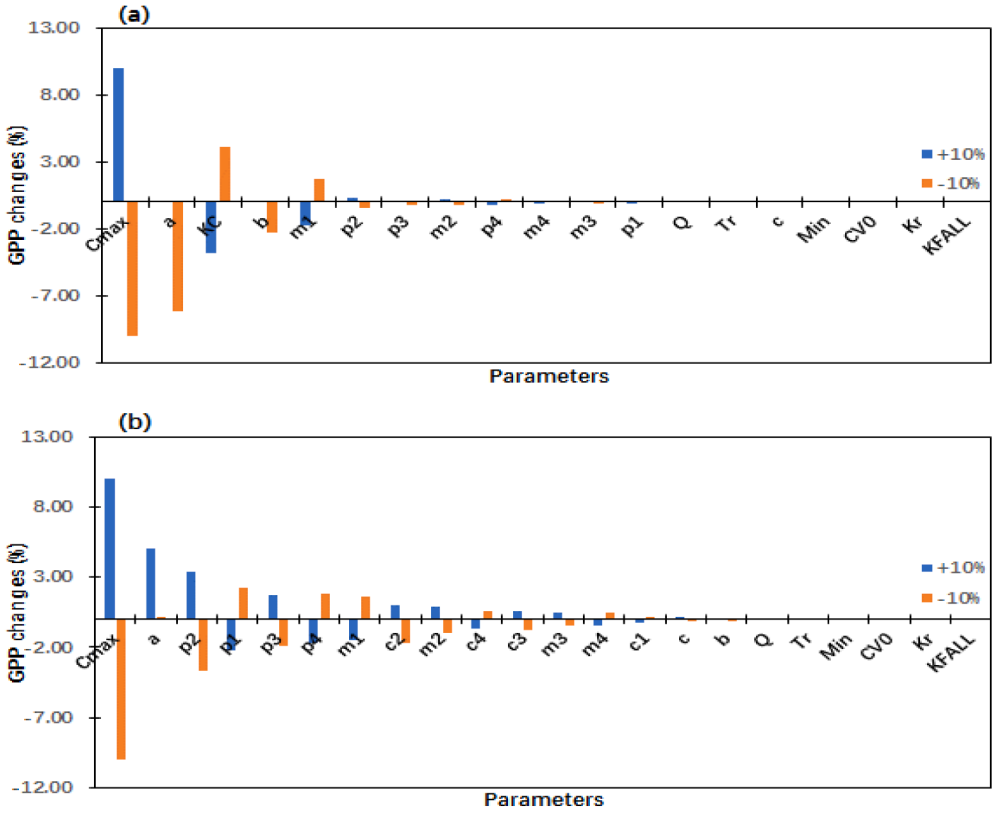

Fig. 4. Histogram are shown for the global sensitivity test of model parameters. (a) When other parameters remain unchanged, the estimated value changes of the Model-133 when each parameter is successively increased by 10% and decreased by 10%. (b) When other parameters remain unchanged, the estimated value changes of the Model-137 when each parameter is successively increased by 10% and decreased by 10%.

of74.57 g C m^{-2} $month^{-1}$, then used per-site optimization with a RMSE of73.60 g C m^{-2} $month^{-1}$. Wu et al. (2010) combined remote sensing and climate data to develop three GPP assessment models to assess GPP in the Harvard Forest from 2003 to 2006, and the RMSE of MOD GPP is71.55 g C m^{-2} $month^{-1}$, then VPM's RMSE decreased to 22.44 g C m^{-2} $month^{-1}$, TGM's RMSE decreased to22.60 g C m^{-2} $month^{-1}$, and VIM's RMSE decreased to22.58 g C m^{-2} $month^{-1}$.

When comparing the accuracy of each model through the RMSE, we can see from Chen et al. (2011) study that our model has been improved. Moreover, our model's R^2 is higher than that of other models or methods.

4.2. Comparison of the new model with the MODIS quantitative model

MODIS GPP is widely used in the Harvard Forest region (Heinsch et al., 2006; Wu et al., 2010; Xiao et al., 2010; Yebra et al., 2015) as a method for quantifying GPP as a Light Utilization Efficiency Model (LUE) (Schaefer et al., 2012). This model used PAR multiplied by the remote sensing PAR portion (fPAR) of vegetation absorption and the biomass conversion factor (commonly referred to as light utilization efficiency) to estimate GPP (Field et al., 1995; Goetz et al., 1999; Heinsch et al., 2003; Landsberg and Waring, 1997; Monteith, 1972; Prince and Goward, 1995; Running et al., 2000, 2004). MODIS GPP's algorithm (MOD17 algorithm) is effectively adjusted by eddy flux observations (Running et al., 2004; Zhao et al., 2005). Eddy flux observations can be used as a benchmark for evaluating both MODIS GPP and our new models. We first selected GPP data covering the EMS towers (latitude: +42.537755, longitude: −72.171478) from the Aqua and Terra satellite products. The 8-day temporal resolution data under a 1 km × 1 km grid cell are aggregated into the monthly GPP from 2004 to 2014. As shown in Fig. 5, compared with EMS eddy flux observation of GPP, from 2004 to 2014, the GPP estimation RMSEs for MODIS-Aqua and MODIS-Terra are 73.0 g C m^{-2} $month^{-1}$ and 74.8 g C m^{-2} $month^{-1}$, respectively, while the GPP estimation RMSEs for the two new models are 34.8 g C m^{-2} $month^{-1}$ and 34.9 g C m^{-2} $month^{-1}$, respectively.

Table 7
Other modeling studies related to the Harvard Forest flux data.

Model or method	Start year	Stop year	RMSE	R^2	Reference
Model-133	2011	2014	<25.50 g C m^{-2} $month^{-1}$	>0.97	
Model-137	2011	2014	<25.00 g C m^{-2} $month^{-1}$	>0.97	
SAT-TEM	2002	2006	45.62 g C m^{-2} $month^{-1}$	0.90	(Chen et al., 2011)
TEM	2002	2006	58.63g C m^{-2} $month^{-1}$	0.87	(Chen et al., 2011)
EK models plus LUE models	1991	2006	80.35 g C m^{-2} $month^{-1}$		(Schaefer et al., 2012)
Fc and Fr comparison method cross-site optimization			74.57 g C m^{-2} $month^{-1}$	0.74	(Yebra et al., 2015)
Fc and Fr comparison method per-site optimization			73.60 g C m^{-2} $month^{-1}$	0.75	(Yebra et al., 2015)
MOD_GPP	2003	2006	71.55 g C m^{-2} $month^{-1}$	0.88	(Wu et al., 2010)
VPM	2003	2006	22.44g C m^{-2} $month^{-1}$	0.94	(Wu et al., 2010)
TGM	2003	2006	22.60 g C m^{-2} $month^{-1}$	0.92	(Wu et al., 2010)
VIM	2003	2006	22.58 g C m^{-2} $month^{-1}$	0.90	(Wu et al., 2010)

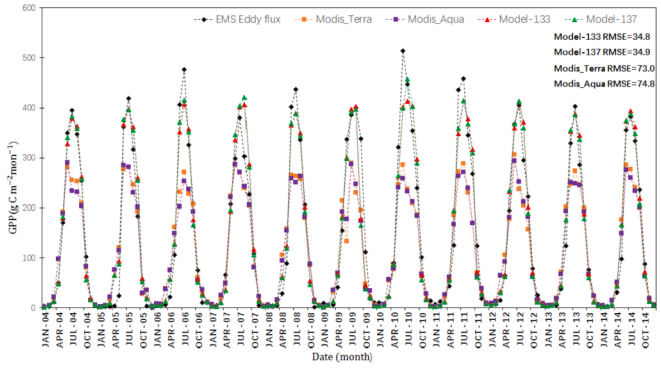

Fig. 5. GPP comparisons between the MODIS satellite products (Terra and Aqua), Model-133 and Model-137 with the observations of EMS Eddy flux from 2004 to 2014 (units:g C m^{-2} $month^{-1}$). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Therefore, the accuracy of GPP estimated by the two new models is more than double that of MODIS GPP. As can be seen from Fig. 5. MODIS underestimated GPP in the middle of the growing season (from June to August), and overestimated GPP in the early growing season (from March to May). Except for the years 2005 and 2013, MODIS underestimated GPP in the senescence season (September). When compared with the observed values, the estimation values of Model-133 and Model-137 in the middle of the growing season (from June to August) are closer to the observed values than the MODIS estimated values. These two new models also show a few deviations in other months, but the overall loss is smaller than that of MODIS GPP.

MODIS GPP is beneficial for large-area estimation (Huang et al., 2015a, 2015b). Thus, if we can use MODIS GPP, which is spatially available for a large region to train our models, we can provide a wider and more accurate GPP estimation at regional scales (Goetz et al., 1999; Xiao et al., 2010).

4.3. Future studies

This paper proposes a new model for quantifying GPP, which has better performance than the initial model and many existing methods. However, there are still many imperfections. Therefore, in the future study, the following four aspects can be carried out.

Find a more suitable basis function. In this study, 10 other basis functions are constructed on the basis of the initial model to establish the basis-function set, and two models (Model-133 and Model-137) with better performance than the initial model are obtained. However, whether they are the best and whether there are more suitable basis functions to construct the model with better performance still need to be further explored. In addition, we only use the parametric model in this paper, and it takes a lot of time for parameter training and selection. In the future, we will consider applying non-parametric probabilistic models to construct basis functions more flexibly (Martino and Read, 2021; Silverman, 1985; Svendsen et al., 2018; Tipping, 2001). Therefore, it is still necessary to focus on the construction of a more suitable basis-function set for environment variables.

Exploration of different types of GPP estimation models. The background of this study is the GPP process model, and no exploration has been made for the EK model and the LUE model. This is mainly due to the different observation principles, leading to differences in the acquisition methods in each type of environmental variable. Therefore, in future studies, other environmental variable data or basis-functions of different types of GPP estimation models can be collected, and then based on the ideas of this paper, they can be further explored to obtain a more accurate estimation model, so as to further verify the extensibility of the method in this paper.

Generalize to a wide range of GPP estimates. The advantage of this study is that, as long as flux data is available, the method presented in this study can be applied to any region to find the optimal model. Therefore, the method is not only applicable to the deciduous broad-leaved forest ecosystem of Harvard Forest, but also applicable to other global vegetation types, such as temperate steppe, polar tundra, temperate mixed forest, and tropical evergreen forest, etc. (McGuire et al., 1992). The aim of trying different vegetation types is that the optimal GPP quantification model may be different depending on the climate zone in which the vegetation type is located. Therefore, in the future, this method can be used to further explore the GPP quantitative models of various vegetation types in different climatic zones.

Continuity measurement of GPP combined with empirical model and MODIS. The empirical model can only estimate the GPP in a small range using measured data with high accuracy; the MODIS GPP algorithm can perform continuous estimation in a large range, but its accuracy is relatively low. Therefore, the two methods can be combined for continuous GPP measurement to obtain an estimation method with wider range and higher accuracy, such as the study of J. Xiao et al. (Xiao et al., 2010). The eddy current observation is the most accurate and can be used as a benchmark to compare the accuracy of model and then effectively adjust the algorithm. However, eddy currents are not observed in all regions of the world. For example, China lacks open eddy flux towers. Therefore, in the future, research areas and data from other countries can be used to train continuous measurement models with strong generalization, and then migrate it for the estimation of China's agricultural yield and forest wood biomass estimation.

5. Conclusions

Based on SCE-UA and Minimum Loss Screening Method, we proposed a new method to optimize the environment variable function for

GPP model. Firstly, a wide range of candidates were built by establishing basis-function sets. Then SCE-UA algorithm and the Minimum Loss Screening Method were used to find the optimal model from a large number of candidates. A cross-validation method was used to test the performance of models selected. We identified two optimum models out of 144 candidates, providing more accurate GPP estimates than the initial model. These two optimal models show that (1) the photosynthetically active radiation function in the initial model can be replaced with a Sigmoid-like function, (2) the temperature function can be replaced with a Q10 equation, and (3) the carbon dioxide equation can use either the semi-saturation equation in the initial model or a Sigmoid-like function. The GPP estimation was more accurate by our new model than other models, validated by the GPP data of Harvard Forest. Our approach was tested efficient, robust, and can be extended to the optimization of other terrestrial ecosystem models.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

This work was supported by the National Key Research and Development Plan of China [grant numbers 2018YFE0122700], which are gratefully acknowledged.

References

Allen, R.G., Pereira, L.S., Raes, D., Smith, M., 1998. Crop evapotranspiration-guidelines for computing crop water requirements. In: FAO Irrigation and Drainage Paper, 56,300, 9: D05109.
Arain, M.A., Yuan, F., Black, T., Meteorology, F., 2006. Soil–plant nitrogen cycling modulated carbon exchanges in a western temperate conifer forest in Canada. Agric. Forest Meteorol. 140 (1–4), 171–192.
Baker, I., Prihodko, L., Denning, A., Goulden, M., Miller, S., Da Rocha, H.R., 2008. Seasonal drought stress in the Amazon: Reconciling models and observations. J. Geophys. Res. Biogeosci. 113, G1.
Ball, J.T., Woodrow, I.E., Berry, J.A., 1987. A Model Predicting Stomatal Conductance and its Contribution to the Control of Photosynthesis under Different Environmental Conditions. Springer, City.
Beer, C., Reichstein, M., Tomelleri, E., Ciais, P., Jung, M., Carvalhais, N., Rödenbeck, C., Arain, M.A., Baldocchi, D., Bonan, G.B., 2010. Terrestrial gross carbon dioxide uptake: global distribution and covariation with climate. Science 329 (5993), 834–838.
Bonan, G.B., Oleson, K.W., Vertenstein, M., Levis, S., Zeng, X., Dai, Y., Dickinson, R.E., Yang, Z.-L.J., 2002. The land surface climatology of the Community Land Model coupled to the NCAR Community Climate Model. J. Clim. 15 (22), 3123–3149.
Chen, M., Zhuang, Q., Cook, D.R., Coulter, R., Pekour, M., Scott, R.L., Munger, J., Bible, K., 2011. Quantification of terrestrial ecosystem carbon dynamics in the conterminous United States combining a process-based biogeochemical model and MODIS and AmeriFlux data. Biogeosciences 8 (9), 2665–2688.
Collatz, G.J., Ball, J.T., Grivet, C., Berry, J.A., meteorology, F., 1991. Physiological and environmental regulation of stomatal conductance, photosynthesis and transpiration: a model that includes a laminar boundary layer. Agric. Forest Meteorol. 54 (2–4), 107–136.
Collatz, G.J., Ribas-Carbo, M., Berry, J., 1992. Coupled photosynthesis-stomatal conductance model for leaves of C4 plants. Aust. J. Plant Physiol. 19 (5), 519–538.
Collins, D.C., Avissar, R.J., 1994. An evaluation with the Fourier amplitude sensitivity test (FAST) of which land-surface parameters are of greatest importance in atmospheric modeling. J. Clim. 7 (5), 681–703.
Field, C.B., Randerson, J.T., Malmström, C.M., 1995. Global net primary production: combining ecology and remote sensing. Remote Sens. Environ. 51 (1), 74–88.
Goetz, S.J., Prince, S.D., Goward, S.N., Thawley, M.M., Small, J., 1999. Satellite remote sensing of primary production: an improved production efficiency modeling approach. Ecol. Model. 122 (3), 239–255.
Guo, B., Yang, F., Han, B., Chen, S., Liu, Y., Yang, X., He, Y., Chen, X., Liu, C., Gong, R., 2020. Spatial and temporal change patterns of net primary productivity and its response to climate change in the Qinghai–Tibet plateau of China from 2000 to 2015. J. Arid Land. 12 (1), 1–17.
Han, J., Moraga, C., 1995. The influence of the sigmoid function parameters on the speed of backpropagation learning. From Nat. Artif. Neural Comput. 930, 195–201.
Hao, G.C., 2015. Study on Medeling the Response of Soil Heterotrophic Respiration to Climate Scenario Using Process-Based Model. Nanjing Agricultural University.
Hayes, D.J., Kicklighter, D.W., McGuire, A.D., Chen, M., Zhuang, Q., Yuan, F., Melillo, J. M., Wullschleger, S.D., 2014. The impacts of recent permafrost thaw on land–atmosphere greenhouse gas exchange. Environ. Res. Lett. 9 (4), 045005.
Heinsch, F., Reeves, M., Bowker, C., Votava, P., Kang, S., Milesi, C., Zhao, M., Glassy, J., Nemani, R., Running, S., 2003. User's guide: GPP and NPP (MOD17A2/A3) products NASA MODIS land algorithm. Version 2. In: MOD17 User's Guide, pp. 1–57.
Heinsch, F.A., Zhao, M., Running, S.W., Kimball, J.S., Nemani, R.R., Davis, K.J., Bolstad, P.V., Cook, B.D., Desai, A.R., Ricciuto, D.M., 2006. Evaluation of remote sensing based terrestrial productivity from MODIS using regional tower eddy flux network observations. IEEE Trans. Geosci. Remote Sens. 44 (7), 1908–1925.
Hoyle, Rick H., 1995. The structural equation modeling approach: basic concepts and fundamental issues. In: Structural Equation Modelling: Concepts, Issues and Applications, pp. 1–15.
Huang, J., Ma, H., Su, W., Zhang, X., Huang, Y., Fan, J., Wu, W., 2015a. Jointly assimilating MODIS LAI and ET products into the SWAP model for winter wheat yield estimation. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 8 (8), 4060–4071.
Huang, J., Tian, L., Liang, S., Ma, H., Becker-Reshef, I., Huang, Y., Su, W., Zhang, X., Zhu, D., Wu, W., 2015b. Improving winter wheat yield estimation by assimilation of the leaf area index from Landsat TM and MODIS data into the WOFOST model. Agric. For. Meteorol. 204, 106–121.
Hurrell, J.W., Holland, M.M., Gent, P.R., Ghan, S., Kay, J.E., Kushner, P.J., Lamarque, J.-F., Large, W.G., Lawrence, D., Lindsay, K.J., 2013. The community earth system model: a framework for collaborative research. Bull. Am. Meteorol. Soc. 94 (9), 1339–1360.
Kan, G., Lei, T., Liang, K., Li, J., Ding, L., He, X., Yu, H., Zhang, D., Zuo, D., Bao, Z., 2016. A multi-core CPU and many-core GPU based fast parallel shuffled complex evolution global optimization approach. IEEE Trans. Parallel Distrib. Syst. 28 (2), 332–344.
Kohavi, R., 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: International Joint Conference on Artificial Intelligence.
Landsberg, J., Waring, R., 1997. A generalised model of forest productivity using simplified concepts of radiation-use efficiency, carbon balance and partitioning. For. Ecol. Manag. 95 (3), 209–228.
Li, H., Qiu, J., Wang, L., Tang, H., Li, C., Van Ranst, E., 2010. Modelling impacts of alternative farming management practices on greenhouse gas emissions from a winter wheat–maize rotation system in China. Agric. Ecosyst. Environ. 135 (1–2), 24–33.
Li, M., He, Y.T., Fu, G., Shi, P.L., Zhang, X.Z., Sun, J., Li, R.Q., Wang, J.B., 2016. Livestock-forage balance in the three river headwater region based on the terrestrial ecosystem model. Ecol. Environ. Sci. 25 (12), 1915–1921.
Liu, H., Cocea, M., 2017. Semi-random partitioning of data into training and test sets in granular computing context. Gran. Comput. 2 (4), 357–386.
Liu, H., Gegov, A., Cocea, M., 2016a. Rule Based Systems for Big Data: A Machine Learning Approach, vol. 13, pp. 1–121.
Liu, H., Gegov, A., Cocea, M., 2016b. Rule-based syste1ms: a granular computing perspective. Gran. Comput. 1 (4), 259–274.
Liu, H., Gegov, A., Cocea, M., 2017. Unified framework for control of machine learning tasks towards effective and efficient processing of big data. In: Data Science Big Data: An Environment of Computational Intelligence., vol. 24, pp. 123–140.
Martino, L., Read, J., 2021. A joint introduction to Gaussian processes and relevance vector machines with connections to Kalman filtering and other kernel smoothers. Inform Fusion 74, 17–38.
McGuire, A.D., Melillo, J.M., Joyce, L., Kicklighter, D.W., Grace, A., Moore, I.I.I.B., Vorosmarty, C.J., 1992. Interactions between carbon and nitrogen dynamics in estimating net primary productivity for potential vegetation in North America. Glob. Biogeochem. Cycles 6 (2), 101–124.
McGuire, A.D., Melillo, J.M., Kicklighter, D.W., Pan, Y., Xiao, X., Helfrich, J., Moore, I.I.I. B., Vorosmarty, C.J., Schloss, A.L., 1997. Equilibrium responses of global net primary production and carbon storage to doubled atmospheric carbon dioxide: sensitivity to changes in vegetation nitrogen concentration. Glob. Biogeochem. Cycles 11 (2), 173–189.
Monteith, J., 1972. Solar radiation and productivity in tropical ecosystems. J. Appl. Ecol. 9 (3), 747–766.
Munger, W., Wofsy, S., 1999. Canopy-Atmosphere Exchange of Carbon, Water and Energy at Harvard Forest EMS Tower since 1991 (LTER).
Pan, Y., Melillo, J.M., McGuire, A.D., Kicklighter, D.W., Pitelka, L.F., Hibbard, K., Pierce, L.L., Running, S.W., Ojima, D.S., Parton, W.J., 1998. Modeled responses of terrestrial ecosystems to elevated atmospheric CO 2: a comparison of simulations by the biogeochemistry models of the vegetation/ecosystem modeling and analysis project (VEMAP). Springer 114 (3), 389–404.
Prince, S.D., Goward, S.N., 1995. Global primary production: a remote sensing approach. J. Biogeogr. 22 (4–5), 815–835.
Raich, J.W., Schlesinger, W.H., 1992. The global carbon dioxide flux in soil respiration and its relationship to vegetation and climate. Tellus Ser. B Chem. Phys. Meteorol. 44 (2), 81–99.
Raich, J.W., Rastetter, E., Melillo, J.M., Kicklighter, D.W., Steudler, P., Peterson, B., Grace, A., Moore, B., Vorosmarty, C.J., 1991. Potential net primary productivity in South America: application of a global model. Ecol. Appl. 1 (4), 399–429.
Riley, W., Still, C., Torn, M., Berry, J., 2002. A mechanistic model of H218O and C18OO fluxes between ecosystems and the atmosphere: model description and sensitivity analyses. Glob. Biogeochem. Cycles 16 (4), 42-41-42-14.
Running, S.W., Thornton, P.E., Nemani, R., Glassy, J.M., 2000. Global terrestrial gross and net primary productivity from the earth observing system. Meth. Ecosyst. Sci. 44–57.
Running, S.W., Nemani, R.R., Heinsch, F.A., Zhao, M., Reeves, M., Hashimoto, H., 2004. A continuous satellite-derived measure of global terrestrial primary production. Bioscience 54 (6), 547–560.

Saltelli, A., Tarantola, S., Campolongo, F., 2000. Sensitivity naysis as an ingredient of modeling. Stat. Sci. 15 (4), 377–395.
Schaefer, K., Collatz, G.J., Tans, P., Denning, A.S., Baker, I., Berry, J., Prihodko, L., Suits, N., Philpott, A., 2008. Combined simple biosphere/Carnegie-Ames-Stanford approach terrestrial carbon cycle model. J. Geophys. Res. Biogeosci. 113, G3.
Schaefer, K., Schwalm, C.R., Williams, C., Arain, M.A., Barr, A., Chen, J.M., Davis, K.J., Dimitrov, D., Hilton, T.W., Hollinger, D.Y., 2012. A model-data comparison of gross primary productivity:results from the north American carbon program site synthesis. J. Geophys. Res. Biogeosci. 117, G3.
Silverman, B.W., 1985. Some aspects of the spline smoothing approach to non-parametric regression curve fitting. J. R. Stat. Soc. Ser. B Methodol. 47, 1–52.
Sun, Y., Frankenberg, C., Wood, J.D., Schimel, D.S., Jung, M., Guanter, L., Drewry, D., Verma, M., Porcar-Castell, A., Griffis, T.J., 2017. OCO-2 advances photosynthesis observation from space via solar-induced chlorophyll fluorescence. Science 358 (6360) eaam5747.
Svendsen, D.H., Martino, L., Campos-Taberner, M., Garcia-Haro, F.J., Camps-Valls, G., 2018. Joint Gaussian processes for biophysical parameter retrieval. IEEE Trans. Geosci. Remote Sens. 56, 1718–1727.
Tian, H., Melillo, J., Kicklighter, D., McGuire, A., Helfrich, J., 1999. The sensitivity of terrestrial carbon storage to historical climate variability and atmospheric CO2 in the United States. Tellus Ser. B Chem. Phys. Meteorol. 51 (2), 414–452.
Tian, H., Chen, G., Liu, M., Zhang, C., Sun, G., Lu, C., Xu, X., Ren, W., Pan, S., Chappelka, A., management, 2010. Model estimates of net primary productivity, evapotranspiration, and water use efficiency in the terrestrial ecosystems of the southern United States during 1895–2007. For. Ecol. Manag. 259 (7), 1311–1327.
Tipping, M.E., 2001. Sparse Bayesian learning and the relevance vector machine. J. Mach. Learn. Res. 1, 211–244.
Walter, B.P., Heimann, M., 2000. A process-based, climate-sensitive model to derive methane emissions from natural wetlands: application to five wetland sites, sensitivity to model parameters, and climate. Glob. Biogeochem. Cycles 14 (3), 745–765.
Wang, Y.-P., Leuning, R., Meteorology, F., 1998. A two-leaf model for canopy conductance, photosynthesis and partitioning of available energy I: model description and comparison with a multi-layered model. Agric. For. Meteorol. 91 (1–2), 89–111.
Wang, S., Zhuang, Q., Lähteenoja, O., Draper, F.C., Cadillo-Quiroz, H., 2018. Potential shift from a carbon sink to a source in Amazonian peatlands under a changing climate. Proc. Natl. Acad. Sci. U. S. A. 115 (49), 12407–12412.
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 (12), 2925–2939.
Xiao, J., Zhuang, Q., Law, B.E., Chen, J., Baldocchi, D.D., Cook, D.R., Oren, R., Richardson, A.D., Wharton, S., Ma, S., 2010. A continuous measure of gross primary production for the conterminous United States derived from MODIS and AmeriFlux data. Remote Sens. Environ. 114 (3), 576–591.
Yang, X., Wittig, V., Jain, A.K., Post, W., 2009. Integration of nitrogen cycle dynamics into the integrated science assessment model for the study of terrestrial ecosystem responses to global change. Glob. Biogeochem. Cycles 23 (4).
Yebra, M., Van Dijk, A.I., Leuning, R., Guerschman, J.P., 2015. Global vegetation gross primary production estimation using satellite derived light-use efficiency and canopy conductance. Remote Sens. Environ. 163, 206–216.
Zhang, S., Lin, G., 2018. Robust data-driven discovery of governing physical laws with error bars. Proc. R. Soc. A 474 (2217), 20180305.
Zhang, S., Lin, G., 2019. Robust data-driven discovery of governing physical laws using a new subsampling-based sparse Bayesian method to tackle four challenges (large noise, outliers, data integration, and extrapolation). J. arXiv Preprint arXiv: 1907.07788.
Zhao, M., Heinsch, F.A., Nemani, R.R., Running, S.W., 2005. Improvements of the MODIS terrestrial gross and net primary production global data set. Remote Sens. Environ. 95 (2), 164–176.
Zhuang, Q., Romanovsky, V., McGuire, A., 2001. Incorporation of a permafrost model into a large-scale ecosystem model: Evaluation of temporal and spatial scaling issues in simulating soil thermal dynamics. J. Geophys. Res.-Atmos. 106 (D24), 33649–33670.
Zhuang, Q., McGuire, A., O'neill, K., Harden, J., Romanovsky, V., Yarie, J., 2002. Modeling soil thermal and carbon dynamics of a fire chronosequence in interior Alaska. J. Geophys. Res.-Atmos. 108, D1.
Zhuang, Q., Melillo, J.M., Kicklighter, D.W., Prinn, R.G., McGuire, A.D., Steudler, P.A., Felzer, B.S., Hu, S., 2004. Methane fluxes between terrestrial ecosystems and the atmosphere at northern high latitudes during the past century: a retrospective analysis with a process-based biogeochemistry model. Glob. Biogeochem. Cycles 18, 3.
Zhuang, Q., He, J., Lu, Y., Ji, L., Xiao, J., Luo, T., 2010. Carbon dynamics of terrestrial ecosystems on the Tibetan plateau during the 20th century: an analysis with a process-based biogeochemical model. Glob. Ecol. Biogeogr. 19 (5), 649–662.
Zhuang, Q., McGuire, A., Melillo, J., Clein, J.S., Dargaville, R., Kicklighter, D., Myneni, R. B., Dong, J., Romanovsky, V., Harden, J., 2011. Carbon cycling in extratropical terrestrial ecosystems of the northern hemisphere during the 20th century: a modeling analysis of the influences of soil thermal dynamics. Tellus Ser. B Chem. Phys. Meteorol. 55 (3), 751–776.
Zhuang, Q., Chen, M., Xu, K., Tang, J., Saikawa, E., Lu, Y., Melillo, J.M., Prinn, R.G., McGuire, A.D., 2013. Response of global soil consumption of atmospheric methane to changes in atmospheric climate and nitrogen deposition. Glob. Biogeochem. Cycles 27 (3), 650–663.