Applied Energy 262 (2020) 114566

Contents lists available at ScienceDirect

Applied Energy

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

Prediction and analysis of net ecosystem carbon exchange based on gradient boosting regression and random forest

Jianchao Cai[a,b], Kai Xu[b], Yanhui Zhu[d], Fang Hu[c,*], Liuhuan Li[c]

[a] State Key Laboratory of Petroleum Resources and Prospecting, China University of Petroleum, Beijing 102249, PR China
[b] Institute of Geophysics & Geomatics, China University of Geosciences, Wuhan 430074, PR China
[c] College of Information Engineering, Hubei University of Chinese Medicine, Wuhan 430065, PR China
[d] Department of Mathematics and Statistics, University of West Florida, Pensacola 32514, USA

HIGHLIGHTS

- Gradient boosting regression and random forest are combined to analyse net ecosystem carbon exchange.
- The model considers 22 environmental variables, more than other similar works.
- The extremums are employed to analyse the corresponding variables' importance.

ARTICLE INFO

Keywords:
Net ecosystem carbon exchange
Variable importance analysis
Gradient boosting regression
Random forest
Prediction model

ABSTRACT

Carbon balance is essential to keep ecosystems sustainable and healthy. Net ecosystem carbon exchange (*NEE*), which is affected by a bunch of meteorological variables to different extent, helps to gauge the balance of the carbon cycle between biological organisms and atmosphere. In this study, the *NEE* data is collected from two flux measuring sites. Gradient boosting regression algorithm is employed to predict *NEE* based on the meteorology and flux data from site UK-Gri. During the training process, KFold cross-validation algorithm is implemented to avoid overfitting, and random forest algorithm is implemented to identify the important variables influencing *NEE* mostly. The four most important variables are found to be global radiation, photosynthetic active radiation, minimum soil temperature, and latent heat. The regression model was compared with three state-of-the-art prediction models: support vector machine, stochastic gradient descent, and bayesian ridge to verify its performance. The experimental results show that this regression model outperforms the other three models, and gives higher value of R-squared, lower values of mean absolute error and root mean squared error. To verify the regression model's generalization ability, the data from the second flux site, NL-Loo, was employed, and the hybrid data of the two sites was used. The results show that this model performs well on the hybrid data, too. In practical terms, the gradient boosting regression model provides many tunable hypterparameters and loss functions, which make it more flexible and accurate compared to the other three models. This study has conclusively demonstrated for the first time that the combination of gradient boosting regression and random forest models should be considered as valuable tools to make effective prediction for *NEE* and acquire reliable important variables influencing *NEE* mostly. The methodologies could be useful in the research fields of ecosystem stability evaluation, environmental restoration, trend analysis of climate change, and global warming monitoring.

1. Introduction

Net ecosystem carbon exchange (*NEE*, also called net ecosystem productivity) is a measure of the difference between net primary production (*NPP*, the net production of organic carbon fixed by plants in an ecosystem) and the carbon losses in heterotrophic respiration [1]. To maintain life on the earth, carbon in the atmosphere and carbon fixed in biological organisms need to be balanced. Animals add more carbon to the ecosystem by simply breathing. Atmospheric carbon is also produced by decaying, as dead animals and plant matter release the carbon stored in their tissues, and by the combustion of trees, plants and fossil fuels, such as oil and coal. Living plants remove carbon dioxide from the

* Corresponding author.
E-mail address: naomifang@hbtcm.edu.cn (F. Hu).

https://doi.org/10.1016/j.apenergy.2020.114566
Received 1 November 2019; Received in revised form 16 January 2020; Accepted 25 January 2020

atmosphere and transform it into oxygen and food energy, which is the reason they are referred to as "carbon sinks" [2]. *NEE* helps to gauge the balance of carbon cycle. As it is calculated by subtracting how much carbon plants fix or remove, from how much carbon is put into the atmosphere, the best result would be a negative value. *NEE* is a key measure to examine the impact of meteorological variability on ecosystem carbon balance [3].

FLUXNET is a global network of micrometeorological tower sites that use eddy covariance method to measure flux data (e.g. *NEE*, the exchange of carbon dioxide, water vapour, and energy between terrestrial ecosystems and the atmosphere) [4], as well as meteorology data (e.g. sum of global radiation (*Rg*), average air temperature (*Ta*), and sum of precipitation (*PPT*)). The original flux and meteorology data is recorded at half-hour scale. Reichstein et al. [5] defined a preprocessing algorithm to optimize the procedure of upscaling the half-hour scale data to seasonal scale. More than 500 tower sites around the world have been operating on a long-term and continuous basis until now. It was launched by NASA as a means of validating the Earth Observation Satellite [4]. Teklemariam et al. [3] evaluated the link between inter-annual variability in ecosystem CO_2 exchange and meteorological conditions using a decade of eddy covariance (flux) and meteorology data recorded at the Mer Bleue peatland. Chu et al. [6] analysed the dynamic characteristic of ecosystem CO_2 exchange and its influence factors at different time scales using eddy covariance data. Sun et al. [7] analysed the relation between diurnal variations of *NEE* and environmental variables by measuring CO_2 flux continuously in 2011–2012 using eddy covariance method in the Yangtze River Delta of China. They found that the main environmental variable impacting *NEE* was photosynthetically active radiation (*PAR*) during daytime, and *Ta* at nighttime.

All the aforementioned methods use process-based models, which are based on theories and hypotheses about biosphere function and are limited by the prescribed and fixed model structure with simplified representation of selected processes and ecosystem components [8]. In order to acquire more accurate results about the objective relationship behind the flux and meteorology data, machine learning is employed in *NEE* study. Machine learning is the scientific study of algorithms and statistical models. It performs a specific task effectively without using explicit instructions, which is called data-oriented approach. This technique has been widely discussed over the past few years and been applied in various research domains [9–11]. Li and Zhang [12], Zhang et al. [13], and Li et al. [14] used machine learning methods to study and predict CO_2 solubility in kinds of physical, chemical and biological conditions. Wijk and Bouten [15] employed a three layer back-propagation neural network to model water and carbon flux using 5 variables measured by eddy flux sites in 6 different Northwestern Europe coniferous forests. They found both short term water and carbon fluxes can be modeled without using detailed physiological and site specific information, which is one of the pioneers in the interdiscipline of *NEE* study and machine learning. Artificial neural network (ANN) has been one of the most prevailing algorithms used in *NEE* prediction. Melesse and Hanley [16], He et al. [17], Qin et al. [18], and Safa et al. [19] used multilayer ANN to study and predict *NEE* in different ecosystems (forest, grassland, and cropland) based on 5~12 meteorological variables. Moffat et al. [20] made a comparison between process-based and data-driven models. Besides, they used data from daytime carbon fluxes of the deciduous broadleaf forest Hainich in Germany and selected 14 variables as inputs of ANN. They found that total photosynthetic photon flux density is the dominant variable and vapor pressure deficit (VPD) is the most important non-radiative control. Ichii et al. [21] empirically estimated terrestrial carbon fluxes, global primary productivity (*GPP*), and *NEE* in Asia using support vector machine (SVM) algorithm. Liu et al. [22] and Zhou et al. [23] used random forest (RF) model to study the drivers of carbon fluxes. They found the most important variables are *PAR*, *Ta*, soil temperature (*Ts*), leaf area index (*LAI*), and soil moisture. Dou and Yang [24] and Dou et al. [25] introduced two new machine learning algorithms to predict *NEE*: adaptive neuro-fuzzy inference system (ANFIS) and extreme learning machine (ELM). They collected data from three boreal forest ecosystems with 5-year span, then selected 4 inputs and combined them to generate 7 different inputs to simulate the other 3 variables (*GPP*, *NEE*, and ecosystem respiration). By changing the internal functions for different algorithms (ANN, SVM, ANFIS, and ELM) and comparing among them, they concluded that there is no such a single universal model with the same training function that could guarantee the most accurate estimations at all sites. Díaz et al. [26] used both gradient boosting regression (GBR) tree and principal component regression models to make day-ahead predictions of the electricity price in Spain. GBR tree produced remarkably low prediction errors when using the median as point prediction method. Hassan et al. [27] implemented 6 different machine learning algorithms, including gradient boosting (GB), bagging, RF, multi-layer perceptron, SVM, and decision tree (DT), to model solar radiation in daily and hourly timescales. They found GB was the most stable one with acceptable computational costs. Besides, genetic neural network [28], ELM [29], model tree ensembles method [8], and fuzzy rough set algorithm [30] are also employed to predict and analyse *NEE*. A brief introduction for approaches identifying main drivers of carbon flux was also included in the work of Xue et al. [30]. They pointed out that the relative roles of controlling variables of carbon flux are critical to devise ecosystem management strategies for mitigating global warming effects and to improve process models. Although researches on prediction of environmental analysis using various machine learning models achieved a lot, it is still an open question whether an accurate and efficient prediction model is applied for *NEE* prediction.

According to the literature aforementioned, ANN and SVM are two common ways to study *NEE* simulation and production, although their shortcomings are well known, e.g., slow learning speed, over-fitting, local minima, difficulties to determine some hyper-parameters for ANN and high memory requirement, large amount of computing time for SVM [31]. ANFIS, ELM, DT, GBR, and RF are employed to overcome the disadvantages of ANN and SVM, among which the combination of GBR and RF is rarely studied. GB [32] is an ensemble learner for both regression and classification, which means it creates a final model based on a collection of individual models. The predictive power of these individual models is weak and prone to overfitting, but combining many such weak models in an ensemble leads to an overall much improved result. The work of Díaz et al. [26] and Hassan et al. [27] showed the advantages of GBR: low prediction errors and nice stability. The construction of GBR generates concomitantly the RFs [33] with some specific meaningful nodes, which can be evaluated by the mathematical approaches to find the hidden relationships. Almost all modelling approaches focus on the relations of few variables (from 4 to 14), which may ignore some true important variables at the very beginning of the work, because there is no sufficient knowledge for how much the quantity of inputs could affect *NEE* [19]. Besides, all the variables considered by the above models use average values and just erase the information of extremums, which deny the impacts of the extreme meteorology components on *NEE* before constructing the models.

Taking the merits and demerits of the above models into account, the objectives of this manuscript are: 1) to implement GBR model for the prediction of *NEE* and RF model for the identification of the important variables that influence *NEE* mostly; 2) to include as many environmental variables (22 features) as possible in the modeling process; 3) to employ some extremums (maximums and minimums) of the variables for more accurate importance analysis.

2. Materials and measures

2.1. Summary of data

The original data set is collected from two flux towers located in Griffin Forest, Aberfeldy, Scotland and Loobos, Netherlands (UK-Gri, NL-Loo, shown in Fig. 1), the former of which is a sub-dataset of the FLUXNET Marconi Conference Gap-Filled Flux and Meteorology Data, 1992–2000 [34], and the latter is from FLUXNET2015 [35]. The details about the two sites are listed in Table 1 and Table 2, (e.g., locations, elevations, and climate types). It shows that the latitude difference of the two sites is about 4.4°, the longitude difference is about 9.5°, and UK-Gri is nearly 320 m higher than NL-Loo, with the latter closing to sea level. Although the location difference of the two sites, both of them belong to mid-latitude region, with the same Koeppen–Geiger climate classification, IGBP land use, NPP land cover, LAI fpar, and plant function type (shown in Table 1). The main process of model construction is based on data of UK-Gri. Data of NL-loo is only employed as further verification of the accuracy of GBR model in Section 4.3.

The meteorology and flux dynamics of UK-Gri were continuously observed in the whole year 1997 with one-day interval. After data cleaning, 730 samples were acquired with 22 features (environmental variables) considered. The 22 features are rearranged according to their importance to *NEE*. The 3-year data of NL-loo is processed in the same way, from which 1, 143 hybrid data samples of the two sites were acquired for model verification.

For gap-filled flux data (.flx) of UK-Gri, three gap filling methods were applied, as shown in Table 2. For gap-filled meteorology data of UK-Gri, one gap filling method was applied (no more details about the method were mentioned for the data file). For the flux and meteorology data of NL-Loo, the detailed gap-filling process is introduced in reference [5]. It should be noted that these gap-filling procedures have been done by the data provider, not by the authors. Therefore, there are six sub-datasets used in this study, as shown in Table 3.

Here, for site UK-Gri, “AB97” is the site and year identification (means Aberfeldy 1997); “dc, lu, re” represent different gap filling methods; “u0” shows the raw data was corrected when published. “dd” means time resolution is days. For site NL-Loo, “NL-Loo” is the site’s Fluxnet ID. “dd” and “hh” mean that the time solutions are day and half-hour. “101213” means the used data is from years 2010, 2012, and 2013.

For the 6 files, the missing values were marked by − 9, 999. In the process of data cleaning, the first step is to remove the samples when the measured values include − 9, 999 from the original data set. The features of samples after removing procedure are shown in Table 4.

2.2. Statistical features of data

Data set from UK-Gri is treated as the main analysis data for *NEE* prediction based on GBR model. Table 5 shows the summary of the data statistics including mean, standard deviation (stdev), minimum, 25^{th} percentile, 50^{th} percentile, 75^{th} percentile, maximum, and variance threshold. Mean refers to the average of each feature that is used to derive the central tendency of the data. Standard deviation is a measure of the degree of variation or dispersion of each feature. Variance threshold is a simple baseline approach to feature selection and all features whose variances do not meet the threshold are removed in this procedure. Further, the values of minimum, 25^{th} percentile, 50^{th} percentile, 75^{th} percentile, and maximum are plotted in Fig. 2.

Fig. 2 shows the 23 features’ statistics of the UK-Gri data set as box plots, which display the data discrete distribution without influence of outliers. A box plot shows the five-number summary including minimum, 25^{th} percentile, 50^{th} percentile, 75^{th} percentile, and maximum. For each box, the minimum is represented as leftmost short vertical line; the 25^{th}, 50^{th}, and 75^{th} percentiles are denoted as left vertical line, middle vertical line, and right vertical line of the box, respectively; the maximum is expressed as rightmost short vertical line. The outliers are plotted separately as points in Fig. 2.

2.3. Pearson correlation coefficient analysis of data set

In statistics, the Pearson correlation coefficient (or Pearson product-moment correlation coefficient) is a measure of the linear correlation among two variables [36]. It has the value between − 1 and + 1, where + 1 is totally positive linear correlation, 0 is non-linear correlation, and − 1 is totally negative linear correlation. The definition of Pearson correlation coefficient r is as follows:

$$r = \frac{\sum_{i=1}^{n} (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{y}_i - \bar{\mathbf{y}})}{\sqrt{\sum_{i=1}^{n} (\mathbf{x}_i - \bar{\mathbf{x}})^2} \sqrt{\sum_{i=1}^{n} (\mathbf{y}_i - \bar{\mathbf{y}})^2}} \tag{1}$$

where n is the sample size; $\mathbf{x}_i$ and $\mathbf{y}_i$ are the individual sample points indexed with i; $\bar{\mathbf{x}}$ is the sample mean representing as $\frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i$, and analogously for $\bar{\mathbf{y}}$.

Fig. 3 shows the Pearson correlation coefficients between two variables in UK-Gri, where correlation values are represented with different shades.

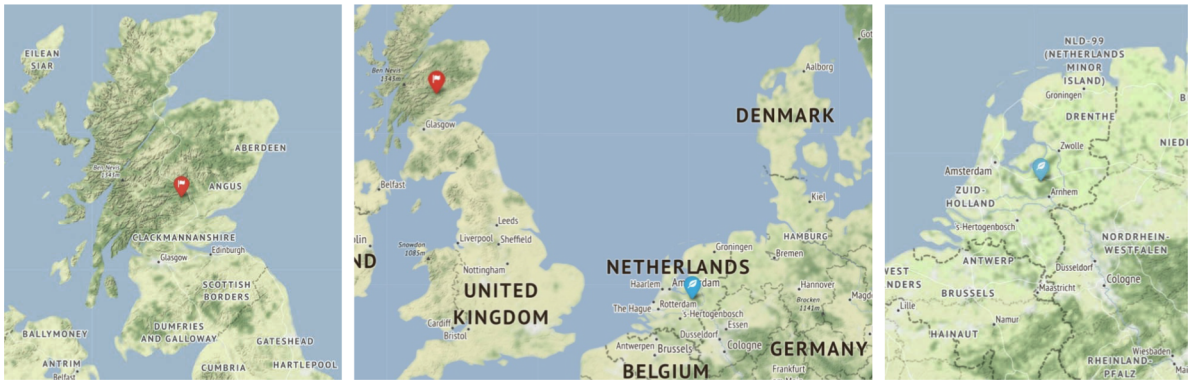

Fig. 1. The location of the two sites. (The map is plotted with Leaflet. Map tiles by Stamen Design, under CC-BY-3.0. Data by ©OpenStreetMap, under CC-BY-SA.) (left: the red marker with a flag is the location of UK-Gri; center: relative locations of the two sites; right: the blue marker with a leaf is NL-loo). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 1
Summary of the flux measuring sites.

Fluxnet ID	UK-Gri	NL-Loo
Site name	Griffin-Aberfeldy-Scotland	Loobos
Location	Aberfeldy, Scotland	Loobos, Netherlands
Time span	1997, the whole year	2010, 2012, 2013, three years
Latitude	56.607222	52.16658
Longitude	− 3.798056	5.743556
Site elevation	343 m	25 m
Koeppen–Geiger climate classification	Cfb - Warm temperate fully humid with warm summer	
IGBP land use	Evergreen needleleaf forest	
UMD land cover	Evergreen needleleaf forest	
LAI fpar	Evergreen needleleaf forest	
NPP land cover	Evergreen needleleaf vegetation	
Plant functional type	Evergreen needleleaf trees	

Table 2
Summary of the data set from measuring sites in UK-Gri and NL-Loo.

Site	UK-Gri	NL-Loo
Time span	1997	2010, 2012, 2013
Data format	Gap-filled flux data (.flx) Gap-filled meteorology data (.met)	Flux and meteorology data in one file (.csv)
Gap filling methods	Nonlinear regression (re) Look up tables (lu) Mean daily courses (dc)	(Method introduced in [5])
Time resolution	Daily (dd)	Daily and half-hourly

years. However, the method is not suitable for measuring fluxes in rough mountainous terrain or near distinct landscape transitions such as lakes [4].

Vertical flux densities of CO_2 (F_c), latent heat (LE) and sensible heat (H) between vegetation and the atmosphere are proportional to the mean covariance between vertical velocity (w') and the respective scalar (c') fluctuations (e.g., CO_2, water vapor, and temperature) [4]. Therefore, the eddy covariance flux of CO_2 could be calculated by [37]:

$$F_c \approx \bar{\rho}_c \, \bar{w'} \, \bar{s'} \tag{4}$$

where F_c is the amount of flux, mol/(m^2 s); $\bar{\rho}_c$ is the mean density of CO_2, mol/m^3; $\bar{w'}$ is the mean vertical velocity of CO_2, m/s; $\bar{s'} = \rho_c/\rho_a$ is the mean mixing ratio of CO_2 in air, dimensionless. For LE and H, Eq. (4) can be changed to:

Table 3
Six sub-data sets employed in the two sites.

FluxnetID	File name	Data format	Gap filling methods
UK-Gri	AB97_dc_u0_dd.flx	Gap-filled flux data	Mean daily courses (dc)
UK-Gri	AB97_lu_u0_dd.flx	Gap-filled flux data	Look up tables (lu)
UK-Gri	AB97_re_u0_dd.flx	Gap-filled flux data	Nonlinear regression (re)
UK-Gri	AB97_dd.met	Gap-filled meteorology data	(Not mentioned)
NL-Loo	NL-Loo_dd_101213.csv	Flux, and meteorology data	Method introduced in [5]
NL-Loo	NL-Loo_hh_101213.csv	Flux, and meteorology data	

2.4. Concept of NEE

NEE refers to the amount of carbon transferred between biological organisms and atmosphere. To define *NEE* accurately, we need to define other variables firstly [2]:

$$NPP = GPP - R_a \tag{2}$$

$$NEE = NPP - R_h \tag{3}$$

where *GPP* refers to the total amount of carbon fixed in the process of photosynthesis by plants in an ecosystem; R_a refers to the amount of carbon respired by plants themselves in autotrophic respiration; *NPP* refers to the net production of organic carbon by plants in an ecosystem (usually measured over a period of a year or more); R_h refers to the carbon losses in heterotrophic respiration. All the exchange rates described above can be expressed in units of carbon amount per unit of area per unit of time, e.g., g C m^{-2}d^{-1} for this study.

2.5. Eddy covariance method

The eddy covariance method is able to measure mass and energy fluxes over short and long timescales, from hours, days, seasons to

$$LE = \bar{\rho}_a \, L_v \, \bar{w'} \, \bar{e'} \tag{5}$$

$$H = \bar{\rho}_a \, C_p \, \bar{w'} \, \bar{T'} \tag{6}$$

where $\bar{\rho}_a$ is the mean density of air, kg/m^3; L_v is the latent heat of vaporization, J/kg; $\bar{e'}$ is the ratio of water vapor in the air; the unit of LE[1] is W/m^2. C_p is the specific heat of air at constant pressure, J/(kg °C); $\bar{T'}$ is fluctuation about the mean of air temperature, °C; the unit of H is W/m^2. Positive flux densities represent mass and energy transfer into the atmosphere and away from the surface; negative values denote the reverse (Ecologists use an opposite sign convention where the uptake of carbon by the biosphere is positive).

3. Methods and evaluations

3.1. Summary of data processing

The summary of the methodologies for *NEE* data processing is

[1] The units of *LE* and *H* are derived using SI. In actual measurement, units could change. For example, MJ/(m^2 day) for *LE*.

Table 4
Summary of features employed in the two sites.

Features	Abbreviation	Units
Sum of net ecosystem exchange	*NEE*	gC m^{-2} d^{-1}
Sum of latent heat	*LE*	M J m^{-2} d^{-1}
Sum of global radiation	*Rg*	M J m^{-2} d^{-1}
Sum of photosynthetic active radiation	*PAR*	Mol m^{-2} d^{-1}
Average air temperature (tower top)	*Ta*	°C
Minimum air temperature of time period	Ta_{mi}	°C
Maximum air temperature of time period	Ta_{mx}	°C
Average soil temperature	*Ts*	°C
Minimum soil temperature of time period	Ts_{mi}	°C
Maximum soil temperature of time period	Ts_{mx}	°C
Average relative humidity (tower top) of time period	*RH*	%
Minimum relative humidity of time period	RH_{mi}	%
Maximum relative humidity of time period	RH_{mx}	%
Average vapor pressure deficit (tower top) of time period	*VPD*	kPa
Minimum vapor pressure deficit of time period	VPD_{mi}	kPa
Maximum vapor pressure deficit of time period	VPD_{mx}	kPa
Average CO_2 concentration in air (tower top) of time period	*Ca*	ppm
Minimum CO_2 concentration in air of time period	Ca_{mi}	ppm
Maximum CO_2 concentration in air of time period	Ca_{mx}	ppm
Sum of precipitation	*PPT*	mm d^{-1}
Average wind speed	*WS*	m s^{-1}
Average air pressure	*Pa*	kPa
Average friction velocity	U^*	m s^{-1}

(Summary of the 23 features employed in the model. *NEE* is the output and the other 22 features are inputs. For the data from UK-Gri, the final data size after data cleaning process is 730. For the hybrid data of UK-Gri and NL-Loo sites, the final data size is 1, 143).

KFold cross-validation method [38]. In the second section, the regression model was constructed based on GBR; the RFs were generated concomitantly after the construction of GBR; the important variables were identified using Gini coefficient; the regression model was trained using **S** and the evaluation metrics were calculated, including R-Squared (R^2), mean absolute error (*MAE*), and root mean squared error (*RMSE*); the best trained model *f* was acquired by iterative optimization process. In the third section, *NEE* was predicted using *f* based on the test set **X** and the results were evaluated with the three metrics.

3.2. Environmental variable importance identification based on RF

Important environmental variables have enormous influence on *NEE* evaluation and analysis. The data type of the variables was transformed from continuous values into discrete values using one-hot encoding method [39]. Then, the RF models of environmental variables were generated after the construction of GBR model. The contribution degree of each feature in each RF tree was calculated using Gini coefficient. The 22 features are ranked according to their Gini coefficients. The most important environmental variables are identified through comparative analysis of the contribution degree. The Gini coefficient (Gini index) [40] is a single number aiming at measuring the degree of inequality in a distribution. The Gini index and variable importance measures are denoted as *GI* and *VIM*, respectively. There are $m = 22$ features (environmental variables) in the data set. The definition of Gini coefficient is as follows:

$$GI_m = \sum_{n=1}^{|N|} \sum_{n' \neq n} p_{mn} p_{mn'} = 1 - \sum_{n=1}^{|N|} p_{mn}^2 \tag{7}$$

where *N* represents the categories; p_{mn} denotes the proportion of category *n* occupying the node *m*, i.e., the proportion of label inconsistency about two samples randomly selected from node *m*. For each feature F_j,

Table 5
Statistical features of data set in UK-Gri.

No.	Features	Mean	Stdev	Minimum	25^{th} percentile	50^{th} percentile	75^{th} percentile	Maximum	Variance Threshold
1	*LE*	1.113	1.241	− 0.1266	0.0647	0.4495	2.153	4.497	1.538
2	*Rg*	6.481	7.579	0	0	2.81	12.14	28.9	57.39
3	*PAR*	11.94	13.85	0	0	5.37	22.61	51.2	191.5
4	*Ta*	9.649	4.163	0.08	6.27	9.78	12.84	21.55	17.31
5	Ta_{mi}	7.410	3.877	− 1.32	4.76	7.25	10.2	16.58	15.01
6	Ta_{mx}	11.71	4.794	1.26	7.77	11.64	15.205	25.25	22.97
7	*Ts*	7.592	2.936	1.6	5.33	7.78	10.15	13.71	8.609
8	Ts_{mi}	6.876	2.955	0.92	4.82	6.77	9.55	13.21	8.717
9	Ts_{mx}	8.267	3.099	1.61	5.67	8.58	10.9	15.08	9.597
10	*RH*	76.87	10.31	48.5	69.70	77.06	84.98	98.16	106.2
11	RH_{mi}	57.54	16.58	30.06	42.77	59.98	69.6	96.47	274.8
12	RH_{mx}	90.96	7.935	59.36	84.33	93.01	98.19	100	62.95
13	*VPD*	0.2957	0.1609	0.022	0.188	0.259	0.38	0.844	0.0259
14	VPD_{mi}	0.095	0.0787	0	0.025	0.077	0.1515	0.405	0.0062
15	VPD_{mx}	0.5804	0.3274	0.042	0.325	0.515	0.746	1.804	0.1072
16	*Ca*	366.0	13.65	333.4	359.9	365.4	371.6	526.6	186.3
17	Ca_{mi}	356.6	12.81	257.9	350.3	357.6	364.6	391	164.1
18	Ca_{mx}	377.5	22.01	337.3	371.8	375.5	379.9	673.7	484.6
19	*PPT*	1.924	4.380	0	0	0	1.4	38.6	19.18
20	*WS*	2.499	1.2849	0	1.545	2.29	3.205	7.47	1.65
21	*Pa*	97.04	0.9839	93.8	96.62	97	97.61	99.8	0.9672
22	U^*	0.3682	0.2170	0.047	0.2175	0.334	0.4515	1.433	0.047
23	*NEE*	− 1.212	2.887	− 9.095	− 3.455	− 0.5145	1.171	3.377	8.325

(Statistics of 730 samples with 23 features in UK-Gri, which reflect the discrete distribution of data set).

outlined in Fig. 4. The data processing procedure is composed of three sections: data preparation, data training, and data predicting. In the first section, the original data set was cleaned by deleting all the samples with missing values (missing values are marked as − 9, 999), and then was split into training set **S** and test set **X**; **S** was analysed with its statistics, normalized, and split into training set and validation set using

the Gini index score represents the average change in node splitting impurity of the j^{th} feature for all decision trees in RF. The importance of feature F_j in node *m*, i.e., the index value change before and after the node *m* branching, is defined as follows:

$$VIM_{jm}^{(Gini)} = GI_m - GI_l - GI_r \tag{8}$$

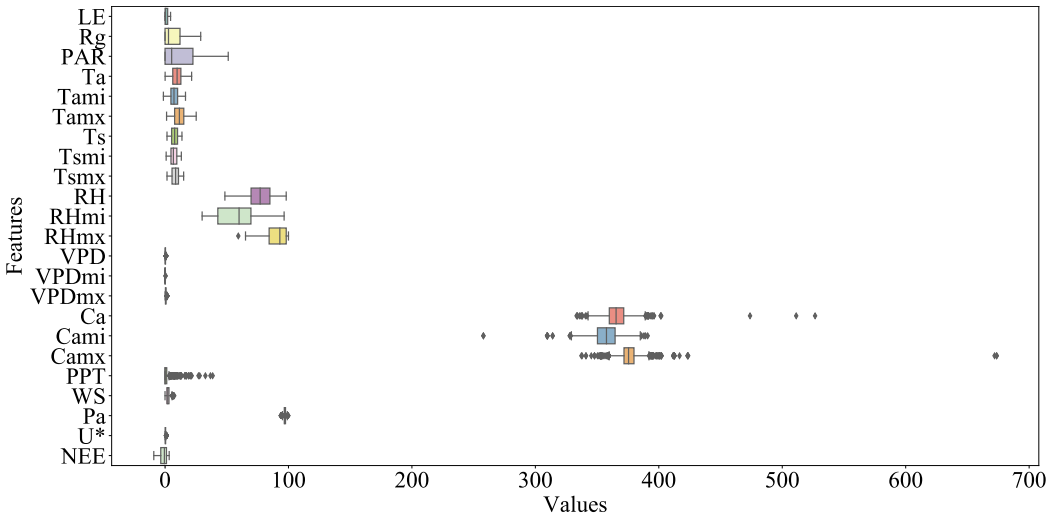

Fig. 2. Statistical features of the UK-Gri data set.

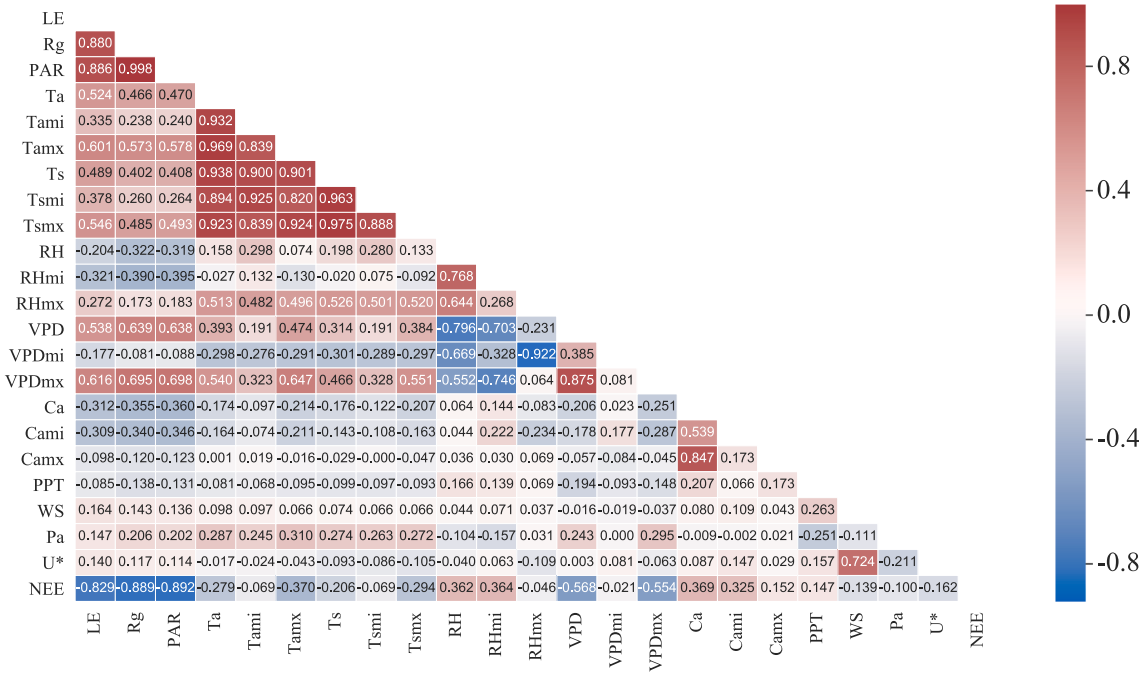

Fig. 3. Pearson correlation coefficients between two variables of *NEE* data set in UK-Gri. (Red color indicates the two variables are positive correlated. Blue color indicates the two variables are negative correlated. White color indicates the two variables are non-linear correlated). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

where GI_l and GI_r represent the Gini index values of two new nodes after node m branching, respectively.

If the node with feature F_j appears in the decision tree i and this node belongs to the assemble M, then the importance of F_j in the i^{th} tree is defined as:

$$VIM_{ij}^{(Gini)} = \sum_{m \in M} VIM_{jm}^{(Gini)} \tag{9}$$

If RF has t decision trees, then,

$$VIM_{j}^{(Gini)} = \sum_{i=1}^{t} VIM_{ij}^{(Gini)} \tag{10}$$

Finally, the acquired importance scores VIM_j are normalized as follows (denoted by VIM'_j):

$$VIM'_j = \frac{VIM_j}{\sum_{i=1}^{m} VIM_i} \tag{11}$$

3.3. NEE prediction based on GBR

3.3.1. Idea of NEE prediction

The data type analysis result shows that the *NEE* values are continuous values. The regression models have the characteristics to predict continuous values for various domains. The GBR model [32,41] integrates weak prediction models, such as DT, RF, etc., and generates a strong prediction model. In each phase, this algorithm constructs a series of weak models by optimizing a differentiable loss function. This model trains the current base learner focusing on the errors of learning previous base learners. For GBR model, in which the RFs are used as

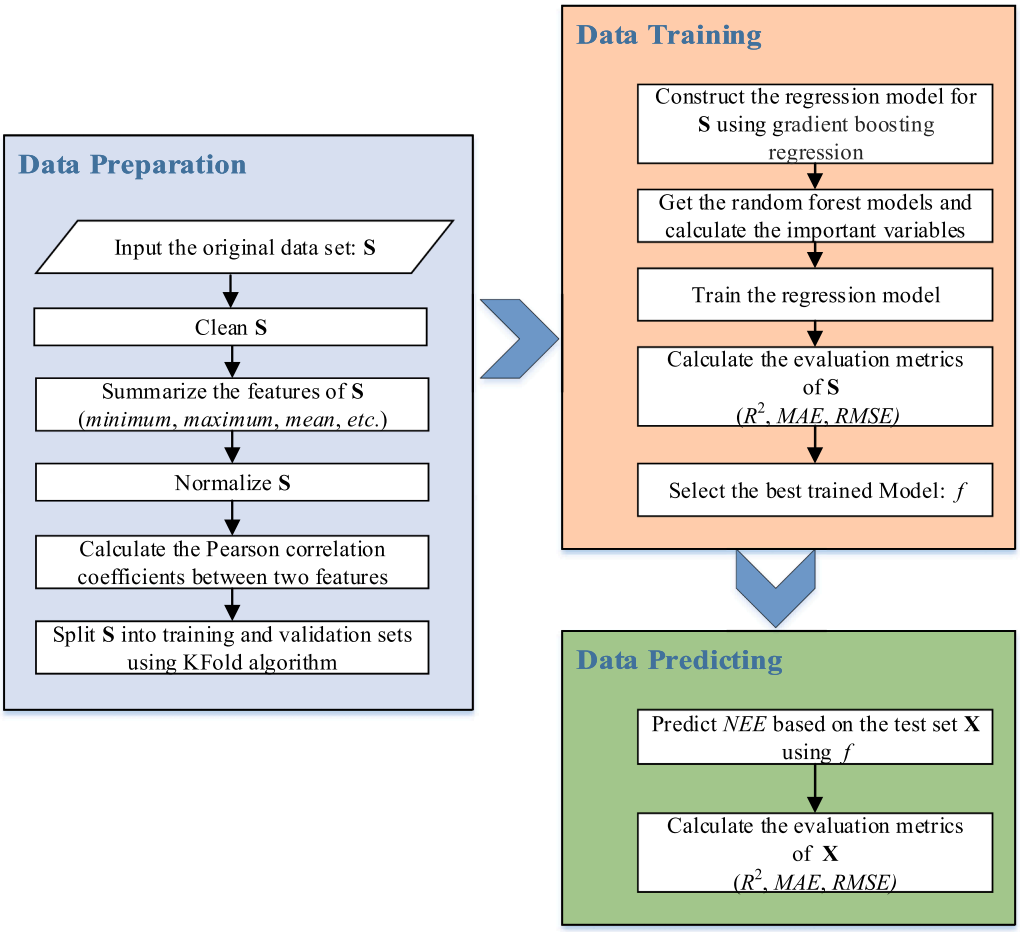

Fig. 4. Flowchart of data processing.

base learners, the negative gradient is taken as an evaluation index to measure the errors of the previous base learners. In the next learning, the previous errors will be updated by fitting the negative gradient. Compared with other regression techniques, GBR generally provides better accuracy and requires minimal data preprocessing, which is useful in implementing the model faster with less complexity. Further, GBR provides many tunable hyperparameters and loss functions, which make the model more flexible, meaning that it could be used to solve a wide variety of problems.

The 730 *NEE* samples from UK-Gri is split into training set **S** with 550 samples and test set **X** with 180 samples. Then the GBR model was trained using the training set **S**, and *NEE* was predicted based on test set **X** using this trained regression model. The RFs of environmental variables were generated concomitantly after the construction of GBR model, and the quantified importance value of each variable in RFs was calculated by Gini coefficient. In order to avoid overfitting of prediction model, the KFold cross-validation was implemented to split the training set further. Cross validation is a powerful preventative measure against overfitting. It combines (averages) measures of fitness in prediction, and derives a more accurate estimation of model prediction performance. In the processing steps of KFold algorithm, the original training set is redivided into k subgroups; for each iteration, $k-1$ subgroups are taken as the training data and the remainder one is regarded as validation data; after k iterations, the average accuracy of k models is acquired as the final accurate rate of the training model. Therefore, the KFold algorithm can efficiently avoid the overfitting problem of regression model. The target of the training process is to minimize the loss function by iterative optimization approaches. The model is trained repeatedly until the minimum of the loss function is acquired. Then the strong prediction model f is found and used to predict *NEE* based on the test set **X**.

3.3.2. Steps of NEE prediction

Step 1: Split the training set **S** into 50 disjoint subsets. The number of training samples is 550. Therefore, the number of samples N in the i^{th} subset S_i is 11. Denote these subsets as $\{S_1, S_2, ..., S_{50}\}$. Set the parameter $k = 50$ in the KFold algorithm.

Step 2: Construct the regression model based on GBR and initialize the parameters. The RFs of environmental variables are generated concomitantly after the construction of GBR model. Then the importance values of environmental variables are calculated in each RF using Gini coefficient.

Step 3: Randomly select one subset S_j from training set **S** as a validation set. The feature matrix of training set is denoted as **x** and the expected results is presented as **y**. Take the remainder $k-1$ subsets from **S** as the new training set $\{S_1, S_2, ..., S_{j-1}, S_{j+1}, ..., S_{50}\}$.

Step 4: In the m^{th} iteration of GBR, an additive model, the previous $m-1$ base learners are fixed, i.e.,

$$f_m(\mathbf{x}) = f_{m-1}(\mathbf{x}) + \rho_m h_m(\mathbf{x}) \tag{12}$$

where $h_m(\mathbf{x})$ is the m^{th} base learner; ρ is the parameter to be optimized; $f_m(\mathbf{x})$ is the generated regression model at the m^{th} iteration.

Step: 5 The target at the m^{th} step is to minimize the loss function (sometimes called empirical error function) denoted as follows,

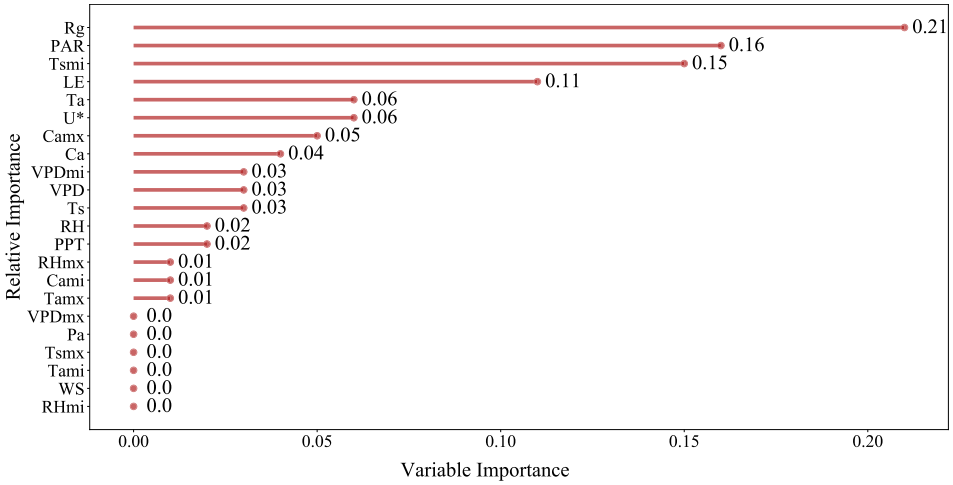

Fig. 5. Plot of environmental variable importance in UK-Gri. (Quantified importance for the 22 environmental variables. The first four variables are the most important, with importance more than 0.1, including *Rg*, *PAR*, Ts_{mi}, and *LE*).

$$L(f) = \sum_{i=1}^{N} L(\mathbf{y}_i,\ f_m(\mathbf{x}_i)) \tag{13}$$

where $\mathbf{x_i}$ is the i^{th} sample, and $\mathbf{y}_i$ is the expected result of i^{th} sample. Eq. (12) can be transformed into Eq. (14) using gradient descent approach.

$$f_m(\mathbf{x}) = f_{m-1}(\mathbf{x}) - \rho_m \frac{\partial}{\partial f_{m-1}(\mathbf{x})} L(\mathbf{y},\ f_{m-1}(\mathbf{x}))h_m(\mathbf{x}) \tag{14}$$

Step: 6 Train $m+1$ base learners from f_0 to f_m, and update the parameter ρ using the gradient descent approach.

Step: 7 For each $S_j, j \in [1,\ k]$, we can get k empirical errors $L(f)$ in Step 5. For a base learner f_i, the empirical error is the mean value er_i of these k empirical errors.

Step: 8 Select a batch of models with the minimum of empirical errors er_i, which will be used to train **S** again.

Step: 9 After finishing the Step 2 to Step 7, output the final trained model $f_m(\mathbf{x})$.

Step: 10 Predict *NEE* using the test set **X** based on the trained model $f_m(\mathbf{x})$, and output the predicted results.

3.3.3. Pseudocode of NEE prediction

The pseudocode of data processing using the GBR model is as follows:

Algorithm 1. *NEE* prediction using the GBR model

Input:
training set: **S** // *NEE* data set with 550 samples
test set: **X** // *NEE* data set with 180 samples
Parameter in KFold algorithm: $k = 50$
Maximum number of iteration: $M = 50$
Output:
Trained model: $f_M(\mathbf{x})$
Predicted results of test set **X**: $\hat{\mathbf{Y}}$

1: Construct the GBR model and acquire the RFs of environmental variables
2: Calculate the importance value of each variable in RF using Gini coefficient
3: $\{S_1,\ S_2,\ ...,\ S_k\} = KFold(\mathbf{S},\ k)$
4: Initial base learner: $f_0(\mathbf{x}) = \arg\min_{\rho} \sum_{i=1}^{N} L(\mathbf{y}_i,\ \rho)$
5: **for** m = 1: M **do**
6: Calculate the negative gradient: $\tilde{\mathbf{y}}_i = -\frac{\partial L(\mathbf{y}_i,\ f_{m-1}(\mathbf{x}_i))}{\partial f_{m-1}(\mathbf{x}_i)},\ i = 1,\ 2,\ ...,\ N$ // $\mathbf{y}_i$ is the actual value and $\tilde{\mathbf{y}}_i$ is the predicted value
7: Minimize the mean absolute error, fit $\tilde{\mathbf{y}}_i$ using one of base learners: $h_m(\mathbf{x}_i),\ w_m = \arg\min_{w} \sum_{i=1}^{N} [\tilde{\mathbf{y}}_i - h_m(\mathbf{x}_i; w)]^2$ // w_m represents the parameter w to evaluate at the m iteration; compared to other accuracy metrics, the mean absolute error has the continuously differentiable characteristic
8: Confirm the step length ρ_m to minimize the loss function $L,\ \rho_m = \arg\min_{\rho} \sum_{i=1}^{N} L(\mathbf{y}_i,\ f_{m-1}(\mathbf{x}_i) + \rho h_m(\mathbf{x}_i; w_m))$
9: Update model: $f_m(\mathbf{x}) = f_{m-1}(\mathbf{x}) + \rho_m h_m(\mathbf{x}; w_m)$
10: **end for**
11: Predict **X** using the trained model $f_M(\mathbf{x})$
12: **return** $\hat{\mathbf{Y}}$ // Predicted results

3.4. Regression evaluation metrics

Generally, researchers use *MSE*, *RMSE*, *MAE*, and R^2 [42] to verify the performance of regression algorithms. The definitions of four metrics are as follows:

① *MSE*

MSE is a metric to evaluate the error between predicted value and actual value, which is defined as follows:

$$MSE = \frac{1}{m} \sum_{i=1}^{m} (\mathbf{y}_i - \hat{\mathbf{y}}_i)^2 \tag{15}$$

where m is the number of samples, $\mathbf{y}_i$ is the actual value and $\hat{\mathbf{y}}_i$ is the predicted value of the i^{th} sample.

② *RMSE*

RMSE is another index to evaluate, which is defined as follows:

$$RMSE = \sqrt{\frac{1}{m} \sum_{i=1}^{m} (\mathbf{y}_i - \hat{\mathbf{y}}_i)^2} \tag{16}$$

③ *MAE*

MAE is defined as follows:

$$MAE = \frac{1}{m} \sum_{i=1}^{m} |(\mathbf{y}_i - \hat{\mathbf{y}}_i)| \tag{17}$$

In general, the lower the *MSE*, *RMSE*, and *MAE* values are, the better the fitting results are.

④ R^2

R^2 is the most important index to verify the accuracy of the predicted result of a regression algorithm, of which the range is [0, 1]. The definition of R^2 is as follows:

Table 6
Environmental variable importance analysis.

No.	Features	Importance
1	Rg	**0. 21**
2	PAR	**0. 16**
3	Ts_{mi}	**0. 15**
4	LE	**0. 11**
5	Ta	0.06
6	U^*	0.06
7	Ca_{mx}	0.05
8	Ca	0.04
9	VPD_{mi}	0.03
10	VPD	0.03
11	Ts	0.03
12	RH	0.02
13	PPT	0.02
14	RH_{mx}	0.01
15	Ca_{mi}	0.01
16	Ta_{mx}	0.01
17	VPD_{mx}	0
18	Pa	0
19	Ts_{mx}	0
20	Ta_{mi}	0
21	WS	0
22	RH_{mi}	0

(Quantified importance for the 22 environmental variables. The most important four are marked with bold font, including Rg, PAR, Ts_{mi}, and LE).

$$R^2 = 1 - \frac{\sum_{i=1}^{m} (\mathbf{y}_i - \hat{\mathbf{y}}_i)^2}{\sum_{i=1}^{m} (\mathbf{y}_i - \bar{\mathbf{y}}_i)^2} \tag{18}$$

where $\bar{\mathbf{y}}_i$ denotes the mean value of actual value $\mathbf{y}_i$. The result of R^2 value equalling to 1 represents the regression model gives predictions without any error. In general, the higher the R^2 value is, the better the fitting result is.

4. Simulation experiments and results analysis

4.1. Environmental variable importance analysis

Fig. 5 and Table 6 show the acquired importance order of environmental variables. The results show that the 22 features influence the NEE to different extents, among which Rg (0.21) is the most important environmental variable, followed by PAR (0.16), Ts_{mi} (0.15), and LE (0.11). This conclusion is similar to Liu et al. [22], Sun et al. [7], and Zhou et al. [23], the former two of which found PAR is the most important variable, while the latter found Rg. Besides, this research considers not only the average values of the variables, but also the extremums, e.g., Ta_{mi}, Ts_{mx}. While Ta is more important than Ts, which is the same as Liu et al. [22], it is notable that Ts_{mi} is more important than Ta. Ts_{mi} influences NEE more importantly than the corresponding average value, Ts. Ta_{mx} influences NEE less importantly than the average value. Ca and its maximum influence NEE in the similar degree. Because LAI data is not available for this study, this variable could not be evaluated and compared. As mentioned in Table 1, the regional climate of the site UK-Gri belongs to Cfb type, which is characterized by rare subzero temperatures, abundant and steady precipitation in all seasons, mild summer, and changeable, often overcast weather (for more details, refer to [43,44]). Because of sufficient precipitation and relatively high humidity, NEE would not be limited by PPT and RH in this circumstance. Solar energy is the power source of the whole earth. The vegetation photosynthesis requires suitable temperature, and low temperature is the main limitation factor for carbon fixation. These empirical analyses correspond to the model results, which give PPT, RH only the importance index of 0.02, while give Rg, PAR, Ts_{mi} the highest importance ranks. Based on the importance rank for environmental variables, researchers could establish the similarity and difference among different climate zones, evaluate the ecosystem stability in local, regional, and even global scales, make strategies for environmental restoration. It should be pointed out that the application of the analysis method proposed requires systematic research and comparisons among all climate zones.

4.2. NEE prediction analysis based on GBR model

To verify the performance of the regression model, three state-of-the-art prediction models: SVM [45,46], stochastic gradient descent (SGD) [47,48], and bayesian ridge (BR) [49] are employed to predict NEE as comparison. The comparison of these four algorithms is shown in Table 7, in terms of model features, hyperparameters, and shapes of regression functions.

Fig. 6 shows the scatter plots of predicted and actual values of NEE based on GBR, SVM, SGD, and BR models using the data from UK-Gri. If the orange point just appears on the black diagonal dash line, it indicates that the predicted value is identical to the actual value; if the orange point appears closely to this dash line, it expresses that the error

Table 7
Comparison of different prediction algorithms.

Algorithms	Model features	Hyperparameters	Regression function shapes
GBR	Discriminative	**n_estimators**: Number of decision trees in the ensemble. **learning_rate**: Shrinks the contribution of each successive decision tree in the ensemble. **loss**: Loss function to be optimized via gradient boosting. **max_depth**: Maximum depth of the decision trees. **max_features**: Number of features to consider when computing the best node split.	axis-aligned partition of feature space
SVM	Discriminative	**kernel**: “linear”, “poly”, “sigmoid”, or “rbf”. **c**: Penalty parameter for regularization. **gamma**: Kernel coefficient for different kernels. **degree**: Degree for the “poly” kernel.	depends on kernel
SGD	Discriminative	**loss**: Loss function to be optimized. **penalty**: Whether to use Lasso, Ridge, or ElasticNet regularization. **alpha**: Regularization strength. **learning_rate**: Shrinks the contribution of each successive training update.	linear
BR	Generative	**_iter**: Maximum number of iterations. **tol**: Stop the algorithm if convergence condition is satisfied.	linear

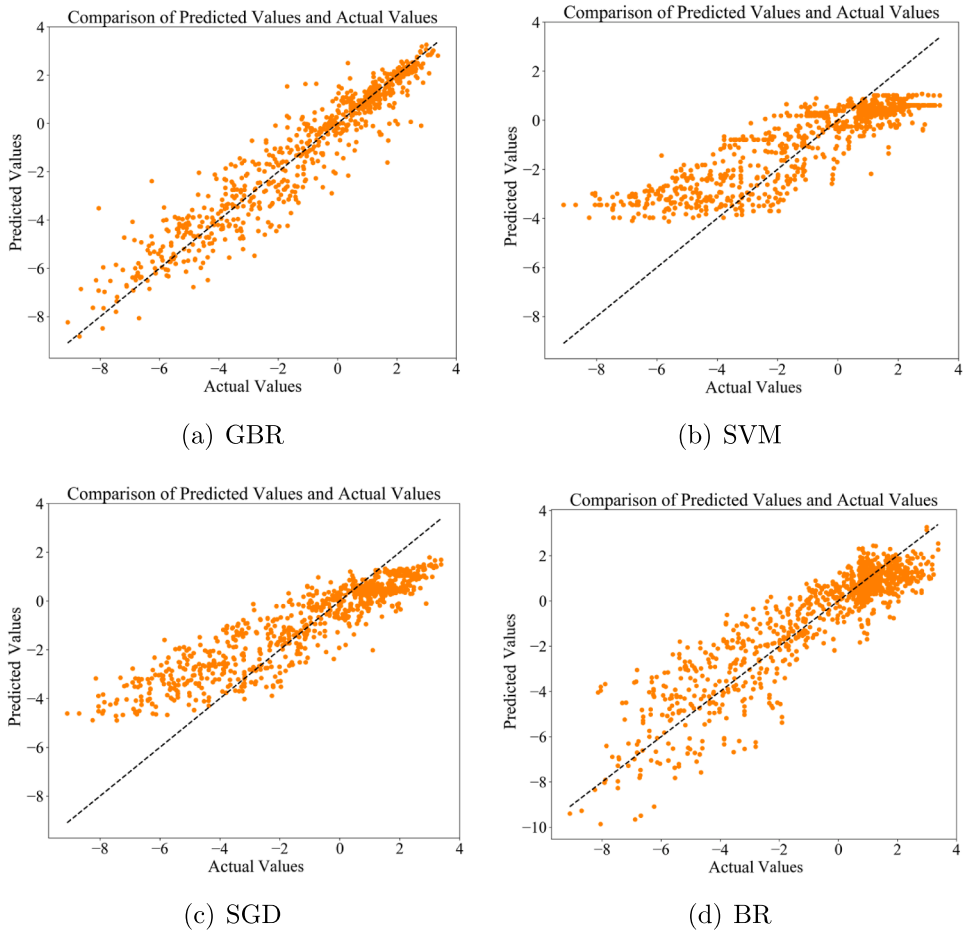

Fig. 6. Comparison of predicted and actual values using GBR, SVM, SGD, and BR models on the data of UK-Gri. (The x-axis represents the actual *NEE* values and y-axis denotes the predicted values calculated by the prediction models. Orange points are the results, and the dash line represents the equality of predicted and actual values. The best predicted results are given by GBR). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

between predicted one and actual one is small. GBR, SVM, and SGD give larger prediction deviations for the negative values, among which GBR gives the least deviations. SVM and SGD give obvious deviate predictions for positive extremums, too, which are reduced to acceptable ranges in GBR. BR gives good results for positive values but large deviations for negatives. The comparison results represent that the GBR model gives better predicted results than other three models. In Fig. 6(a), some points with the actual values less than − 6 are far away from the line. These deviate points give larger predicted values than the actual ones. To find out the reason for the deviations, we made further analysis of the statistical values, including minimum, mean, and maximum, of the 730 samples extracted from UK-Gri. According to the analysis results, there are 50 samples with the normalized actual *NEE* values less than − 6, including 31 samples in training set and 19 in test set. The 22 features of the 50 samples have relatively large deviations from the means of the features for the whole 730 samples. Compared with the whole 730 samples, there are 3 distribution patterns for the 22 features in the 50 samples:

- Right-skewed distribution with fewer negative extremums, leading to larger means, which includes *LE*, *Rg*, *PAR*, *Ta*, Ta_{mi}, Ta_{mx}, *Ts*, Ts_{mi}, Ts_{mx}, RH_{mx}, VPD_{mx}, *WS*, and U^*;
- Left-skewed distribution with fewer positive extremums, leading to smaller means, which includes *RH*, RH_{mi}, VPD_{mi}, *Ca*, Ca_{mi}, Ca_{mx}, and *PPT*;
- Right-skewed distribution with both fewer positive and negative extremums, leading to larger means, which includes *VPD* and *Pa*.

Table 8
Comparative analysis on different models in UK-Gri.

No.	Models	R^2	*MAE*	*RMSE*
1	GBR	**0. 8702**	**0. 8178**	**1. 0722**
2	SVM	0.7433	1.195	2.121
3	SGD	0.7657	1.123	1.936
4	BR	0.7848	1.062	1.778

(Comparison GBR model with three state-of-the-art models. The highest R^2, lowest *MAE* and *RMSE* are marked with bold font. The best results are given by GBR).

These deviations exist in some machine learning approaches to predict *NEE*, e.g., models proposed in papers [16,24,28]. The researches just represented the objective experimental results without explaining the reasons for the deviations. This problem would be analysed further in future works. Table 8 and Fig. 7 show the comparative results of GBR and other three models, which verify that GBR outperforms the other three, with higher value of R^2 (0.8702), lower values of *MAE* (0.8178) and *RMSE* (1.0722).

The 730 samples are split into different scales of training sets and test sets to verify the optimal splitting strategy. Table 9 and Fig. 8 show the results of three splitting strategies. The GBR model acquired the highest value of R^2 (0.8702) when the data set is split into 550 training set and 180 test set. Although the splitting of 400 training set and 330 test set obtained the lowest values of *MAE* and *RMSE*, this splitting got the worst result of R^2 (0.6576), which is the most important index for

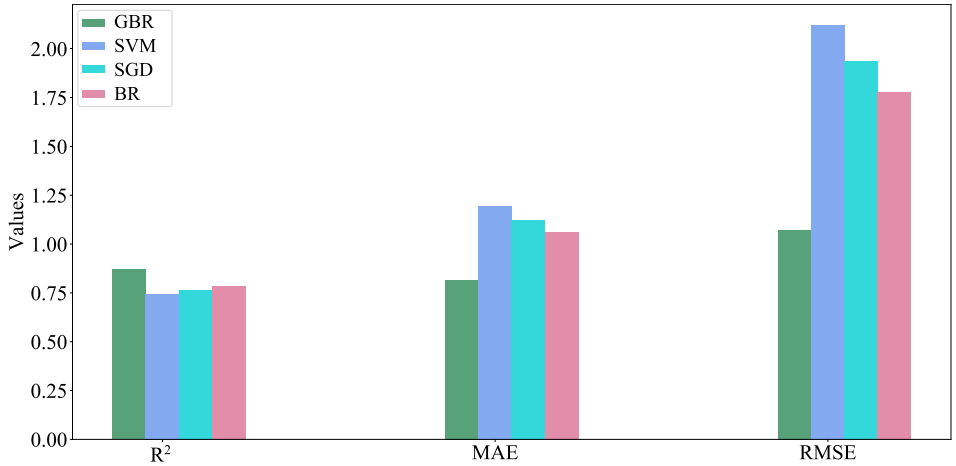

Fig. 7. Comparison plots of GBR, SVM, SGD, and BR models in R^2, *MAE*, and *RMSE* on the data of UK-Gri. (GBR gives the highest R^2, lowest *MAE* and *RMSE*, which shows GBR model outperforms the others).

Table 9
Comparative analysis on GBR model with different scales of training and test sets in UK-Gri.

Training Set	Test Set	R^2	*MAE*	*RMSE*
400	330	0.6576	**0. 4874**	**0. 4770**
500	230	0.7646	0.6292	0.8513
550	180	**0. 8702**	0.8178	1.0722

(The highest R^2, lowest *MAE* and *RMSE* are marked with bold font).

evaluating the result of prediction. Overall, the splitting of 550 training set and 180 test set is considered as the best division for the GBR model to predict *NEE*, after the trade-off among three evaluation metrics. With the prediction model, researchers could make predictions for future *NEE* dynamics by days, months, years, decades, or longer. After that, the trend of climate change would be revealed.

4.3. Verification of GBR model with hybrid data

Further, to verify the GBR model's generalization ability, the data from another flux site (NL-Loo) is employed. Site NL-Loo is located in Netherlands and is similar with UK-Gri in climate type, Koeppen–Geiger climate classification, IGBP land use, UMD land cover, LAI fpar, NPP land cover, and plant functional type (as shown in Table 1). It is known that both Scotland and Netherlands belong to the temperate oceanic climate (warm temperate fully humid with warm summer, more precisely). Generally speaking, the regional climate is consistent in decades, therefore, it may be reasonable to make a contrast between UK-Gri and NL-Loo, and the hybrid data of the two sites is used. Therefore, we combined the year 1997 data (730 samples) in UK-Gri with the year 2010 data (72 samples) and year 2012 data (105 samples) in NL-Loo as the training set, and took the year 2013 data (236 samples) in NL-Loo as the test data.

Fig. 9 shows the scatter plots of predicted and actual *NEE* values based on GBR, SVM, SGD, and BR models. Fig. 9(a) shows that the orange data points distribute uniformly at both sides of the dash line, which means the predicted values are in good agreement with the actual ones. GBR model is also compared with SVM, SGD, and BR on the hybrid data set (Fig. 9(b), (c), (d)). The four model all make well predictions for positive values. When predicting negative values, the deviations become obvious. Among the four, GBR gives relatively smaller deviations. The results (Table 10 and Fig. 10) also show that the GBR model outperforms other three models, with higher value of R^2 (0.9020), lower values of *MAE* (0.5405) and *RMSE* (0.6860), which are good results for regression models. The results imply that this regression model is suitable for predicting *NEE* based on different sites.

5. Conclusions

The GBR model was constructed for *NEE* prediction based on one-

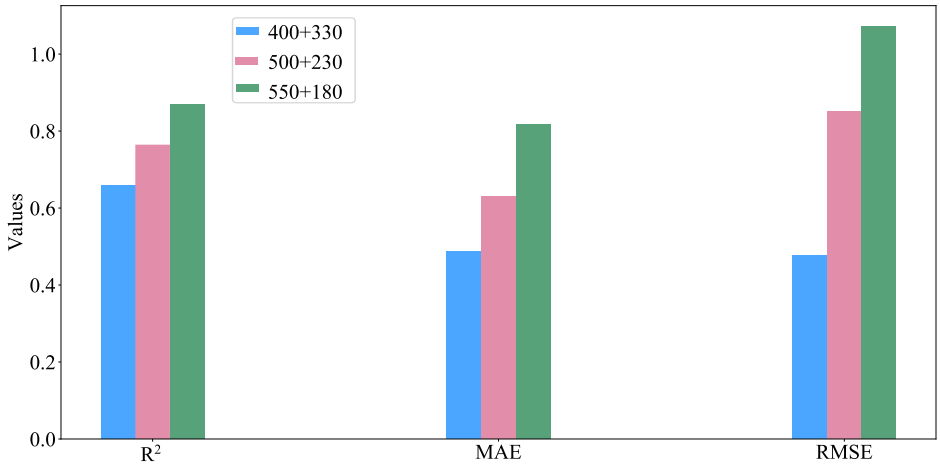

Fig. 8. Comparison of GBR results on different sizes of training and test sets. (The splitting of 550 training set and 180 test set gets the highest R^2).

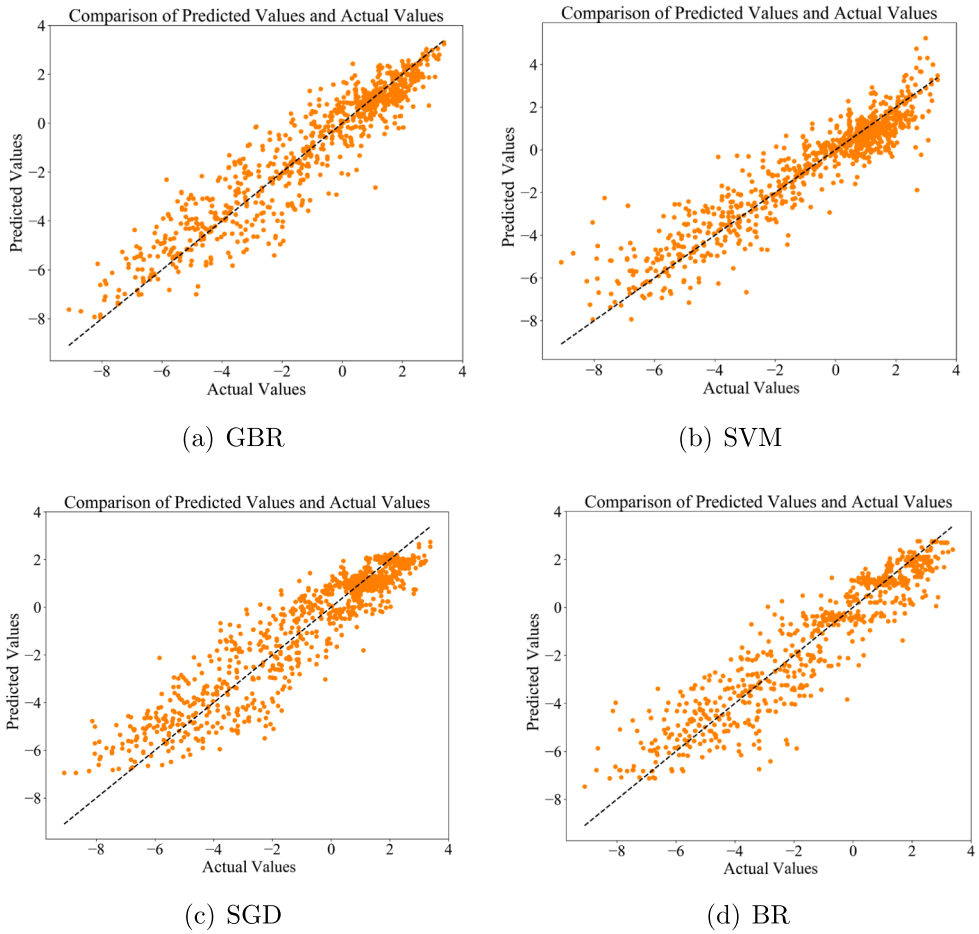

Fig. 9. Comparison of predicted and actual values using GBR, SVM, SGD, and BR models on the hybrid data of UK-Gri and NL-Loo. (The best predicted results are given by GBR).

Table 10
Comparative analysis of different models on the hybrid data of UK-Gri and NL-Loo.

No.	Models	R^2	*MAE*	*RMSE*
1	GBR	**0. 9020**	**0. 5405**	**0. 6860**
2	SVM	0.8521	1.125	1.485
3	SGD	0.8217	1.859	1.659
4	BR	0.7590	1.986	2.243

(Comparison GBR model with other three state-of-the-art models. The highest R^2, lowest *MAE* and *RMSE* are marked with bold font. The best results are given by GBR).

year duration flux and meteorology data of site UK-Gri. During the training process, KFold algorithm was implemented to avoid overfitting. The best regression model was acquired by repeatedly training and minimizing the empirical error. The GBR prediction results are compared with SVM, SGD, BR models in terms of R^2, *MAE*, and *RMSE*, which confirms the superiority of GBR to other three models. RFs were constructed concomitantly with GBR to identify the important environmental variables influencing *NEE* mostly. To verify the performance of the GBR model, the hybrid data of UK-Gri and NL-Loo were employed. The conclusions are summarized as follows:

- The experimental results indicate that GBR model makes more accurate predictions than other three models (SVM, SGD, and BR), which implies that GBR model is better for predicting *NEE* based on one-day interval meteorology data from the sites located in the temperate oceanic climate regions.
- Among the 22 variables, there are four with the importance index more than 0.1 (*LE*, *Rg*, *PAR*, and Ts_{mi}), which can be considered as the four most important variables. Ts_{mi} is found to be more important than its corresponding average observation values, *Ts*. This result reveals that the extremums of environmental variables have vital influence on *NEE*.
- The GBR model acquires good prediction results on the hybrid data from two sites, UK-Gri and NL-Loo, which are located in different geographic positions with similar environments.

It should be noted that this study focuses on one day interval data with one year duration, which is small climate scale, and the study area is a temperate oceanic evergreen needleleaf forest ecosystem. The spatial–temporal scale guarantees the precise prediction for *NEE* dynamics on year-scale evergreen needleleaf forest ecosystems, while limits the model's generalization at the same time. The prediction ability for long-term scales (e.g., years, decades) and other ecosystems (e.g., tropical, dry, and continental climates) may remain questionable. In practical terms, the GBR model provides many tunable hypterparameters and loss functions, which make it more flexible and accurate, when applied to large scale *NEE* predictions. Further, The GBR model can be applied to predict *NEE* from the individual site or from the multiple sites located in different regions with the similar climate type. At the same time, the RFs of environmental variables can also be used to identify the important factors influencing the *NEE* dynamics from the global perspective. Consequently, when upscaling the fluxes from ecosystem to regional or global scale, the combination models should be considered as the valuable tools for predicting and analyzing *NEE*. This

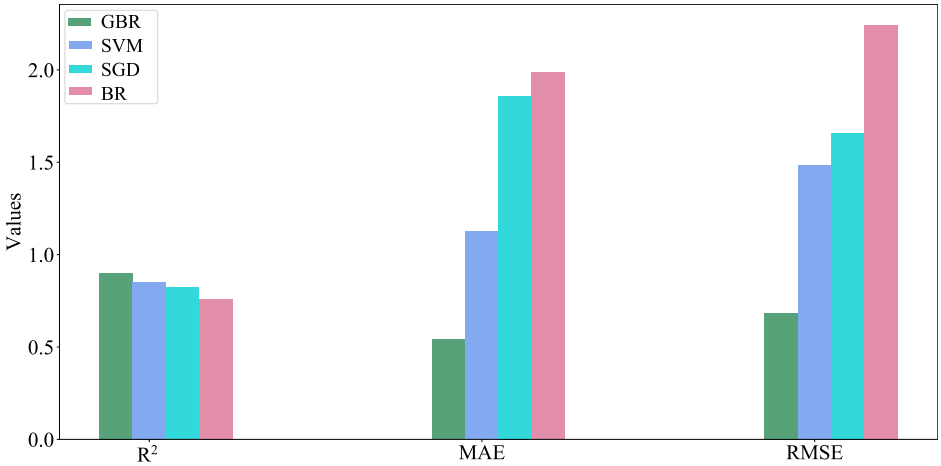

Fig. 10. Comparison plots of GBR, SVM, SGD, and BR models in terms of R^2, *MAE*, and *RMSE* on the hybrid data of UK-Gri and NL-Loo. (GBR gives the highest R^2, lowest *MAE* and *RMSE*, which show GBR model outperforms the others).

would be helpful in understanding the reason behind the climate zones. The knowledge could also be applied to make strategies for environmental restoration. Besides, the prediction ability of GBR could be used to monitor the climate change trend, and to evaluate the ecosystem stability. This is also the future study prospects in the field.

Declaration of Competing Interest

The authors have declared no conflict of interest.

Acknowledgement

We acknowledge the funding supportted by the National Natural Science Foundation of China (41722403), the Natural Science Foundation of Hubei Province (2018CFB259, 2018CFA051), and the Fundamental Research Funds for the Central Universities (2462019YJRC011). This work used eddy covariance data acquired and shared by the FLUXNET community, including these networks: AmeriFlux, AfriFlux, AsiaFlux, CarboAfrica, CarboEuropeIP, CarboItaly, CarboMont, ChinaFlux, Fluxnet-Canada, GreenGrass, ICOS, KoFlux, LBA, NECC, OzFlux-TERN, TCOS-Siberia, and USCCC. The ERA-Interim reanalysis data are provided by ECMWF and processed by LSCE. The FLUXNET eddy covariance data processing and harmonization was carried out by the European Fluxes Database Cluster, AmeriFlux Management Project, and Fluxdata project of FLUXNET, with the support of CDIAC and ICOS Ecosystem Thematic Center, and the OzFlux, ChinaFlux and AsiaFlux offices.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.apenergy.2020.114566.

References

[1] Kirschbaum MU, Mueller R. Net ecosystem exchange: workshop proceedings, cooperative research centre for greenhouse accounting, 2001.
[2] Kane J. Definition of net ecosystem exchange, Website, https://sciencing.com/definition-net-ecosystem-exchange-6802053.html, 2017.
[3] Teklemariam T, Lafleur P, Moore T, Roulet N, Humphreys E. The direct and indirect effects of inter-annual meteorological variability on ecosystem carbon dioxide exchange at a temperate ombrotrophic bog. Agric For Meteorol 2010;150(11):1402–11.
[4] Baldocchi D, Falge E, Gu L, Olson R, Hollinger D, Running S, et al. Fluxnet: A new tool to study the temporal and spatial variability of ecosystem–scale carbon dioxide, water vapor, and energy flux densities. Bull Am Meteorol Soc 2001;82(82):2415–34.
[5] Reichstein M, Falge EM, Baldocchi DD, Papale D, Aubinet M, Berbigier P, et al. On the separation of net ecosystem exchange into assimilation and ecosystem respiration: review and improved algorithm. Global Change Biol 2005;11:1424–39.
[6] Chu X. The response mechanism of ecosystem co2 exchange on precipitation distribution over a supra-tidal wetland in the yellow river delta. Ph.D. thesis, Yantai Institute of Coastal Zone Research, University of Chinese Academy of Sciences, 2018 [In Chinese].
[7] Sun X. Characteristics of net ecosystem exchange and environmental factors of rice-wheat rotation system in the yangtze river delta of china. Chinese J Eco-Agric 2015;23(7):803–11. [In Chinese].
[8] Jung M, Reichstein M, Margolis HA, Cescatti A, Richardson AD, Arain MA, et al. Global patterns of land-atmosphere fluxes of carbon dioxide, latent heat, and sensible heat derived from eddy covariance, satellite, and meteorological observations. J Geophys Res: Biogeosci 2011;116:G00J07.
[9] Allamanis M, Barr ET, Devanbu P, Sutton C. A survey of machine learning for big code and naturalness. ACM Comput Surv 2018;51(4):1–37.
[10] Su Q, Zhu Y, Jia Y, Li P, Hu F, Xu X. Sedimentary environment analysis by grain-size data based on mini batch k-means algorithm. Geofluids 2018;2018:8519695.
[11] Hu F, Wang M, Zhu Y, Liu J, Jia Y. A time simulated annealing-back propagation algorithm and its application in disease prediction. Mod Phys Lett B 2018;32(25):1850303.
[12] Li H, Zhang Z. Mining the intrinsic trends of CO_2 solubility in blended solutions. J CO_2 Utilizat 2018;26:496–502.
[13] Zhang Z, Li H, Chang H, Pan Z, Luo X. Machine learning predictive framework for CO_2 thermodynamic properties in solution. J CO_2 Utilizat 2018;26:152–9.
[14] Li H, Yan D, Zhang Z, Lichtfouse E. Prediction of CO_2 absorption by physical solvents using a chemoinformatics-based machine learning model. Environ Chem Lett 2019;17(3):1397–404.
[15] van Wijk M, Bouten W. Water and carbon fluxes above european coniferous forests modelled with artificial neural networks. Ecol Model 1999;120(2):181–97. https://doi.org/10.1016/S0304-3800(99)00101-5.
[16] Melesse AM, Hanley RS. Artificial neural network application for multi-ecosystem carbon flux simulation. Ecol Model 2005;189(3):305–14. https://doi.org/10.1016/j.ecolmodel.2005.03.014.
[17] He H, Yu G, Zhang L, Sun X, Su W. Simulating CO_2 flux of three different ecosystems in chinaflux based on artificial neural networks. Sci China Series D: Earth Sci 2006;49(2):252–61.
[18] Qin Z, li Su G, en Zhang J, Ouyang Y, Yu Q, Li J. Identification of important factors for water vapor flux and CO_2 exchange in a cropland. Ecol Model 2010;221(4):575–81. https://doi.org/10.1016/j.ecolmodel.2009.11.007.
[19] Safa B, Arkebauer TJ, Zhu Q, Suyker A, Irmak S. Net ecosystem exchange (nee) simulation in maize using artificial neural networks. IFAC J Syst Control 2019;7:100036. https://doi.org/10.1016/j.ifacsc.2019.100036.
[20] Moffat AM, Beckstein C, Churkina G, Mund M, Heimann M. Characterization of ecosystem responses to climatic controls using artificial neural networks. Glob Change Biol 2010;16(10):2737–49. https://doi.org/10.1111/j.1365-2486.2010.02171.x.
[21] Ichii K, Ueyama M, Kondo M, Saigusa N, Kim J, Alberto MC, et al. New data-driven estimation of terrestrial CO_2 fluxes in asia using a standardized database of eddy covariance measurements, remote sensing data, and support vector regression. J Geophys Res: Biogeosci 2017;122(4):767–95.
[22] Liu Y, Zhou G, Du H, Berninger F, Mao F, Li X, et al. Response of carbon uptake to abiotic and biotic drivers in an intensively managed lei bamboo forest. J Environ Manage 2018;223:713–22.
[23] Zhou Q, Fellows A, Flerchinger GN, Flores AN. Examining interactions between and among predictors of net ecosystem exchange: A machine learning approach in a semi-arid landscape. Sci Rep 2019;9(1):2222.
[24] Dou X, Yang Y. Comprehensive evaluation of machine learning techniques for estimating the responses of carbon fluxes to climatic forces in different terrestrial ecosystems. Atmosphere 2018;9(3). https://doi.org/10.3390/atmos9030083.
[25] Dou X, Yang Y, Luo J. Estimating forest carbon fluxes using machine learning

techniques based on eddy covariance measurements. Sustainability 2018;10(1). https://doi.org/10.3390/su10010203.
[26] Díaz G, Coto J, Gómez-Aleixandre J. Prediction and explanation of the formation of the spanish day-ahead electricity price through machine learning regression. Appl Energy 2019;239:610–25.
[27] Hassan MA, Khalil A, Kaseb S, Kassem MA. Exploring the potential of tree-based ensemble methods in solar radiation modeling. Appl Energy 2017;203:897–916.
[28] Wen X, Zhao Z, Deng X, Xiang W, Tian D, Yan W, et al. Applying an artificial neural network to simulate and predict chinese fir (cunninghamia lanceolata) plantation carbon flux in subtropical china. Ecol Model 2014;294:19–26. https://doi.org/10.1016/j.ecolmodel.2014.09.006.
[29] Sun W, Wang Y, Zhang C. Forecasting CO_2 emissions in Hebei, China, through moth-flame optimization based on the random forest and extreme learning machine. Environ Sci Pollut Res 2018;25(29):28985–97.
[30] Xue Y, Chen Y, Hu Y, Chen H. Fuzzy rough set algorithm with binary shuffled frog-leaping (bsfl-frsa): An innovative approach for identifying main drivers of carbon exchange in temperate deciduous forests. Ecol Indicat 2017;83:41–52.
[31] Dou X, Yang Y. Estimating forest carbon fluxes using four different data-driven techniques based on long-term eddy covariance measurements: Model comparison and evaluation. Sci Total Environ 2018;627:78–94. https://doi.org/10.1016/j.scitotenv.2018.01.202.
[32] Friedman JH. Greedy function approximation: a gradient boosting machine. Ann Stat 2001:1189–232.
[33] Breiman L. Random forests. Machine Learn 2001;45(1):5–32.
[34] Falge E, Aubinet M, Bakwin P, Baldocchi D, Berbigier P, Bernhofer C, Black T, et al. Fluxnet marconi conference gap-filled flux and meteorology data, 1992–2000 (2005). https://doi.org/10.3334/ORNLDAAC/811. http://daac.ornl.gov/cgi-bin/dsviewer.pl?ds_id=811.
[35] Fluxnet2015, 2005. https://doi.org/10.18140/FLX/1440178. http://sites.fluxdata.org/NL-Loo/.
[36] Nesselroade KP, Grimm LG. Statistical applications for the behavioral and social sciences. Wiley Online Library; 2019.
[37] Burba G. Eddy covariance method for scientific, industrial, agricultural and regulatory applications: A field book on measuring ecosystem gas exchange and areal emission rates. LI-Cor Biosci, 2013.
[38] McLachlan G, Do K-A, Ambroise C. Analyzing microarray gene expression data vol. 422. John Wiley & Sons; 2005.
[39] Harris D, Harris S. Digital design and computer architecture. Morgan Kaufmann; 2010.
[40] Gini C. Concentration and dependency ratios, English translation in Rivista di. Politica Econ 1997;87(8–9):769–89.
[41] Friedman JH. Stochastic gradient boosting. Comput Stat Data Anal 2002;38(4):367–78.
[42] Willmott CJ. Some comments on the evaluation of model performance. Bull Am Meteorol Soc 1982;63(11):1309–13.
[43] Kottek M, Grieser J, Beck C, Rudolf B, Rubel F. World map of the köppen-geiger climate classification updated. Meteorol Z 2006;15:259–63. https://doi.org/10.1127/0941-2948/2006/0130.
[44] Wikipedia, Köppen climate classification, http://https://en.wikipedia.org/wiki/K%C3%B6ppen_climate_classification, accessed January 6, 2020.
[45] Cortes C, Vapnik V. Support-vector networks. Mach Learn 1995;20(3):273–97.
[46] Chen PH, Lin CJ, Scholkopf B. A tutorial on v-support vector machines. Appl Stochastic Models Bus Ind 2005;21(2):111–36.
[47] Cauchy A. Méthode générale pour la résolution des systemes d'équations simultanées. Comptes Rendus Mathematique Academie des Sciences, Paris 1847;25:536–8.
[48] Polyak BT, Juditsky AB. Acceleration of stochastic approximation by averaging. SIAM J Control Opt 1992;30(4):838–55.
[49] MacKay DJ. Bayesian interpolation. Neural Comput 1992;4(3):415–47.