forests

MDPI

Article

Modeling and Predicting Carbon and Water Fluxes Using Data-Driven Techniques in a Forest Ecosystem

Xianming Dou [1,2] and Yongguo Yang [1,2,*]

[1] Key Laboratory of Coalbed Methane Resources and Reservoir Formation Process of Ministry of Education, China University of Mining and Technology, Xuzhou 221116, China; xianmingdou@cumt.edu.cn

[2] School of Resources and Geosciences, China University of Mining and Technology, Xuzhou 221116, China

* Correspondence: yongguoyang@163.com; Tel.: +86-516-8359-1016

Received: 30 October 2017; Accepted: 8 December 2017; Published: 12 December 2017

Abstract: Accurate estimation of carbon and water fluxes of forest ecosystems is of particular importance for addressing the problems originating from global environmental change, and providing helpful information about carbon and water content for analyzing and diagnosing past and future climate change. The main focus of the current work was to investigate the feasibility of four comparatively new methods, including generalized regression neural network, group method of data handling (GMDH), extreme learning machine and adaptive neuro-fuzzy inference system (ANFIS), for elucidating the carbon and water fluxes in a forest ecosystem. A comparison was made between these models and two widely used data-driven models, artificial neural network (ANN) and support vector machine (SVM). All the models were evaluated based on the following statistical indices: coefficient of determination, Nash-Sutcliffe efficiency, root mean square error and mean absolute error. Results indicated that the data-driven models are capable of accounting for most variance in each flux with the limited meteorological variables. The ANN model provided the best estimates for gross primary productivity (GPP) and net ecosystem exchange (NEE), while the ANFIS model achieved the best for ecosystem respiration (*R*), indicating that no single model was consistently superior to others for the carbon flux prediction. In addition, the GMDH model consistently produced somewhat worse results for all the carbon flux and evapotranspiration (ET) estimations. On the whole, among the carbon and water fluxes, all the models produced similar highly satisfactory accuracy for GPP, *R* and ET fluxes, and did a reasonable job of reproducing the eddy covariance NEE. Based on these findings, it was concluded that these advanced models are promising alternatives to ANN and SVM for estimating the terrestrial carbon and water fluxes.

Keywords: carbon fluxes; evapotranspiration; forest ecosystem; data-driven techniques; group method of data handling; extreme learning machine; adaptive neuro-fuzzy inference system

1. Introduction

The exchanges of carbon dioxide and water vapor fluxes between the atmosphere and forest ecosystems are dominated by a variety of abiotic and biotic factors [1,2]. At present, a large body of available direct measurements of these fluxes as well as their related energy and environmental variables in forest stands using the eddy covariance method are being compiled, archived and distributed [3]. In recent two decades, these measurements have been extensively utilized by the science community to explore the broad mechanisms controlling the dynamic variation of carbon and water fluxes from hourly to decadal time scales [4,5]. In addition, numerous efforts have focused on better quantifying and exploiting the magnitude and behavior involved in the interactions of carbon and water cycles [6,7]. Although remarkable advances have been obtained in our understanding of the nature and mechanisms leading to the ongoing evolution of carbon and water fluxes, their responses to different types of disturbances, such as land use changes, nutrient deposition, CO_2

elevation and fires [8–11] remain uncertain. In this context, accurate estimation of carbon and water fluxes of forest ecosystems is of particular importance for addressing the problems originated from global environmental change, and providing helpful information about carbon and water budgets for analyzing and diagnosing past and future climate change.

Process-based models that characterize complex biophysical and ecophysiological processes of land-atmosphere coupling at different temporal and spatial scales are commonly employed to quantify carbon and water budgets. According to several recent studies in relation to data-model inter-comparison of carbon and water fluxes, the evidence is convincing that appreciable errors do indeed exist between observed and estimated values by process-based models [12–14]. More recently, data-model assimilation techniques have been proposed in order to reduce their predictive errors via assimilating high resolution multi-source remote sensing observations (e.g., surface temperature and soil moisture) into process-based models with various optimization algorithms, such as ensemble Kalman filter, three dimensional and four dimensional variational assimilation methods. However, these ecosystem models depend on a large number of constants and input variables, which are comparatively difficult to be acquired due to high heterogeneity over land surface [15,16]. It is therefore vital that much attention should be given to further enhance the estimates of carbon and water fluxes.

Generally, when addressing the improvement of the predictability of different issues in ecological, hydrological and environmental fields, one of the most important challenges that needs to be considered is the specific non-linear problems in the dynamically varying systems [17,18]. In the last decade, the use of data-driven techniques has become increasingly popular in abovementioned fields [19–21], due to their ability to deal with nonlinearity in time series [22,23]. Among all data-driven techniques, artificial neural network (ANN) and support vector machine (SVM) have been broadly utilized [24–30], aiming at a better understanding the underlying mechanisms that dominate the carbon and water fluxes, and representing the complex processes involved in the carbon and water exchanges. These studies have concentrated primarily on exploiting the relative contributions of the driving forces for the evolution of carbon and water fluxes [24], interpolating the missing carbon and water data according to the flux tower measurements [25], and quantifying the nonlinear processes of carbon and water interactions mainly at site scale [26] and regional scale using the upscaling techniques [27,28]. In addition, these data-driven techniques are increasingly used to correct the carbon and water flux errors estimated by the process-based models [29,30].

With the recent advances in machine learning, many state-of-the-art data-driven techniques have been presented, mainly including group method of data handling (GMDH), generalized regression neural network (GRNN), adaptive neuro-fuzzy inference system (ANFIS) and extreme learning machine (ELM). The ability of these methods to perform with limited climatic inputs in terms of elucidating the complex processes of water vapor fluxes between land surface and the atmosphere, such as reference evapotranspiration (ET) and evaporation, has been adequately demonstrated by previous studies [31–33]. Moreover, these methods are also widely used for regression analysis in many other fields, such as modeling of hydrological time series (e.g., stream-flow and rainfall) [34–36], forecasting of renewable energy (e.g., solar radiation and wind power) [37,38], and modeling of meteorological time series (e.g., air temperature and precipitation) [39,40]. In contrast, comparatively few studies relevant to these methods have been carried out to map the actual ET and carbon fluxes, here including gross primary production (GPP), ecosystem respiration (R) and net ecosystem exchange (NEE), in terrestrial ecosystems based on the data measured at the flux tower sites by the eddy covariance technique.

Despite the considerable technological efforts involved in addressing the aforementioned issues of carbon and water flux exchanges, however, relatively little work has been conducted to compare the ability of these traditional techniques (ANN and SVM). Moreover, with the increasing number of novel invented methods, the need for such a comprehensive comparison of conventional with advanced methods should to be more crucial. Namely, it can be considered as a helpful benchmark for ensuring

the estimated accuracy of solving different issues by properly selecting an effective method from a variety of machine learning techniques. Therefore, the comparison of these methods, including ANN, SVM, GRNN, GMDH, ELM and ANFIS was carefully considered in this study.

Our major goal is to utilize six different data-driven techniques based on machine learning to estimate the carbon and water fluxes with continuous six-year observation data from a flux tower site in a forest ecosystem. The specific aims are threefold as follows: (1) to investigate the feasibility and capability of various data-driven models, including GRNN, GMDH, ELM and ANFIS, for simulating the daily carbon and water fluxes at the ecosystem level; (2) to demonstrate that these newly proposed modern models can be used as desirable complements to the traditionally accepted ANN and SVM models; and (3) to examine the modeling differences between the ET and three primary components of carbon fluxes (GPP, *R* and NEE). In addition, this study focuses on examining the modeling ability of aforementioned models only at a single flux tower site. It should be pointed out that these methods can also be used to upscale the carbon and water fluxes from site to regional scale with remote sensing data. However, it is beyond the scope of the present work and will be carried out in our follow-up investigation.

2. Materials and Methods

2.1. Site Description and Data Presentation

As part of our on-going research, carbon and water flux measurements for a forest ecosystem are being carried out at a flux tower site in Canada. This study site (53.99° N, 105.12° W) is situated approximately 80 km east-northeast of Prince Albert National Park, Saskatchewan, Canada. This site belongs to the representative boreal evergreen needle-leaf forest and it is predominantly covered by black spruce regenerated after a fire in 1879. Its average canopy height was 7.2 m. It had an average stand age of 110 years in 2004, ranging from 91 to 130 years. Based on the statistics from the nearest long-term weather station during 1971 to 2000 [41], mean annual air temperature (T_a) and cumulative annual precipitation were 0.4 °C and 467 mm, respectively. More descriptions regarding the site history, topography, soil, and vegetation can be found in Swanson and Flanagan [42]. At this site, continuous half-hourly carbon and water vapor fluxes have been measured by using the eddy covariance technique. Half-hourly climate variables were also observed, mainly including T_a, soil temperature (T_s), net radiation (R_n), relative humidity (R_h), wind speed and volumetric soil moisture content. Further details on the characteristics of the used instruments and measurement procedure of CO_2 flux can be found in Krishnan et al. [43].

The method of gap-filling for eddy covariance-measured NEE and its partitioning into gross primary productivity (GPP) and *R* was in accordance with the Fluxnet-Canada Research Network protocol [44]. In brief, *R* was calculated from nighttime and cold-season NEE. The gap-filling of night *R* and estimation of daytime *R* was based on an empirical relationship between *R* and T_s at a shallow depth (2 cm). GPP was derived according to the daytime *R* and daytime-measured NEE. During the cold season, GPP was set to zero. More details regarding NEE gap-filling and its separation scheme as well as the random errors in NEE, *R* and GPP are given in Krishnan et al. [43]. In addition, as stated above, both GPP and *R* were not measured directly by the eddy covariance technique but were calculated from the eddy covariance-measured NEE. In the following sections, for the sake of distinguishing these calculated GPP and *R* from modeled or predicted GPP and *R* by our proposed models, the estimated GPP and *R* based on NEE are hereinafter referred to as measured or observed GPP and *R*.

Daily values of environmental variables during the study period are shown in Figure 1. The available data during the period of six years (2004–2009) were divided into three parts, respectively for training (2004–2007), validation (2008) and testing (2009). Table 1 shows the statistical parameters of daily data in the three datasets. Based on the table, it can be seen that each segment has similar statistical features. The climate variables correlate strongly with each flux. However, the correlation values for GPP, *R* and ET are generally greater than those of values for NEE, which may lead to

difficulty in accurately estimating the ET. The meteorological variables, T_a, T_s and R_n, show strong positive correlation with GPP, R and ET, while present negative correlation with NEE.

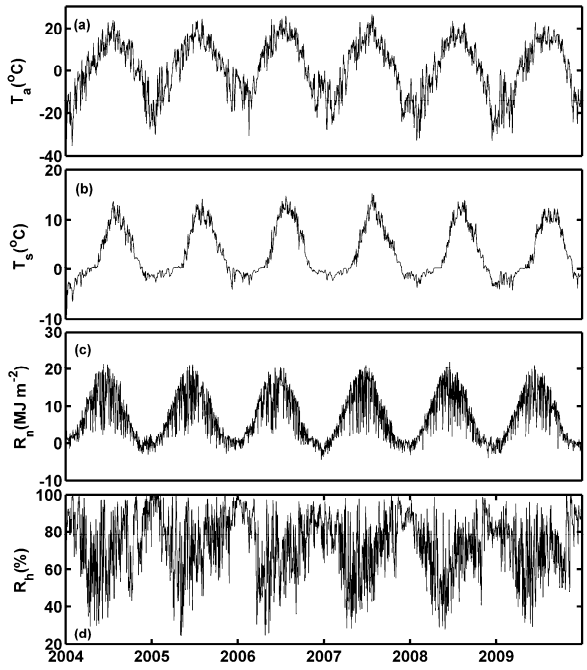

Figure 1. Daily values of environmental variables during the study period. (**a**) Air temperature above the canopy (T_a); (**b**) 5-cm depth soil temperature (T_s); (**c**) net radiation (R_n); (**d**) relative humidity (R_h).

Table 1. Statistical parameters of flux tower-measured daily environmental variables including air temperature (T_a, °C), net radiation (R_n, mol·m^{-2}), relative humidity (R_h, %), soil temperature (T_s, °C), gross primary productivity (GPP, g·C·m^{-2}·Day^{-1}), ecosystem respiration (R, g·C·m^{-2}·Day^{-1}), net ecosystem exchange (NEE, g·C·m^{-2}·Day^{-1}) and evapotranspiration (ET, mm·Day^{-1}).

Period	Variable	X_{mean}	X_{max}	X_{min}	X_{sd}	X_{ku}	X_{sk}	CC_{GPP}	CC_R	CC_{NEE}	CC_{ET}
Training	T_a	1.87	26.37	−35.31	12.68	2.27	−0.36	0.83	0.81	−0.56	0.79
	R_n	6.59	21.39	−4.28	6.27	2.04	0.51	0.74	0.59	−0.70	0.82
	R_h	74.05	100.00	24.56	17.24	2.52	−0.57	−0.38	−0.24	0.45	−0.43
	T_s	3.31	15.41	−9.18	5.00	2.04	0.56	0.87	0.94	−0.46	0.80
	GPP	2.16	7.94	−0.06	2.44	1.91	0.66	1.00	0.90	−0.78	0.91
	R	1.71	6.88	0.00	1.69	2.63	0.96	0.90	1.00	−0.44	0.82
	NEE	−0.45	3.90	−4.12	1.16	2.96	−0.78	−0.78	−0.44	1.00	−0.71
	ET	0.81	4.05	−0.03	0.83	2.99	1.02	0.91	0.82	−0.71	1.00
Validation	T_a	−0.49	23.47	−33.01	14.07	2.12	−0.41	0.83	0.80	−0.64	0.81
	R_n	6.71	21.90	−3.86	6.49	1.89	0.39	0.69	0.51	−0.72	0.74
	R_h	70.07	98.86	28.29	15.90	2.52	−0.47	−0.32	−0.15	0.45	−0.35
	T_s	3.12	14.12	−3.67	5.05	1.93	0.60	0.88	0.95	−0.56	0.84
	GPP	2.43	9.41	−0.15	2.82	1.87	0.67	1.00	0.91	−0.85	0.93
	R	1.79	7.28	0.04	1.78	2.82	1.01	0.91	1.00	−0.55	0.87
	NEE	−0.64	3.01	−5.20	1.42	2.57	−0.77	−0.85	−0.55	1.00	−0.76
	ET	0.91	4.01	−0.01	0.95	2.93	0.97	0.93	0.87	−0.76	1.00
Testing	T_a	−0.97	22.81	−29.72	13.73	1.70	−0.08	0.87	0.84	−0.64	0.84
	R_n	7.16	20.98	−2.72	5.38	2.31	0.49	0.68	0.49	−0.72	0.76
	R_h	67.80	96.54	29.62	13.88	2.70	−0.07	0.00	0.16	0.24	−0.09
	T_s	2.79	12.37	−4.15	4.87	1.84	0.63	0.88	0.94	−0.52	0.80
	GPP	2.29	8.42	−0.05	2.74	1.88	0.69	1.00	0.90	−0.81	0.93
	R	1.58	6.77	−0.26	1.84	2.54	0.96	0.90	1.00	−0.48	0.83
	NEE	−0.71	3.02	−4.43	1.33	2.99	−0.86	−0.81	−0.48	1.00	−0.78
	ET	0.86	3.79	−0.04	0.91	2.59	0.92	0.93	0.83	−0.78	1.00

Note: X_{mean}, X_{max}, X_{min}, X_{sd}, X_{ku} and X_{sk} refer to the mean, maximum, minimum, standard deviation, kurtosis and skewness of each variable, respectively; CC_{GPP}, CC_R, CC_{NEE} and CC_{ET} refer to the correlation coefficient between each variable and GPP, R, NEE and ET, respectively.

2.2. Data-Driven Techniques

2.2.1. Artificial Neural Network

ANN models with the similar functions and structures to the human nervous system have great parallel computational capability in dealing with the complex nonlinear processes in different natural systems [45]. ANN method is widely recognized as an important supervised technique for addressing various issues, such as regression, classification and pattern recognition, in a wide range of fields. A feed-forward ANN with a single hidden layer was carefully examined in the present work, due to its adequate functional approximation ability in the practical applications [31,46,47]. In addition, the widely applied back-propagation algorithm was also used with the intention of obtaining the optimum model parameters (weights and thresholds) through a large amount of iterative processes. During the back-propagation learning procedure for a specific network, its related topology structures, key inner parameters and algorithms were ultimately determined by the trial and error method in order to efficiently achieve a desired error range. The network error can be calculated using the mean squared error (MSE):

$$\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N}(o_i - t_i)^2 \quad (1)$$

where N is the number of the examples; o_i is the network output and t_i is the target.

2.2.2. Support Vector Machine

The SVM method is a popular computational intelligence technique and has been employed in many research fields [48–50]. The versatile functions and powerful generalization ability of SVM technique in dealing with complex nonlinear problems are mainly attributed to the theoretical foundation based on both the statistical learning theory and structural risk minimization principle. Its fundamental principle described in detail can be found in Cortes and Vapnik [51]. The kernel function is used to calculate the inner product through mapping the inputs into high-dimensional feature spaces and it is universally accepted as an innovative and crucial trick for the SVM method. To effectively capture useful information from training data, the choice of proper kernel function is particularly important. The radial basis function (RBF) is a simple, efficient, reliable and extensively adopted kernel and its role in the generalization ability of SVM method has been proven by previous studies [49,52]. Therefore, in the present study, RBF algorithm was used for the developed SVM models to approximate the nonlinear processes of the input-output space. In addition, to obtain an optimal SVM model, its related parameters including regularization factor, insensitive error band width and kernel width should be taken into account. In this research, the values of both regularization factor and kernel width were determined through the grid search approach [53]. The value of insensitive error band width was set to 0.01 by default.

2.2.3. Adaptive Neuro-Fuzzy Inference System

ANFIS approach, viewed as a hybrid data-driven technique, was proposed by Jang [54] on the basis of the abovementioned ANN and fuzzy inference system (FIS). Specifically, the powerful capability of self-learning of supervised neural network is systematically integrated with the unsupervised fuzzy system. The sophisticated ANFIS technique is capable of qualitatively and quantitatively characterizing the nonlinear processes of carbon and water interactions as shown later in this study. Accordingly, ANFIS is a computationally more efficient model and also can generate more robust results than conventional data-driven models (e.g., ANN). Given a training data set, a primary task is to appropriately choose an FIS. At present, Takagi and Sugeno [55] and Mamdani [56] are two major methods for building FISs. The former is considered to be more computationally efficient and compact, due to its superiority over the latter regarding the calculation of consequent parts. Consequently, Takagi and Sugeno-based ANFIS models were developed by the present research.

Optimizing the parameters in the initial ANFIS is another important task, which can be performed by the hybrid learning algorithm using two alternating passes. More specifically, the least-squares method is conducted with the objective of identifying the consequent parameters related to the outputs of the system in the forward pass, whereas the gradient descent algorithm is used in the backward pass for optimizing the premise parameters involved in the membership functions. In addition, especially noteworthy is the fact that the efficiency and potential of ANFIS is dependent on the methods of generating FIS. Although the grid partitioning method has been broadly adopted, the usage of this approach is appreciably hindered due to the curse of dimensionality. Therefore, the fuzzy c-means clustering (FCM) algorithm was employed in this investigation. Further information about FCM algorithm may be found in Dariane and Azimi [35].

2.2.4. Generalized Regression Neural Network

GRNN firstly proposed by Specht [57] is considered as a new extension to the feed-forward radial basis function neural network. It can be used to deal with the regression tasks and thus has been widely utilized as an attractive alternative to other data-driven techniques. Compared with the traditional ANN method, the GRNN has several distinct advantages, mainly including: (1) the simplicity of the network architecture that strongly relies upon the number of the used variables as well as the sample size of a given training dataset, and does not require to be amended during the learning; (2) the fast learning speed in that the GRNN does not need to be trained by using an iterative process which must be implemented through the back-propagation algorithm for ANN with an enormous amount of computation; and (3) the solution of the local minimum problem based on the radial basis function in the pattern layer. In addition to these advantages, another especially noteworthy strength of the GRNN is that it does not require intensive efforts to determine its intrinsic functions or parameters, except for the smoothing factor that appreciably impacts the model generalization ability. To date, there has been no universally accepted approach for obtaining the best smoothing factor for the purpose of ensuring the modeling precision of the GRNN model. In the current research, an iterative procedure with fourfold cross validation was adopted to determine the optimal smoothing factor which was set in the range of 0.01 to 1 based on our experience and the findings from other studies in the applications of the GRNN [58,59].

2.2.5. Extreme Learning Machine

ELM is a relatively new data-driven technique and has received great attention within the last few years. Unlike conventional computational intelligence methods, ELM was innovatively designed by Huang et al. [60] with a single-hidden-layer feed-forward network structure. More importantly, the hidden nodes are independent of the training dataset and their related parameters do not need to be updated during the model training procedure. Nevertheless, the ELM method has been used successfully to deal with regression and classification problems [61]. In addition, another noteworthy point is that the ELM method is able to handle the nonlinear problems in a comparatively shorter time compared with other data-driven methods and thus is recommended for addressing the emergent issues that should be urgently solved, such as real-time flood forecasting and drought dynamic monitoring and prediction [62,63]. In the present study, for all the developed ELM models, the widely used sigmoid activation function was adopted for the hidden layer, while the linear function was undertaken for the output layer. The detailed principle of ELM method can be found in Huang et al. [60].

2.2.6. Group Method of Data Handling

GMDH method proposed by Ivakhnenko [64] is an extremely high-order polynomial neural network. The network topology of this inductive learning approach can be flexibly developed through the heuristic self-organization algorithm. Specifically, given a training dataset, the best network architecture of a GMDH model is obtained automatically, while for a traditional backpropagation-based

ANN model, it must be determined commonly according to previous experience and a time-consuming trial and error method. Additionally, the possible impact induced by the existing outliers of the training samples does not need to be taken into account for the GMDH method, and the redundancy problem involved in input variables is solved by a sequential learning algorithm. In a GMDH network, all the hidden nodes which act as activation functions are respectively connected by two inputs and a single output, and each node can be represented by a second-rate two-variable polynomial regression equation with five weights and one bias. After achieving an optimal complex network, a high-order iterative polynomial series related to all the nodes used in the network can be generated. The unknown weighting coefficients of this series are solved by the ordinary least square regression method. The fundamental principle of GMDH algorithm in more detail can be found in Müller et al. [65].

2.3. Model Development and Software Availability

Six different data-driven modeling techniques were developed and compared for the present investigation. In order to achieve an accurate and reasonable comparison of different fluxes estimates, all the carbon and water fluxes were modeled and predicted with the same input variables (T_a, T_s, R_n, R_h), according to the strong correlations between environmental variables and each flux as shown in Table 1. Before the training of each applied model, these input variables and the corresponding output variable were normalized to a range between 0 and 1.

In this study, the LIBSVM package written by Chang and Lin [53] was used to develop the SVM model, which is one of the most popular packages. It can be downloaded from https://www.csie.ntu.edu.tw/~cjlin/libsvm/. The ELM model was developed based on the software package written by Huang et al. [60], which can be downloaded from http://www.ntu.edu.sg/home/egbhuang/. In developing other models, MATLAB software (version 8.2, The MathWorks, Inc., Natick, MA, USA) and its toolboxes, including Neural Network Toolbox 8.1 and Fuzzy Logic Toolbox 2.2.18, were utilized.

2.4. Model Evaluation

The performance of developed data-driven models was evaluated on the basis of several statistical indices, including the coefficient of determination (R^2), Nash-Sutcliffe efficiency (NSE), root mean square error (RMSE) and mean absolute error (MAE). The expressions of the abovementioned statistical indices are described as below:

$$R^2 = \left[\frac{\sum_{i=1}^{N} (Y_{o,i} - \overline{Y_o})(Y_{m,i} - \overline{Y_m})}{\sqrt{\sum_{i=1}^{N} (Y_{o,i} - \overline{Y_o})^2 \sum_{i=1}^{N} (Y_{m,i} - \overline{Y_m})^2}} \right]^2 \quad (2)$$

$$\mathrm{NSE} = 1 - \frac{\sum_{i=1}^{N} (Y_{o,i} - Y_{m,i})^2}{\sum_{i=1}^{N} (Y_{o,i} - \overline{Y_o})^2} \quad (3)$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (Y_{o,i} - Y_{m,i})^2} \quad (4)$$

$$\mathrm{MAE} = \frac{1}{N} \sum_{i=1}^{N} |Y_{o,i} - Y_{m,i}| \quad (5)$$

where Y_o and Y_m denote the observed and modeled values, respectively; $\overline{Y_o}$ and $\overline{Y_m}$ are the means of observed and modeled values, respectively; and N is the number of observed values.

3. Results

3.1. GPP Modeling Using the Machine Learning Methods

The performance indices, including R^2, NSE, RMSE and MAE, are used in order to evaluate the efficiency of applied models in predicting the daily GPP, *R*, NEE and ET. R^2 and NSE with larger values and RMSE and MAE with smaller values indicate higher model efficiency. The estimated results in predicting the daily GPP by the six data-driven methods (ANN, GRNN, ELM, ANFIS, SVM and GMDH) for the training, validation and testing periods are summarized in Table 2. It is shown in the table that the ANN model performs superior to the other models in the estimation of daily GPP in the testing period, with the highest values of R^2 and NSE and the lowest values of RMSE and MAE. The SVM model performs slightly inferior to the ANN model and gives the second-best efficiency among all the models. However, the GMDH model yields the worst precision with respect to NSE, RMSE and MAE metrics.

Table 2. Comparisons of data-driven models for gross primary productivity (GPP, g·C·m^{-2}·Day^{-1}) for the training, validation and testing periods.

Model	Training				Validation				Testing			
	R^2	NSE	RMSE	MAE	R^2	NSE	RMSE	MAE	R^2	NSE	RMSE	MAE
ANN	0.9491	0.9491	0.5500	0.3633	0.9565	0.9248	0.7722	0.5063	0.9622	0.9427	0.6548	0.4220
GRNN	0.9705	0.9703	0.4199	0.2461	0.9507	0.9186	0.8034	0.4880	0.9443	0.9254	0.7473	0.4364
ELM	0.9119	0.9119	0.7234	0.5254	0.9289	0.8897	0.9350	0.6667	0.9312	0.9098	0.8218	0.5800
ANFIS	0.9460	0.9460	0.5661	0.3588	0.9521	0.9224	0.7845	0.4878	0.9570	0.9341	0.7024	0.4394
SVM	0.9553	0.9552	0.5156	0.3091	0.9556	0.9242	0.7750	0.4936	0.9574	0.9361	0.6915	0.4290
GMDH	0.9041	0.9040	0.7549	0.5962	0.9157	0.8809	0.9715	0.7308	0.9323	0.8894	0.9100	0.6691

According to the aforementioned four evaluation indices, the overall performance rankings of the applied models for carbon and water fluxes for the testing period are given in Table 3. It can be seen from the table that the model performance rankings for GPP prediction can be ranked as follows: ANN, SVM, ANFIS, GRNN, ELM and GMDH.

Table 3. Model performance rankings of the data-driven models for carbon and water fluxes for the testing period.

Model	GPP	*R*	NEE	ET
ANN	1	4	1	4
GRNN	4	5	3	5
ELM	5	3	5	3
ANFIS	3	1	4	2
SVM	2	2	2	1
GMDH	6	6	6	6

Note: Model performance rankings were determined according to the R^2, NSE, RMSE and MAE indices. For example, for GPP, the ANN performs the best (No. 1), while the GMDH gives the worst accuracy (No. 6).

The comparisons of daily GPP between measured and predicted using the data-driven models over the testing period in the form of scatter plot are illustrated in Figure 2. As shown in Figure 2, the fit lines of ANN and SVM models (slope = 0.86) are closer to the ideal fit lines (1:1 lines), whereas for the GMDH model, the fit line (slope = 0.76) is more inconsistent with the ideal fit line than those for the other models. Furthermore, the estimates from both the ELM and GMDH models fail to match the corresponding observed values around their minimum values.

Although obvious performance differences for GPP prediction exist among the applied models (Table 2), these models provide similar scattered estimates as a whole (Figure 2). Accordingly, we describe only the seasonal and annual variation in carbon and water fluxes estimated by the ANN model. Figure 3 compares the measured and predicted values for daily GPP, *R*, NEE and ET using the

ANN model among the training, validation and testing periods. The over- and under-estimation of the ANN model are clearly demonstrated in the figure. As shown from Figure 3a, the simulated values of daily GPP consistently follow the measured ones, while the under-estimation of the ANN model is clearly seen in the peaks, especially during the validation and testing periods.

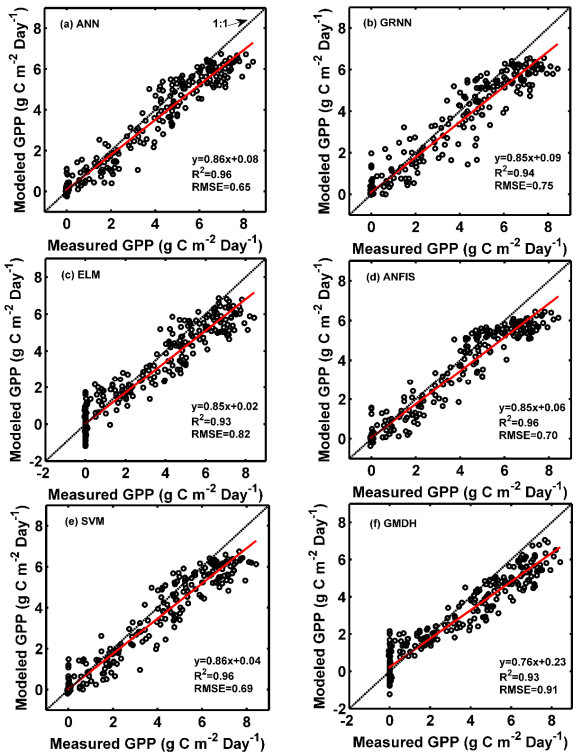

Figure 2. Comparisons of daily GPP measured by eddy covariance technique and predicted by data-driven models for the testing period. (**a**) ANN model; (**b**) GRNN model; (**c**) ELM model; (**d**) ANFIS model; (**e**) SVM model; and (**f**) GMDH model.

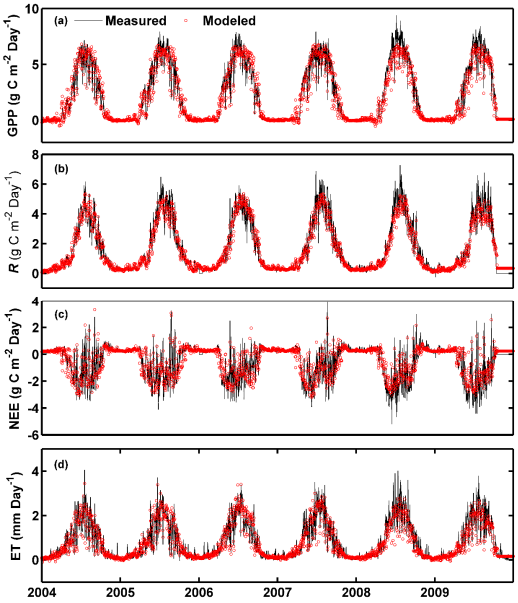

Figure 3. Eddy covariance-measured and ANN model-simulated daily carbon and water fluxes in the whole period from 2004 to 2009. (**a**) GPP; (**b**) *R*; (**c**) NEE; and (**d**) ET.

Furthermore, the comparisons of annual total carbon and water fluxes measured by eddy covariance technique and predicted by different data-driven models in the whole period are shown in Figure 4. As shown in Figure 4a, there is a slight difference among the applied models, which are able to reproduce most of the observed annual GPP. However, in the prediction period (2009), annual total GPP measured by the eddy covariance technique is under-estimated by 10% to 14%.

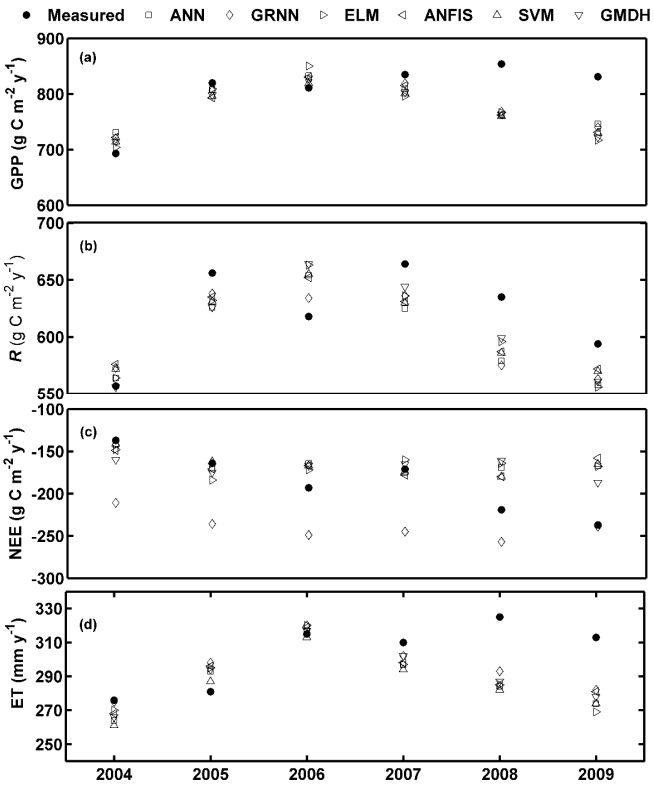

Figure 4. Comparisons of annual total carbon and water fluxes measured by eddy covariance technique and predicted by data-driven models in the whole period. (**a**) GPP; (**b**) *R*; (**c**) NEE; and (**d**) ET.

3.2. R Modeling Using the Machine Learning Methods

The estimated accuracies in the prediction of the daily *R* by the data-driven methods used in this study are shown in Table 4. Similar to the GPP forecasting, the GMDH model in estimating the daily *R* also gives the lowest model efficiency in the testing period, while the ANFIS model provides the best performance based on the R^2, NSE, RMSE and MAE metrics. In addition, the overall performance ranks of the data-driven models used in the present study in the testing period can be concluded as follows: ANFIS, SVM, ELM, ANN, GRNN and GMDH (Table 3).

Table 4. Comparisons of data-driven models for ecosystem respiration (*R*, g·C·m^{-2}·Day^{-1}) for the training, validation and testing periods.

Model	Training				Validation				Testing			
	R^2	NSE	RMSE	MAE	R^2	NSE	RMSE	MAE	R^2	NSE	RMSE	MAE
ANN	0.9373	0.9366	0.4250	0.2656	0.9421	0.9210	0.5000	0.3209	0.9489	0.9303	0.4848	0.3377
GRNN	0.9637	0.9636	0.3223	0.1895	0.9360	0.9128	0.5254	0.3416	0.9413	0.9293	0.4880	0.3423
ELM	0.9297	0.9297	0.4477	0.2874	0.9411	0.9280	0.4775	0.3061	0.9422	0.9310	0.4821	0.3289
ANFIS	0.9391	0.9391	0.4167	0.2619	0.9418	0.9241	0.4902	0.3084	0.9498	0.9393	0.4522	0.3199
SVM	0.9414	0.9413	0.4089	0.2334	0.9425	0.9257	0.4850	0.3102	0.9471	0.9347	0.4692	0.3385
GMDH	0.9277	0.9277	0.4541	0.2958	0.9348	0.9203	0.5023	0.3343	0.9307	0.9099	0.5511	0.3969

As shown in Figure 5, the regression fit line between observed and ANFIS predicted for daily *R* in the testing period (slope = 0.85) is closer to the exact fit line (1:1 line). However, the fit line of the GMDH model (slope = 0.79) tends to deviate from the ideal fit line more than those for the other models. Moreover, as illustrated in Figure 3b, the observed values of daily *R* well match the predicted ones using the ANN model during the whole period, whereas the peaks during the validation and testing periods are again significantly underestimated. In addition, according to the annual estimates by the applied models, the total *R* in 2009 is under-estimated by 4% to 6% (Figure 4b).

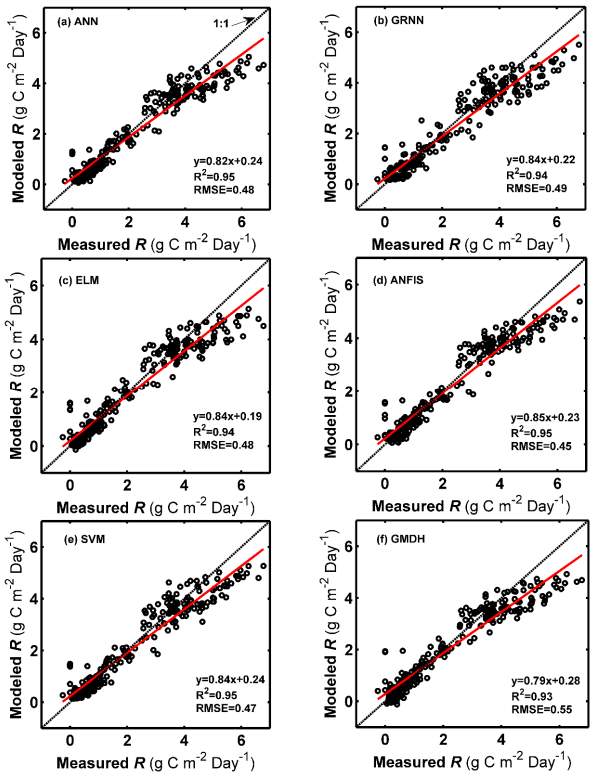

Figure 5. Comparisons of daily *R* measured by eddy covariance technique and predicted by data-driven models for the testing period. (**a**) ANN model; (**b**) GRNN model; (**c**) ELM model; (**d**) ANFIS model; (**e**) SVM model; and (**f**) GMDH model.

3.3. NEE Modeling Using the Machine Learning Methods

Table 5 shows the predicted accuracies of daily NEE by the data-driven models employed in the present study. As clearly seen from the table, the ANN and GMDH models provide the best and worst model efficiency in estimating the NEE, respectively, which are similar to previous GPP predictions. In addition, according to R^2, NSE, RMSE and MAE criteria in the testing period, the overall performance ranks of the models are summarized as follows: ANN, SVM, GRNN, ANFIS, ELM and GMDH (Table 3).

As shown in Figure 6, similar to the aforementioned GPP and *R* estimation, the fit line of the GMDH model (slope = 0.50) also deviates from the ideal fit line more than those for the other models. However, the fit lines of ANN, ANFIS and SVM for daily NEE forecasting in the testing period more closely match the ideal fit line (1:1 lines). Additionally, it is clearly seen from Figure 3c that most of the modeled values of daily NEE closely follow the corresponding observed ones among the three periods, while the peaks are underestimated, mainly in the validation and testing periods. Furthermore, according to the annual estimates by the applied models, the total NEE in 2009 is under-estimated by 21% to 33%, except for the estimates from the GRNN model (Figure 4c).

Table 5. Comparisons of data-driven models for net ecosystem exchange (NEE, g·C·m^{-2}·Day^{-1}) for the training, validation and testing periods.

Model	Training				Validation				Testing			
	R^2	NSE	RMSE	MAE	R^2	NSE	RMSE	MAE	R^2	NSE	RMSE	MAE
ANN	0.7962	0.7961	0.5238	0.3552	0.8482	0.7946	0.6429	0.4370	0.8352	0.7784	0.6270	0.4399
GRNN	0.8819	0.8399	0.4641	0.3668	0.8037	0.7492	0.7103	0.5436	0.8053	0.7665	0.6436	0.4484
ELM	0.6754	0.6754	0.6610	0.4944	0.7611	0.6937	0.7850	0.5716	0.7216	0.6700	0.7651	0.5640
ANFIS	0.7745	0.7745	0.5509	0.3724	0.8302	0.7904	0.6494	0.4321	0.7891	0.7322	0.6893	0.4922
SVM	0.8130	0.8128	0.5019	0.3139	0.8406	0.7990	0.6358	0.4313	0.8175	0.7673	0.6426	0.4746
GMDH	0.6191	0.6187	0.7164	0.5445	0.6701	0.6145	0.8806	0.6726	0.6692	0.6043	0.8379	0.5712

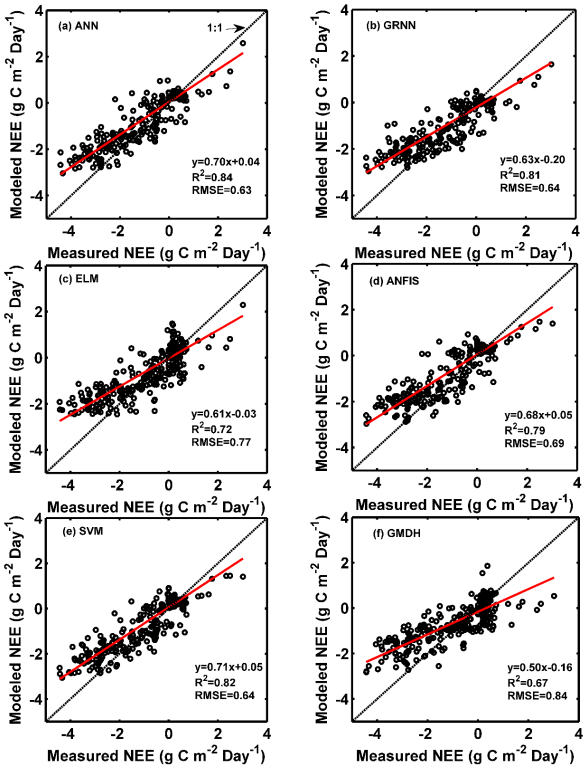

Figure 6. Comparisons of daily NEE measured by eddy covariance technique and predicted by data-driven models for the testing period. (**a**) ANN model; (**b**) GRNN model; (**c**) ELM model; (**d**) ANFIS model; (**e**) SVM model; and (**f**) GMDH model.

3.4. ET Modeling Using the Machine Learning Methods

Table 6 summarizes the modeled accuracies of daily ET estimated by the data-driven models. The GMDH model yields the worst performance in estimating daily ET, which is consistent with the previous forecasting of carbon fluxes. However, the other models perform better than the GMDH model and they have similar model efficiency with respect to R^2, NSE, RMSE and MAE statistics.

It is clear from Figure 7 that the fit line of the GMDH model in the prediction of daily ET tends to deviate from the ideal fit line, while the fit lines of other models are consistently closer to the ideal fit line. As shown in Figure 3d, the simulated values of daily ET estimated by the ANN, ANFIS and SVM models well match the corresponding measured values for the training, validation and testing periods. However, the appreciable underestimations of the peaks from the ANN, ANFIS and SVM models, mainly in the validation and testing periods, are clearly seen. Moreover, according to the annual estimates by the applied models, the total ET in 2009 is under-estimated by 10% to 14% (Figure 4d).

Table 6. Comparisons of data-driven models for evapotranspiration (ET, mm·Day^{-1}) for the training, validation and testing periods.

Model	Training				Validation				Testing			
	R^2	NSE	RMSE	MAE	R^2	NSE	RMSE	MAE	R^2	NSE	RMSE	MAE
ANN	0.9172	0.9171	0.2387	0.1659	0.9206	0.8743	0.3378	0.2160	0.9105	0.8759	0.3210	0.2169
GRNN	0.9188	0.9180	0.2375	0.1667	0.9139	0.8746	0.3374	0.2171	0.9065	0.8739	0.3236	0.2193
ELM	0.9095	0.9095	0.2495	0.1768	0.9186	0.8761	0.3353	0.2233	0.9100	0.8780	0.3183	0.2114
ANFIS	0.9152	0.9152	0.2416	0.1671	0.9163	0.8768	0.3344	0.2106	0.9109	0.8817	0.3134	0.2129
SVM	0.9259	0.9254	0.2265	0.1480	0.9171	0.8766	0.3347	0.2081	0.9120	0.8775	0.3190	0.2129
GMDH	0.8920	0.8920	0.2726	0.1950	0.9032	0.8617	0.3543	0.2379	0.8950	0.8480	0.3553	0.2551

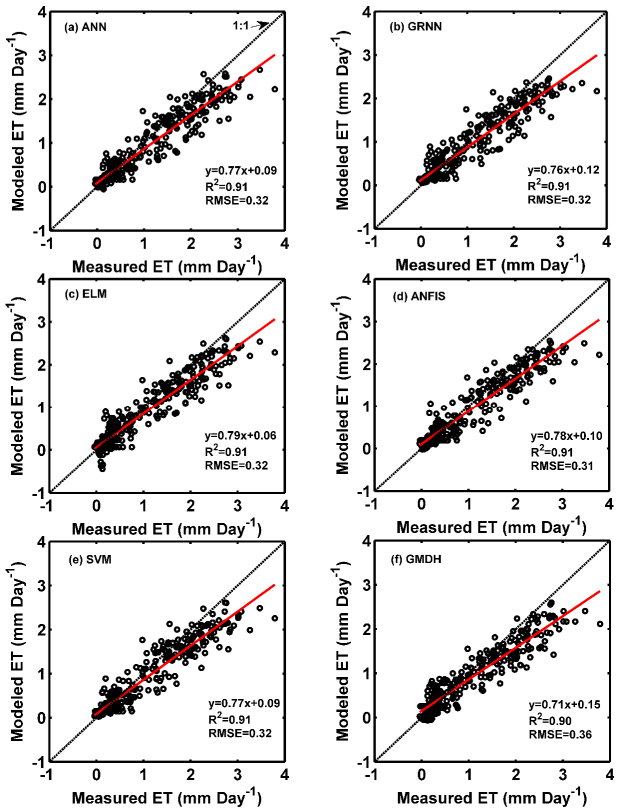

Figure 7. Comparisons of daily ET measured by eddy covariance technique and predicted by data-driven models for the testing period. (**a**) ANN model; (**b**) GRNN model; (**c**) ELM model; (**d**) ANFIS model; (**e**) SVM model; and (**f**) GMDH model.

4. Discussion

The present study, for the first time to our best knowledge, investigated the adaptability and validity of a variety of machine learning techniques, including GRNN, ELM, ANFIS and GMDH, for modeling and predicting the terrestrial carbon and water fluxes for a forest ecosystem based on the data measured with the eddy covariance technique. In addition, to assess the generalization ability of all the approaches in our research, two conventional data-driven modeling techniques, namely ANN and SVM, were also employed as benchmarks. Furthermore, several performance indices involving R^2, NSE, RMSE and MAE were adopted for the model evaluation in order to adequately manifest the efficiency of the applied models. In the following subsections, we concentrate mainly on discussing the predictive capability of various data-driven models and the discrepancy of different carbon and water fluxes in modeling ability, and providing the limitations of the current study and its potential improvements for future research.

Our predictive results in the testing periods from Table 2 and Tables 4–6, demonstrated that a large amount of diurnal variance in each carbon and water flux was accounted for by our used models, with average 95%, 94%, 77% and 91% for GPP, *R*, NEE and ET, respectively. Therefore, these machine learning techniques have adequate capability to describe the complex interactions between the carbon and water fluxes and environmental factors. Moreover, previous studies have also proved the effectiveness of data-driven models, primarily including ANN and SVM methods, for the terrestrial carbon and water flux prediction at ecosystem level [66–69]. In recent years, another important data-driven technique, namely regression tree method, has been successfully utilized to estimate the carbon fluxes [28,70]. Beer et al. [28] used model tree ensemble and ANN models to estimate the spatial distributions of global GPP, and found that these two models obtained similar estimates, which are comparable to those of both process-based and atmospheric inversion models. Xiao et al. [70] utilized regression tree algorithm to exclusively upscale NEE from flux tower to the continental scale with remotely sensed and AmeriFlux data. They found that the 8-day observed NEE can be reproduced reasonably well by this model at the site level (R^2 = 0.73). Additionally, it is interesting to note that, based on the site investigated in this study, Fu et al. [71] used the regression tree model to predict the NEE using two different types of remote sensing data and obtained satisfactory results with R^2 = 0.70 for Landsat data and R^2 = 0.68 for MODIS data during 2005–2006. By comparing with these results, on the whole, our proposed models achieved higher accuracy with the average value of R^2 =0.77 in 2009. In conclusion, our presently proposed models and aforementioned regression tree method have great potential for estimating carbon fluxes, and can be considered as alternative tools to scale up eddy covariance-measured carbon flux data to regional or global scale across different vegetation types.

Furthermore, the seasonal and inter-annual variability in each flux caused by the environmental forcing variables can be satisfactorily captured (Figures 3 and 4). It is noteworthy to point out that appreciable underestimations of the peaks during the growing season and annual total carbon and water fluxes by most of the used models occurred both in the validation and testing periods. In addition, although all the applied models can adequately reproduce each flux for the entire year (2009), substantial differences existed among different seasons, which is consistent with the finding reported by Xiao et al. [70]. In general, our used models provided the worst estimates in winter (December, January and February), while produced satisfactory estimates in both summer and fall seasons. For example, according to our estimates, for GPP, a mean value of R^2 = 0.15, RMSE = 0.13 g·C·m^{-2}·Day^{-1}, and MAE = 0.11 g·C·m^{-2}·Day^{-1} was obtained by the six models in winter, compared with 0.95, 0.51 g·C·m^{-2}·Day^{-1}, and 0.32 g·C·m^{-2}·Day^{-1} in fall. The reasons for these discrepancies may be due to the errors induced by the eddy covariance-measured NEE [72], the partitioning approach of NEE into GPP and *R* [73,74], and the gap-filling methods for missing flux data [69]. Moreover, the annual average values of T_a, T_s and R_h in both 2008 and 2009 were lower compared with the 4-year means during 2004–2007. For example, in 2009, the annual average values of T_a and T_s were lower by 2.84 and 0.52 °C, respectively, while the annual average value of R_n was higher by 0.57 mol·m^{-2}. These non-stationary features in environmental variables may lead to a potential impediment in reproducing the carbon and water fluxes.

According to the modeling performance of various carbon fluxes among all the examined approaches in this work, the ANN model provided the optimal estimates for GPP and NEE, while the ANFIS model achieved the best for *R*, indicating that no single model consistently outperforms others for all the carbon flux estimation. For this reason, it is extremely essential to compare the estimates using a variety of data-driven modeling methods to forecast the carbon fluxes. In contrast, for all the carbon fluxes and ET prediction, the GMDH model consistently produced the worst modeling results. This may be caused in part by its inherent limitations such as selection of input arguments, reduction of complexity, multi-collinearity and over-fitting [75,76].

Taken as a whole, according to the estimates using our proposed techniques, it was found that the performance differences among all the models were comparatively slight for GPP, *R* and ET fluxes, while considerable differences were found for NEE. Specifically, in the prediction of NEE,

the ANN model achieved the optimal estimate with the approximate value of R^2 = 0.84, NSE = 0.78, RMSE = 0.63 g·C·m^{-2}·Day^{-1} and MAE = 0.44 g·C·m^{-2}·Day^{-1}, whereas GMDH model performed the worst with the approximate value of R^2 = 0.67, NSE = 0.60, RMSE = 0.84 g·C·m^{-2}·Day^{-1} and MAE = 0.57 g·C·m^{-2}·Day^{-1}. Moreover, it is noteworthy that evaluating the difference in predicting the carbon fluxes at different time scales (e.g., daily, monthly and annual) should take into account the influences of random errors in half-hourly flux observations. Richardson et al. [72] demonstrated that a total random error in NEE induced by both the eddy covariance measurements and gap filling is roughly 25 g·C·m^{-2}·year^{-1}. In addition to random error, systematic errors can also add to the uncertainty of carbon flux estimates. Therefore, when ranking the performance of our proposed methods for NEE prediction, it seems difficult to find the sources of error in the estimates from these models. Furthermore, in comparison to other fluxes, the diurnal and seasonal variations of NEE in amplitude and phase are strongly affected by the complex interplay between photosynthesis and respiration. On the other hand, through comparing the estimates among the fluxes, the worst results occurred for NEE estimates. A potential explanation for the lower performance in NEE prediction could be attributed to the omission of some important variables such as soil properties and biomass pools for the establishment of our used models. Hence, it is essential to select effective driving variables as model inputs for the further improvement in the predictive ability of data-driven models, especially for NEE, in the future work. In view of the estimates of ET, all the models achieved almost consistent, high modeling accuracy, suggesting that machine learning techniques can be expected as powerful tools to simulate and predict ET. What's more, it should be acknowledged that in recent years, these modeling techniques have been triumphantly applied in numerous branches in hydrology for nonlinear time series analysis, such as reference ET [77,78] and evaporation prediction [33], rainfall [79] and runoff forecasting [80,81].

Our applied models were trained via many attempts in order to determine the optimized internal structure, functions and parameters. And afterwards their corresponding best predictions were derived based on the cross-validation strategy, taking into account the common drawback of over-fitting. Despite all this, the uncertainty issue in estimated results remains a great challenge for further research. To a certain extent, the predictive error caused by such uncertainty could potentially undermine the credibility of the models, and may lead to some problems in the applications of interpolation and extrapolation. Unfortunately, the uncertainty issues existing in the output results are ignored by most studies of machine learning modeling techniques in practical applications. Therefore, to overcome the negative impacts brought by the uncertainty in time series prediction, the uncertainty issues of data-driven models, involving a range of sources such as input data, internal parameters as well as geometry, have been recently addressed by a number of studies [82]. Reasonable uncertainty evaluation is needed to quantify the beneficial information related to confidence bounds and provide more rigorous and credible estimates for policy makers. Specifically, for addressing the uncertainty problem in time series prediction, a suite of algorithms, such as Markov chain Monte Carlo (MCMC) [83] and Bayesian model averaging (BMA) [84], have been integrated into either a single data-driven model or a set of ensemble models. For example, Zhang et al. [85] utilized the MCMC algorithm to train the Bayesian neural networks for estimating different uncertainty sources, and effectively quantified the uncertainty of stream-flow simulation. Chitsazan et al. [86] used a hierarchical BMA approach to estimate the uncertainty of fluoride concentration prediction based on the uncertain sources from the ANN model, and found that the most prediction variance was generated by the uncertain inputs and internal parameters of the ANN model. Consequently, to obtain more beneficial and reliable carbon and water flux prediction in the present work, examining these uncertainty algorithms combined with our applied models is essential for the follow-up study.

5. Conclusions

With the advance of machine learning techniques, many modern data-driven approaches have been developed during the past few decades, which leads to the present predicament of what modeling

technique should be chosen in a practical application, predominantly due to the lack of comprehensive benchmark studies. In order to conquer this predicament, a comparative research plays an essential role in obtaining a full-scale overview of different data-driven methods, specifically aiming at gaining deep insights into their strengths and limitations and drawing some helpful conclusions with regard to their predictability and robustness. In this context, the main basis of the current work was to investigate the suitability of our newly proposed models, including GRNN, GMDH, ELM and ANFIS, in mapping the non-linear relationships that dominate the exchanges of the forest carbon and water fluxes at a flux tower site. In addition, these models were compared and evaluated for the first time with the classical ANN and SVM models in terms of several performance indices (R^2, NSE, RMSE and MAE).

It was found the ANN model provided the best estimates for GPP and NEE, whereas the ANFIS model achieved the best for *R*, indicating that no single model was consistently superior to others for the carbon flux prediction. Therefore, the use of a variety of methods is of particular importance for obtaining adequately accurate estimates for the carbon fluxes. In contrast, for all the carbon fluxes and ET estimation, the GMDH model consistently produced somewhat worse results, and accordingly may be not recommended for the present applications. Moreover, there were considerable differences among all the carbon and water fluxes. When taken as a whole, all the models generated the similar satisfactory predictive accuracy for GPP, *R* and ET fluxes, and did a reasonable job of reproducing the eddy covariance NEE.

The present investigation has manifested the feasibility and validity of several novel data-driven techniques for forecasting the carbon and water fluxes measured by the eddy covariance technique. These models can be used as attractive complements to traditional ANN and SVM models, except for GMDH, which may be due to involvement of large numbers of high-dimensional matrix calculations. In addition, these modern techniques provide many new alternative approaches for data collectors and processers to interpolate the missing data during the long-term eddy covariance measurements, which is of importance for guaranteeing the estimated accuracy in the further studies, when using these related measurements for the parameterization of process-based models and the validation of their estimates for forecasting carbon and water fluxes, and the assessment of published carbon and flux products from remote sensing techniques. More importantly, it is expected that these powerful methods offer novel perspective for up-scaling the carbon and water fluxes from ecosystem to regional or global scale with remote sensing data, as our follow-up investigation, which is particularly essential for the scientific community to intentionally ascertain the carbon and water budgets and further provide helpful information for policy makers responding to present and future climate change.

Acknowledgments: The work was financially supported by the Natural Science Fund of China (No. 41672324, No. 41430317, and No. 41402291), the National Science and Technology Major Projects (No. 2016ZX05044-002), and the Priority Academic Program Development of Jiangsu Higher Education Institutions. The present research used the eddy covariance data acquired and shared by the Fluxnet-Canada Research Network. The authors acknowledge the project investigator who contributed to the eddy covariance flux measurements for providing valuable data for this study. The authors also gratefully acknowledge three anonymous reviewers for their constructive comments.

Author Contributions: Yongguo Yang and Xianming Dou conceived and designed the experiments; Xianming Dou performed the experiments, analyzed the data and wrote the manuscript; Yongguo Yang provided important suggestions for the final manuscript.

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

References

1. Mitchell, S.R.; Emanuel, R.E.; McGlynn, B.L. Land–atmosphere carbon and water flux relationships to vapor pressure deficit, soil moisture, and stream flow. *Agric. For. Meteorol.* **2015**, *208*, 108–117. [CrossRef]
2. Medlyn, B.; De Kauwe, M. Biogeochemistry: Carbon dioxide and water use in forests. *Nature* **2013**, *499*, 287–289. [CrossRef] [PubMed]
3. Baldocchi, D.; Falge, E.; Gu, L.; Olson, R.; Hollinger, D.; Running, S.; Anthoni, P.; Bernhofer, C.; Davis, K.; Evans, R. 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*, 2415–2434. [CrossRef]

4. Jensen, R.; Herbst, M.; Friborg, T. Direct and indirect controls of the interannual variability in atmospheric CO_2 exchange of three contrasting ecosystems in Denmark. *Agric. For. Meteorol.* **2017**, *233*, 12–31. [CrossRef]
5. Thomas, C.K.; Law, B.E.; Irvine, J.; Martin, J.G.; Pettijohn, J.C.; Davis, K.J. Seasonal hydrology explains interannual and seasonal variation in carbon and water exchange in a semiarid mature ponderosa pine forest in central Oregon. *J. Geophys. Res. Biogeosci.* **2009**, *114*, G04006. [CrossRef]
6. Tan, Z.H.; Zhang, Y.P.; Deng, X.B.; Song, Q.H.; Liu, W.J.; Deng, Y.; Tang, J.W.; Liao, Z.Y.; Zhao, J.F.; Song, L. Interannual and seasonal variability of water use efficiency in a tropical rainforest: Results from a 9 year eddy flux time series. *J. Geophys. Res. Atmos.* **2015**, *120*, 464–479. [CrossRef]
7. Amthor, J.S.; Chen, J.M.; Clein, J.S.; Frolking, S.E.; Goulden, M.L.; Grant, R.F.; Kimball, J.S.; King, A.W.; McGuire, A.D.; Nikolov, N.T.; et al. Boreal forest CO_2 exchange and evapotranspiration predicted by nine ecosystem process models: Intermodel comparisons and relationships to field measurements. *J. Geophys. Res. Atmos.* **2001**, *106*, 33623–33648. [CrossRef]
8. Zhao, F.; Zeng, N.; Asrar, G.; Friedlingstein, P.; Ito, A.; Jain, A.; Kalnay, E.; Kato, E.; Koven, C.D.; Poulter, B.; et al. Role of CO_2, climate and land use in regulating the seasonal amplitude increase of carbon fluxes in terrestrial ecosystems: A multimodel analysis. *Biogeosciences* **2016**, *13*, 5121–5137. [CrossRef]
9. Schimel, D.; Stephens, B.B.; Fisher, J.B. Effect of increasing CO_2 on the terrestrial carbon cycle. *Proc. Natl. Acad. Sci. USA* **2015**, *112*, 436–441. [CrossRef] [PubMed]
10. De Vries, W. Forest ecology: Nutrients trigger carbon storage. *Nat. Clim. Change* **2014**, *4*, 425–426. [CrossRef]
11. Balshi, M.S.; McGuire, A.D.; Duffy, P.; Flannigan, M.; Kicklighter, D.W.; Melillo, J. Vulnerability of carbon storage in North American boreal forests to wildfires during the 21st century. *Glob. Change Biol.* **2009**, *15*, 1491–1510. [CrossRef]
12. Hoffman, F.M.; Randerson, J.T.; Arora, V.K.; Bao, Q.; Cadule, P.; Ji, D.; Jones, C.D.; Kawamiya, M.; Khatiwala, S.; Lindsay, K. Causes and implications of persistent atmospheric carbon dioxide biases in earth system models. *J. Geophys. Res. Biogeosci.* **2014**, *119*, 141–162. [CrossRef]
13. Raczka, B.M.; Davis, K.J.; Huntzinger, D.; Neilson, R.P.; Poulter, B.; Richardson, A.D.; Xiao, J.; Baker, I.; Ciais, P.; Keenan, T.F. Evaluation of continental carbon cycle simulations with North American flux tower observations. *Ecol. Monogr.* **2013**, *83*, 531–556. [CrossRef]
14. Schwalm, C.R.; Williams, C.A.; Schaefer, K.; Anderson, R.; Arain, M.A.; Baker, I.; Barr, A.; Black, T.A.; Chen, G.; Chen, J.M.; et al. A model-data intercomparison of CO_2 exchange across North America: Results from the North American Carbon Program site synthesis. *J. Geophys. Res. Biogeosci.* **2010**, *115*, G00H05. [CrossRef]
15. Keenan, T.F.; Carbone, M.S.; Reichstein, M.; Richardson, A.D. The model-data fusion pitfall: Assuming certainty in an uncertain world. *Oecologia* **2011**, *167*, 587–597. [CrossRef] [PubMed]
16. Wang, Y.-P.; Trudinger, C.M.; Enting, I.G. A review of applications of model–data fusion to studies of terrestrial carbon fluxes at different scales. *Agric. For. Meteorol.* **2009**, *149*, 1829–1842. [CrossRef]
17. Michener, W.K.; Jones, M.B. Ecoinformatics: Supporting ecology as a data-intensive science. *Trends Ecol. Evol.* **2012**, *27*, 85–93. [CrossRef] [PubMed]
18. Brown, C.J.; Schoeman, D.S.; Sydeman, W.J.; Brander, K.; Buckley, L.B.; Burrows, M.; Duarte, C.M.; Moore, P.J.; Pandolfi, J.M.; Poloczanska, E.; et al. Quantitative approaches in climate change ecology. *Glob. Change Biol.* **2011**, *17*, 3697–3713. [CrossRef]
19. Olden, J.D.; Lawler, J.J.; Poff, N.L. Machine learning methods without tears: A primer for ecologists. *Q. Rev. Boil.* **2008**, *83*, 171–193. [CrossRef]
20. Cherkassky, V.; Krasnopolsky, V.; Solomatine, D.P.; Valdes, J. Computational intelligence in earth sciences and environmental applications: Issues and challenges. *Neural Netw.* **2006**, *19*, 113–121. [CrossRef] [PubMed]
21. Londhe, S.; Charhate, S. Comparison of data-driven modelling techniques for river flow forecasting. *Hydrol. Sci. J.* **2010**, *55*, 1163–1174. [CrossRef]
22. Jordan, M.; Mitchell, T. Machine learning: Trends, perspectives, and prospects. *Science* **2015**, *349*, 255–260. [CrossRef] [PubMed]
23. Langley, P. The changing science of machine learning. *Mach. Learn.* **2011**, *82*, 275–279. [CrossRef]
24. Qin, Z.; Su, G.-L.; Zhang, J.-E.; 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*, 575–581. [CrossRef]
25. Schmidt, A.; Wrzesinsky, T.; Klemm, O. Gap filling and quality assessment of CO_2 and water vapour fluxes above an urban area with radial basis function neural networks. *Bound. Layer Meteorol.* **2007**, *126*, 389–413. [CrossRef]

26. Tramontana, G.; Jung, M.; Schwalm, C.R.; Ichii, K.; Camps-Valls, G.; Ráduly, B.; Reichstein, M.; Arain, M.A.; Cescatti, A.; Kiely, G.; et al. Predicting carbon dioxide and energy fluxes across global FLUXNET sites with regression algorithms. *Biogeosciences* **2016**, *13*, 4291–4313. [CrossRef]
27. Jung, M.; Reichstein, M.; Margolis, H.A.; Cescatti, A.; Richardson, A.D.; Arain, M.A.; Arneth, A.; Bernhofer, C.; Bonal, D.; Chen, J.; et al. Global patterns of land-atmosphere fluxes of carbon dioxide, latent heat, and sensible heat derived from eddy covariance, satellite, and meteorological observations. *J. Geophys. Res. Biogeosci.* **2011**, *116*, G00J07. [CrossRef]
28. Beer, C.; Reichstein, M.; Tomelleri, E.; Ciais, P.; Jung, M.; Carvalhais, N.; RC6denbeck, C.; Arain, M.A.; Baldocchi, D.; Bonan, G.B. Terrestrial gross carbon dioxide uptake: Global distribution and covariation with climate. *Science* **2010**, *329*, 834–838. [CrossRef] [PubMed]
29. Wang, T.; Brender, P.; Ciais, P.; Piao, S.; Mahecha, M.D.; Chevallier, F.; Reichstein, M.; Ottlé, C.; Maignan, F.; Arain, A.; et al. State-dependent errors in a land surface model across biomes inferred from eddy covariance observations on multiple timescales. *Ecol. Model.* **2012**, *246*, 11–25. [CrossRef]
30. Abramowitz, G.; Pitman, A.; Gupta, H.; Kowalczyk, E.; Wang, Y. Systematic bias in land surface models. *J. Hydrometeorol.* **2007**, *8*, 989–1001. [CrossRef]
31. Kumar, M.; Raghuwanshi, N.S.; Singh, R. Artificial neural networks approach in evapotranspiration modeling: A review. *Irrig. Sci.* **2011**, *29*, 11–25. [CrossRef]
32. Traore, S.; Luo, Y.; Fipps, G. Deployment of artificial neural network for short-term forecasting of evapotranspiration using public weather forecast restricted messages. *Agric. Water Manag.* **2016**, *163*, 363–379. [CrossRef]
33. Kisi, O. Pan evaporation modeling using least square support vector machine, multivariate adaptive regression splines and M5 model tree. *J. Hydrol.* **2015**, *528*, 312–320. [CrossRef]
34. Lima, A.R.; Cannon, A.J.; Hsieh, W.W. Forecasting daily streamflow using online sequential extreme learning machines. *J. Hydrol.* **2016**, *537*, 431–443. [CrossRef]
35. Dariane, A.B.; Azimi, S. Forecasting streamflow by combination of a genetic input selection algorithm and wavelet transforms using anfis models. *Hydrol. Sci. J.* **2016**, *61*, 585–600. [CrossRef]
36. Mekanik, F.; Imteaz, M.A.; Talei, A. Seasonal rainfall forecasting by adaptive network-based fuzzy inference system (ANFIS) using large scale climate signals. *Clim. Dyn.* **2015**, *46*, 3097–3111. [CrossRef]
37. Salcedo-Sanz, S.; Casanova-Mateo, C.; Pastor-Sánchez, A.; Sánchez-Girón, M. Daily global solar radiation prediction based on a hybrid Coral Reefs Optimization—Extreme Learning Machine approach. *Sol. Energy* **2014**, *105*, 91–98. [CrossRef]
38. Osório, G.J.; Matias, J.C.O.; Catalão, J.P.S. Short-term wind power forecasting using adaptive neuro-fuzzy inference system combined with evolutionary particle swarm optimization, wavelet transform and mutual information. *Renew. Energy* **2015**, *75*, 301–307. [CrossRef]
39. Kisi, O.; Shiri, J. Prediction of long-term monthly air temperature using geographical inputs. *Int. J. Climatol.* **2014**, *34*, 179–186. [CrossRef]
40. Partal, T.; Cigizoglu, H.K.; Kahya, E. Daily precipitation predictions using three different wavelet neural network algorithms by meteorological data. *Stoch. Environ. Res. Risk Assess.* **2015**, *29*, 1317–1329. [CrossRef]
41. Zha, T.; Barr, A.G.; van der Kamp, G.; Black, T.A.; McCaughey, J.H.; Flanagan, L.B. Interannual variation of evapotranspiration from forest and grassland ecosystems in Western Canada in relation to drought. *Agric. For. Meteorol.* **2010**, *150*, 1476–1484. [CrossRef]
42. Swanson, R.V.; Flanagan, L.B. Environmental regulation of carbon dioxide exchange at the forest floor in a boreal black spruce ecosystem. *Agric. For. Meteorol.* **2001**, *108*, 165–181. [CrossRef]
43. Krishnan, P.; Black, T.A.; Barr, A.G.; Grant, N.J.; Gaumont-Guay, D.; Nesic, Z. Factors controlling the interannual variability in the carbon balance of a southern boreal black spruce forest. *J. Geophys. Res.* **2008**, *113*, D09109. [CrossRef]
44. Barr, A.G.; Black, T.A.; Hogg, E.H.; Griffis, T.J.; Morgenstern, K.; Kljun, N.; Theede, A.; Nesic, Z. Climatic controls on the carbon and water balances of a boreal aspen forest, 1994–2003. *Glob. Change Biol.* **2007**, *13*, 561–576. [CrossRef]
45. Maier, H.R.; Dandy, G.C. Understanding the behaviour and optimising the performance of back-propagation neural networks: An empirical study. *Environ. Model. Softw.* **1998**, *13*, 179–191. [CrossRef]

46. Maier, H.R.; Jain, A.; Dandy, G.C.; Sudheer, K.P. Methods used for the development of neural networks for the prediction of water resource variables in river systems: Current status and future directions. *Environ. Model. Softw.* **2010**, *25*, 891–909. [CrossRef]
47. Özesmi, S.L.; Tan, C.O.; Özesmi, U. Methodological issues in building, training, and testing artificial neural networks in ecological applications. *Ecol. Model.* **2006**, *195*, 83–93. [CrossRef]
48. Çevik, A.; Kurtoğlu, A.E.; Bilgehan, M.; Gülşan, M.E.; Albegmprli, H.M. Support vector machines in structural engineering: A review. *J. Civ. Eng. Manag.* **2015**, *21*, 261–281. [CrossRef]
49. Raghavendra, N.S.; Deka, P.C. Support vector machine applications in the field of hydrology: A review. *Appl. Soft Comput.* **2014**, *19*, 372–386. [CrossRef]
50. Mountrakis, G.; Im, J.; Ogole, C. Support vector machines in remote sensing: A review. *ISPRS J. Photogramm. Remote Sens.* **2011**, *66*, 247–259. [CrossRef]
51. Cortes, C.; Vapnik, V. Support-vector networks. *Mach. Learn.* **1995**, *20*, 273–297. [CrossRef]
52. Hamidi, O.; Poorolajal, J.; Sadeghifar, M.; Abbasi, H.; Maryanaji, Z.; Faridi, H.R.; Tapak, L. A comparative study of support vector machines and artificial neural networks for predicting precipitation in Iran. *Theor. Appl. Climatol.* **2015**, *119*, 723–731. [CrossRef]
53. Chang, C.-C.; Lin, C.-J. Libsvm: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.* **2011**, *2*, 27. [CrossRef]
54. Jang, J.-S. Anfis: Adaptive-network-based fuzzy inference system. *IEEE Trans. Syst. Man Cybern.* **1993**, *23*, 665–685. [CrossRef]
55. Takagi, T.; Sugeno, M. Fuzzy identification of systems and its applications to modeling and control. *IEEE Trans. Syst. Man Cybern.* **1985**, *SMC-15*, 116–132. [CrossRef]
56. Mamdani, E.H.; Assilian, S. An experiment in linguistic synthesis with a fuzzy logic controller. *Int. J. Man Mach. Stud.* **1975**, *7*, 1–13. [CrossRef]
57. Specht, D.F. A general regression neural network. *IEEE Trans. Neural Netw.* **1991**, *2*, 568–576. [CrossRef] [PubMed]
58. Zhou, Q.; Jiang, H.; Wang, J.; Zhou, J. A hybrid model for PM 2.5 forecasting based on ensemble empirical mode decomposition and a general regression neural network. *Sci. Total Environ.* **2014**, *496*, 264–274. [CrossRef] [PubMed]
59. Ladlani, I.; Houichi, L.; Djemili, L.; Heddam, S.; Belouz, K. Modeling daily reference evapotranspiration (ET0) in the north of Algeria using generalized regression neural networks (GRNN) and radial basis function neural networks (RBFNN): A comparative study. *Meteorol. Atmos. Phys.* **2012**, *118*, 163–178. [CrossRef]
60. Huang, G.-B.; Zhu, Q.-Y.; Siew, C.-K. Extreme learning machine: Theory and applications. *Neurocomputing* **2006**, *70*, 489–501. [CrossRef]
61. Ding, S.; Xu, X.; Nie, R. Extreme learning machine and its applications. *Neural Comput. Appl.* **2014**, *25*, 549–556. [CrossRef]
62. Yaseen, Z.M.; Jaafar, O.; Deo, R.C.; Kisi, O.; Adamowski, J.; Quilty, J.; El-Shafie, A. Stream-flow forecasting using extreme learning machines: A case study in a semi-arid region in Iraq. *J. Hydrol.* **2016**, *542*, 603–614. [CrossRef]
63. Deo, R.C.; Şahin, M. Application of the extreme learning machine algorithm for the prediction of monthly Effective Drought Index in eastern Australia. *Atmos. Res.* **2015**, *153*, 512–525. [CrossRef]
64. Ivakhnenko, A.G. Polynomial theory of complex systems. *IEEE Trans. Syst. Man Cybern.* **1971**, *1*, 364–378. [CrossRef]
65. Müller, J.A.; Ivachnenko, A.G.; Lemke, F. GMDH algorithms for complex systems modelling. *Math. Comput. Model. Dyn. Syst.* **1998**, *4*, 275–316. [CrossRef]
66. Dou, X.; Chen, B.; Black, T.; Jassal, R.; Che, M. Impact of nitrogen fertilization on forest carbon sequestration and water loss in a chronosequence of three douglas-fir stands in the Pacific Northwest. *Forests* **2015**, *6*, 1897–1921. [CrossRef]
67. Moffat, A.M.; Beckstein, C.; Churkina, G.; Mund, M.; Heimann, M. Characterization of ecosystem responses to climatic controls using artificial neural networks. *Glob. Change Biol.* **2010**, *16*, 2737–2749. [CrossRef]
68. Yang, F.; Ichii, K.; White, M.A.; Hashimoto, H.; Michaelis, A.R.; Votava, P.; Zhu, A.X.; Huete, A.; Running, S.W.; Nemani, R.R. Developing a continental-scale measure of gross primary production by combining MODIS and AmeriFlux data through support vector machine approach. *Remote Sens. Environ.* **2007**, *110*, 109–122. [CrossRef]

69. Moffat, A.M.; Papale, D.; Reichstein, M.; Hollinger, D.Y.; Richardson, A.D.; Barr, A.G.; Beckstein, C.; Braswell, B.H.; Churkina, G.; Desai, A.R.; et al. Comprehensive comparison of gap-filling techniques for eddy covariance net carbon fluxes. *Agric. For. Meteorol.* **2007**, *147*, 209–232. [CrossRef]
70. Xiao, J.; Zhuang, Q.; Baldocchi, D.D.; Law, B.E.; Richardson, A.D.; Chen, J.; Oren, R.; Starr, G.; Noormets, A.; Ma, S.; et al. Estimation of net ecosystem carbon exchange for the conterminous United States by combining MODIS and AmeriFlux data. *Agric. For. Meteorol.* **2008**, *148*, 1827–1847. [CrossRef]
71. Fu, D.; Chen, B.; Zhang, H.; Wang, J.; Black, T.A.; Amiro, B.D.; Bohrer, G.; Bolstad, P.; Coulter, R.; Rahman, A.F.; et al. Estimating landscape net ecosystem exchange at high spatial–temporal resolution based on Landsat data, an improved upscaling model framework, and eddy covariance flux measurements. *Remote Sens. Environ.* **2014**, *141*, 90–104. [CrossRef]
72. Richardson, A.D.; Hollinger, D.Y.; Burba, G.G.; Davis, K.J.; Flanagan, L.B.; Katul, G.G.; William Munger, J.; Ricciuto, D.M.; Stoy, P.C.; Suyker, A.E.; et al. A multi-site analysis of random error in tower-based measurements of carbon and energy fluxes. *Agric. For. Meteorol.* **2006**, *136*, 1–18. [CrossRef]
73. Stoy, P.C.; Katul, G.G.; Siqueira, M.B.S.; Juang, J.-Y.; Novick, K.A.; Uebelherr, J.M.; Oren, R. An evaluation of models for partitioning eddy covariance-measured net ecosystem exchange into photosynthesis and respiration. *Agric. For. Meteorol.* **2006**, *141*, 2–18. [CrossRef]
74. Lasslop, G.; Reichstein, M.; Papale, D.; Richardson, A.D.; Arneth, A.; Barr, A.; Stoy, P.; Wohlfahrt, G. Separation of net ecosystem exchange into assimilation and respiration using a light response curve approach: Critical issues and global evaluation. *Glob. Change Biol.* **2010**, *16*, 187–208. [CrossRef]
75. Tamura, H.; Kondo, T. Heuristics free group method of data handling algorithm of generating optimal partial polynomials with application to air pollution prediction. *Int. J. Syst. Sci.* **1980**, *11*, 1095–1111. [CrossRef]
76. Onwubolu, G.C. Design of hybrid differential evolution and group method of data handling networks for modeling and prediction. *Inf. Sci.* **2008**, *178*, 3616–3634. [CrossRef]
77. Abdullah, S.S.; Malek, M.A.; Abdullah, N.S.; Kisi, O.; Yap, K.S. Extreme learning machines: A new approach for prediction of reference evapotranspiration. *J. Hydrol.* **2015**, *527*, 184–195. [CrossRef]
78. Laaboudi, A.; Mouhouche, B.; Draoui, B. Neural network approach to reference evapotranspiration modeling from limited climatic data in arid regions. *Int. J. Biometeorol.* **2012**, *56*, 831–841. [CrossRef] [PubMed]
79. Abbot, J.; Marohasy, J. Input selection and optimisation for monthly rainfall forecasting in Queensland, Australia, using artificial neural networks. *Atmos. Res.* **2014**, *138*, 166–178. [CrossRef]
80. Meng, X.; Yin, M.; Ning, L.; Liu, D.; Xue, X. A threshold artificial neural network model for improving runoff prediction in a karst watershed. *Environ. Earth Sci.* **2015**, *74*, 5039–5048. [CrossRef]
81. Kalteh, A.M. Monthly river flow forecasting using artificial neural network and support vector regression models coupled with wavelet transform. *Comput. Geosci.* **2013**, *54*, 1–8. [CrossRef]
82. Ghahramani, Z. Probabilistic machine learning and artificial intelligence. *Nature* **2015**, *521*, 452–459. [CrossRef] [PubMed]
83. Geyer, C.J. Practical markov chain monte carlo. *Stat. Sci.* **1992**, *7*, 473–483. [CrossRef]
84. Hoeting, J.A.; Madigan, D.; Raftery, A.E.; Volinsky, C.T. Bayesian model averaging: A tutorial. *Stat. Sci.* **1999**, *14*, 382–401.
85. Zhang, X.; Liang, F.; Srinivasan, R.; Van Liew, M. Estimating uncertainty of streamflow simulation using Bayesian neural networks. *Water Resour. Res.* **2009**, *45*, W02403. [CrossRef]
86. Chitsazan, N.; Nadiri, A.A.; Tsai, F.T.C. Prediction and structural uncertainty analyses of artificial neural networks using hierarchical Bayesian model averaging. *J. Hydrol.* **2015**, *528*, 52–62. [CrossRef]