water

Article

Estimating Carbon Dioxide (CO_2) Emissions from Reservoirs Using Artificial Neural Networks

Zhonghan Chen, Xiaoqian Ye and Ping Huang *

Department of Environmental Science, School of Environmental Science and Engineering, Sun Yat-sen University, Guangzhou 510275, China; chenzhh43@mail2.sysu.edu.cn (Z.C.); yexqian@mail2.sysu.edu.cn (X.Y.)

* Correspondence: eeshping@mail.sysu.edu.cn or pinghuang43@foxmail.com; Tel.: +86-132-4974-8826

Received: 3 November 2017; Accepted: 28 December 2017; Published: 1 January 2018

Abstract: Freshwater reservoirs are considered as the source of atmospheric greenhouse gas (GHG), but more than 96% of global reservoirs have never been monitored. Compared to the difficulty and high cost of field measurements, statistical models are a better choice to simulate the carbon emissions from reservoirs. In this study, two types of Artificial Neural Networks (ANNs), Back Propagation Neural Network (BPNN) and Generalized Regression Neural Network (GRNN), were used to predict carbon dioxide (CO_2) flux emissions from reservoirs based on the published data. Input variables, which were latitude, age, the potential net primary productivity, and mean depth, were selected by Spearman correlation analysis, and then the rationality of these inputs was proved by sensitivity analysis. Besides this, a Multiple Non-Linear Regression (MNLR) and a Multiple Linear Regression (MLR) were used for comparison with ANNs. The performance of models was assessed by statistical metrics both in training and testing phases. The results indicated that ANNs gave more accurate results than regression models and GRNN provided the best performance. With the help of this GRNN, the total CO_2 emitted by global reservoirs was estimated and possible CO_2 flux emissions from a planned reservoir was assessed, which illustrated the potential application of GRNN.

Keywords: CO_2; reservoirs; general regression neural network; back propagation neural network

1. Introduction

Since the problem of greenhouse gases (GHGs) emissions from hydroelectric reservoirs was first addressed in the publication in 1993 [1], it has been the focus of research around the world [2], especially in Canada [3,4], Brazil [5], and the United States [6,7]. Over the past two decades, a growing amount of work has documented reservoirs' roles as GHG sources [8,9], after extensive research was carried out in various reservoirs. However, the magnitude of global flux of GHG from reservoirs is still highly debatable [10]. According to the current estimations of global carbon dioxide (CO_2) emissions, the hydroelectric reservoirs were responsible for emitting 48 Tg C yr^{-1} as CO_2 [11], while Demmer et al. [12] estimated that GHG emissions accounted for 36.8 Tg C yr^{-1} as CO_2 ignoring the types of reservoirs. These estimates corresponded roughly to 2% of global carbon emissions from inland waters that reported a flux of 2100 Tg C yr^{-1} as CO_2 [13]. Although there was a minor difference between the estimated global CO_2 flux from hydroelectric systems and all reservoir systems, any significant difference between the areal emissions of CO_2 from hydroelectric and non-hydroelectric reservoirs was not detected by statistical analysis [12]. Depending on reservoir type, GHG emissions from reservoirs are related to various factors, which is crucial for understanding the mechanism and the control over GHG emissions. In a single reservoir system, both depth and temperature might be the important predictors of carbon emissions, and also reflect the spatial and seasonal variability [5,14]. GHG emissions tend to decrease with the increase of reservoirs' latitude and age [11]. Considering the internal source of carbon emissions, the initial organic carbon in the flooded area is another key factor, especially in the young reservoirs [15,16]. Besides, the GHG emissions are also related to dam operation regime [5] and water quality, such as pH [17] and Chlorophyll-*a* [18,19].

Despite significance and uncertainty, there is still a lack of measured CO_2 emissions from reservoirs in many regions, which leaves a critical gap in the global CO_2 budgets. To resolve this problem, statistical models are the appropriate methods to extrapolate the flux of reservoirs without measured data and then derive regional or global estimations relying on a limited number of measurements [20]. One of the most common models is the statistical regression model, which can demonstrate the relationship between CO_2 emissions and one [9] or several [11,16] factors by the regression equations. Other models that can identify more complex non-linear responses, such as Random Forests [6] and Monte Carlo simulation [21], were also used to elucidate the relationship between CO_2 emissions and the factors of environment or dams, and also to predict CO_2 emissions from other under-sampled reservoirs in the nearby geographic region. Unlike these models with the inputs of environmental factors, a mathematical model with the theory of Self-Organized Criticality (SOC) was employed to extrapolate values from one reservoir to another directly without any other features [22]. These pioneering models showed a low degree of precision and regional limitations; therefore, developing and optimizing models of reservoir CO_2 emissions is still one of the future research directions [6,12].

Artificial neural networks (ANNs) were frequently being used for the simulation of both water quality [23,24] and GHG emissions caused by energy consumption [25,26], and showed great potential for prediction. However, few research has been conducted to simulate GHG emissions from reservoirs using ANNs. Actually, ANNs have several advantages that are suitable for this study. ANNs have robust learning and generalization ability, after simulating the learning and decision-making process of human beings. Therefore, ANNs can describe the linear or non-linear relationships precisely even with limited input variables [27,28] and identify complex patterns in dataset without adequate understanding of the interaction among variables [24]. Besides this, because the learning mechanism in ANNs is non-parametric, the structure and distribution of data are not limitations [29]. The precision of fitting by ANNs largely depends on the quantity and quality of data [30]. CO_2 fluxes have been measured in at least 229 reservoir systems in the world until 2016 [12], which supplies the sufficient amount of data for ANNs. Besides, many commonly employed techniques for measurement focus on quantifying the diffusive flux of gases across the air–water interface, which is suitable for CO_2 because of its solubility [12]. There are some other key factors that affect model performance, such as the architecture selection and parameter settings [28]. However, it is difficult to reach any conclusion of which model architecture is absolutely suitable to a particular circumstance. Therefore, ad hoc approaches, such as a trial-and-error approach, might be acceptable to determine the parameters, following the principle that the optimal network structure generally keeps a balance between generalization ability and network complexity [31].

In this study, we make the first attempt to simulate the CO_2 flux emission from reservoirs by ANNs, back propagation neural network (BPNN) and general regression neural network (GRNN), based on the published data from various types of reservoirs with a global distribution. The input variables of models is selected through both the correlation analysis and domain knowledge. The rationality of selected sets of inputs is tested by sensitivity analysis. The model parameters selection are described in detail and the model performance is evaluated using statistical indices. Since the models have the ability to predict CO_2 fluxes from a reservoir without measurements, the global fluxes of CO_2 emissions from reservoirs can be estimated. Besides this, the possible magnitude of CO_2 emissions from a planned reservoir can also be assessed by models based on some reservoirs' features, which gives guidance for dam's construction.

2. Materials and Methods

2.1. Data Collection

In this study, most CO_2 emission fluxes from reservoir surface were based on data collection by Barros et al. [11] in 2011 and Deemer et al. [12] in 2016. CO_2 emission monitoring data from reservoirs in some latest literature were also assembled. Some essential parameters about reservoirs and monitoring were collected from the relevant literature and part of missing values were completed from Global Reservoir and Dam (GRanD) database [19]. Another important parameter, the primary productivity of potential vegetation (NPP0) in the reservoir's location, was extracted from the map of the human appropriation of net primary production (HANPP) [32]. In cases where more than one study measured CO_2 fluxes from the system in the same year, the arithmetic mean of these parameters was used for statistic, and CO_2 fluxes from same reservoir measured in different years were kept at the original values. Therefore, 277 data sets were collected based on the studies in 235 reservoirs, and the information of data set can be found in supplementary Table S1.

In total, we assembled 10 parameters of 235 reservoirs with a global distribution, including latitude (Lat), age, chlorophyll-*a* (Chl-*a*), water temperature (WT), mean depth (MD), residence time (RT), dissolved organic carbon (DOC), total phosphorus (TP), the potential net primary productivity of the area (NPP0), and CO_2 flux. CO_2 flux was used as the output of models in this study. The statistical parameters including minimum value, maximum value, mean value, median, standard deviation, and variation coefficient are given in Table 1.

Table 1. The statistical parameters for data sets.

Parameters	Unit	Min	Max	Mean	Median	SD	VC
Lat	°	−42.93	68.00	31.96	38.17	26.36	0.83
Age	yrs	1.00	95.00	39.09	36.00	24.55	0.63
Chl-*a*	$\mu g\ L^{-1}$	0.20	137.50	12.03	4.13	24.78	1.96
WT	°C	6.30	35.00	17.88	17.40	5.52	0.30
MD	m	0.30	400.00	26.58	15.00	40.26	1.52
RT	days	1.25	13,140.00	665.75	180.00	1689.54	2.54
DOC	$mg\ L^{-1}$	1.25	30.00	4.79	3.82	4.01	0.84
TP	$\mu g\ L^{-1}$	1.40	500.00	62.61	29.00	96.89	1.55
NPP0	$mg\ C\ m^{-2}\ d^{-1}$	151.90	3200.68	1529.21	1574.50	604.70	0.40
CO_2 flux	$mg\ C\ m^{-2}\ d^{-1}$	−356.00	3800.00	400.90	254.75	569.89	1.42

Note: SD, the standard deviation; VC, the coefficient of variation; Lat, latitude; Chl-*a*, chlorophyll-*a*; WT, water temperature; MD, mean depth; RT, residence time; DOC, dissolved organic carbon; TP, total phosphorus; NPP0, potential net primary productivity.

2.2. Input Variables and Data Processing

The selection of a set of appropriate input variables is the precondition of developing a satisfactory ANN for prediction [31]. There are two basic principles for selection: one is the confirmation of input–output relationship, and the other is the independence among input variables [26]. This characteristic of independence is very important since correlated data can provide redundant information, which increases the likelihood of overfitting and the difficulty to find optimal weights [31]. There are two general categories of techniques, model-free and model-based approaches, to examine the relationship between alternative inputs and outputs, such as the correlation analysis and the stepwise analysis [31]. In this study, the correlation analysis, sensitivity analysis, the availability of data, and domain knowledge are combined to select an optimal set of input variables for models.

Since the variables have different units, the variables should be scaled to a uniform range before passing them through the ANNs to avoid convergence problems and extremely small weighting factors [33]. There are no fixed rules for the standardization in literature. In this study, normalization was done by the following equation:

$$Y = (y_{max} - y_{min})\left(\frac{x - x_{min}}{x_{max} - x_{min}}\right) + y_{min} \tag{1}$$

where, Y denotes the data value after normalization, x_{max} and x_{min} are the maximum and the minimum values of each variable, and y_{max} and y_{min} denote the boundary values of the specific range, which are 1 and -1 respectively in this study.

The complete CO_2 emissions data set comprises 235 reservoirs in 21 countries. In this study, 70% of samples with the whole selected input variables were randomly chosen to build and train models, while the remaining data was used to test the models.

2.3. *Artificial Neural Networks (ANNs)*

Artificial neural networks (ANNs) are the abstract computational models that follow the behavior of the human brain. ANNs have been used widely for prediction and forecasting in the environmental area, because they are believed to approximate any finite non-linear function with high accuracy [34]. Traditionally, ANNs are divided into feed-forward and recurrent networks. Feed-forward architectures, which have many types, such as Multilayer Perceptions (MLP) and Radial Basis Function (RBFNN), are the most popular architectures used in researches [31]. In feed-forward ANNs, the input signal propagates through network in a forward direction, from the input layer to the hidden layer and then to the output neurons [33]. Among many types of feed-forward networks, BPNN and GRNN were chosen for this study.

2.3.1. Back Propagation Neural Networks (BPNNs)

Back Propagation Neural Networks (BPNNs) are typical multi-layer perceptron neural networks (MLPs) with error back-propagation (BP) algorithm for network learning [34]. BPNNs are organized as hierarchical networks with several layers including an input layer, hidden layer(s), and an output layer. Generally, it is believed that BPNNs with one hidden layer are able to describe any finite non-linear relationship with acceptable performance [35]. Each layer has several neurons that transmit input values and process to the next layer by a set of weights (Figure 1). In each neuron, the sum of the products of input variables and their weights is transformed to an output value by an activation function. In this study, the classical *tansig* function was selected as the activation function from the input layer to the hidden layer (Equation (2)). The *purelin* linear function was applied from the hidden layer to the output layer (Equation (3)). There is a wide variety of algorithms available for training a network and adjusting its weights, and Levenberg–Marquardt algorithm (LMA) was used in this study.

$$X_j = tansig\left(\sum_{i=1}^{m} x_i w'_{ji} + b'_j\right) \tag{2}$$

$$Y = \sum_{j=1}^{n}\left(X_j w''_j + b''\right) \tag{3}$$

where, m and n are the numbers of neurons in the input and hidden layers, respectively. x_i is the values of the input variables; X_j is the result obtained by activation function (*tansig*) from the input neurons; w'_{ji} is the connection weight between the ith input neuron and jth hidden neuron, and w''_j is the connection weight between the jth hidden neuron and the output neuron; b'_j and b'' are the bias for the jth hidden neuron and the output neuron, respectively.

The reason for selecting BPNN is that it is the most commonly used for simulation among ANNs. Besides, the BP algorithm can efficiently minimize network error by dynamically searching for the optimal weights. The optimal number of hidden nodes for BPNN was selected by trying different integers in a reasonable range separately, which created the balance between complexity and accuracy.

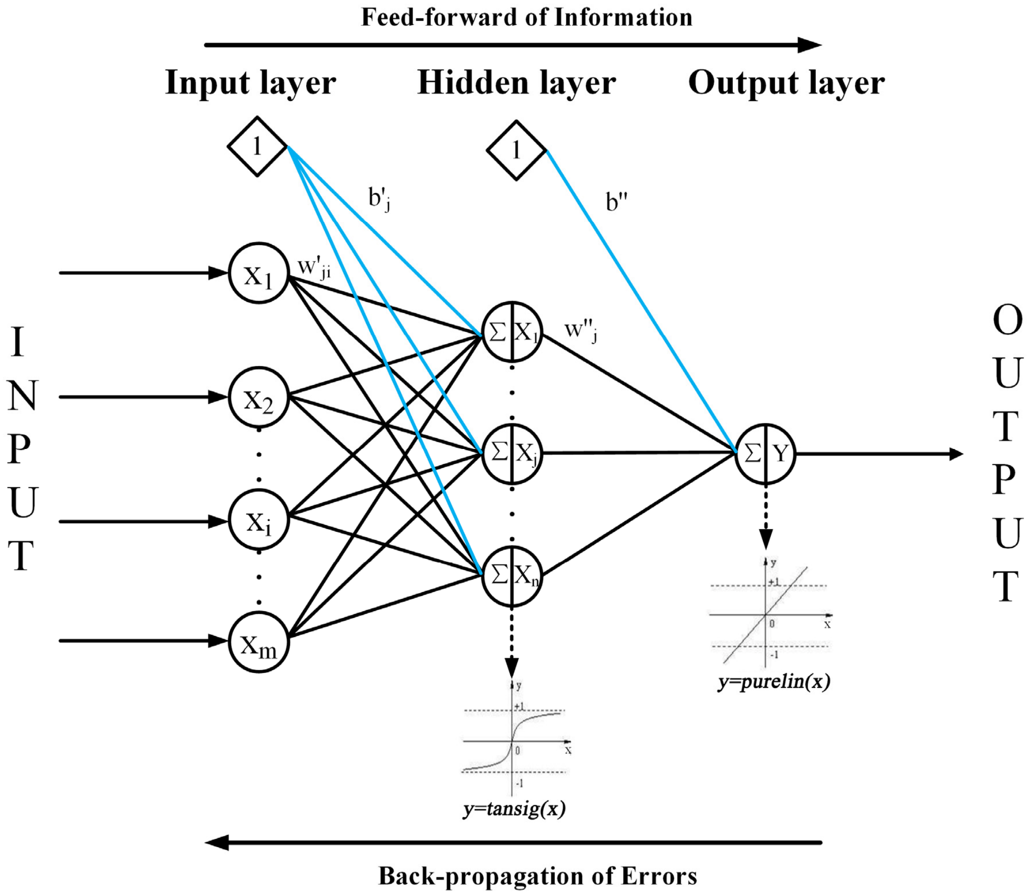

Figure 1. Architecture of Back-Propagation Neural Network (BPNN).

2.3.2. General Regression Neural Networks (GRNNs)

Unlike BPNNs, General Regression Neural Networks (GRNNs) do not rely on iterative procedures for their training but rely on a standard statistical technique called kernel regression [36]. GRNNs consist of four consequent layers, namely input, pattern, summation, and output layers (Figure 2).

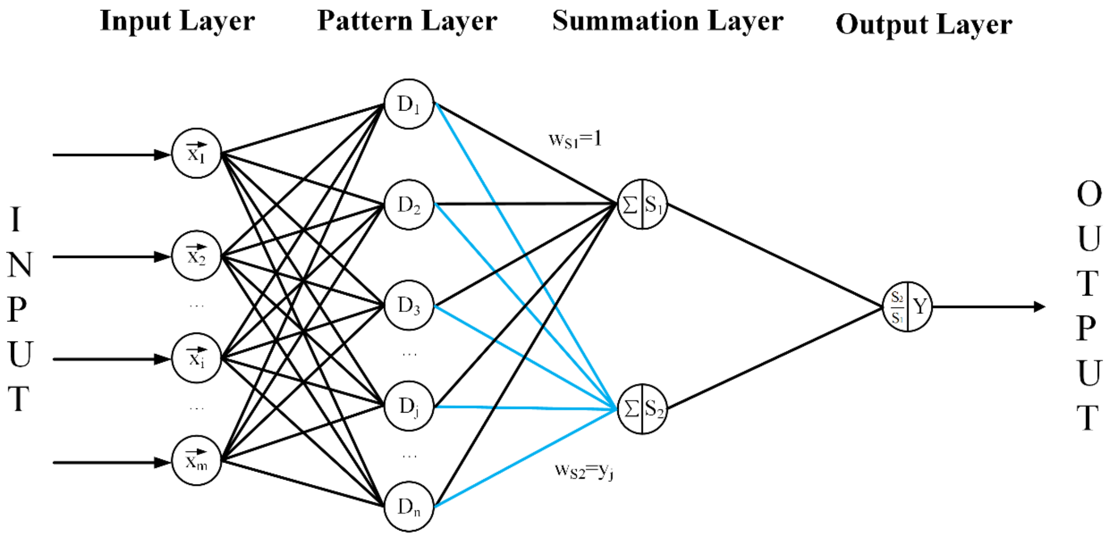

Figure 2. Architecture of General Regression Neural Network (GRNN).

In the first (input) layer, the input units pass input variables to pattern layer through input weights. In the second (pattern) layer, each neuron presents a training pattern, and the similarity between input patterns is calculated using a distance function (Equation (4)). The third (summation) layer consists two different types of summation, including single division and summation units. Each neuron in the pattern layer is connected to both S-summation and D-summation neurons in the summation layer. The S-summation neuron calculates the sum of the weighted responses in pattern layer, whereas the D-summation neuron computes the unweighted output neuron in the pattern layer. The final (output) layer just divides each S-summation neuron on D-summation neuron and represents the network prediction (Equations (5)) [37,38].

$$D_j\left(x, x_j\right) = \sum_{i=1}^{m}\left(\frac{x_j - x_{ji}}{\sigma}\right)^2 \tag{4}$$

$$\mathrm{Y} = \frac{\sum_{j=1}^{n} y_j exp\left[-D_j\left(x, x_j\right)\right]}{\sum_{j=1}^{n} exp\left[-D_j\left(x, x_j\right)\right]} \tag{5}$$

where, m and n are the number of elements of an input vector and the number of training patterns, respectively. D is the Gaussian function. The term of $\left(x_j - x_{ji}\right)$ denotes the difference between the ith training data x_{ji} and the point of estimation x_j. σ is the spread (smoothness) parameter whose optimal value can be experimentally evaluated. y_i represents the weight relationship between ith neuron in the pattern layer and the S-summation neuron in the summation layer.

The GRNN method was also selected because of its fast learning and convergence to the optimal regression surface [39]. Besides, it has another advantage that the network architecture is fixed and only one single parameter named spread needs to be optimized [39]. Different values of spread were also tried separately for each GRNN model in this study.

2.4. Statistical Regression Models

The performance of ANNs was compared with those of the multiple linear regression model (MLR) and multiple non-linear regression model (MNLR). These models had the same inputs as ANNs and were developed and tested using the same data. The MLR is shown in Equation (6). To identify a suitable MNLR, the various numerical transformations were tried, such as reciprocal, logarithm, and square root (Equation (7)).

$$\mathrm{Y} = \alpha_0 + \alpha_1 x_1 + \alpha x_2 + \ldots + \alpha_m x_m \tag{6}$$

$$\mathrm{Y} = \beta_0 + \beta_1 f_1(x_1) + \beta_1 f_2(x_2) + \ldots + \beta_m f_m(x_m) \tag{7}$$

where, x_m is the input, α_m is the coefficients of first degree inputs, f_m represents the transfer function for the input, and β_m is the coefficients of transferred input.

2.5. Performance Metrics

The results of created models were analyzed using multiple performance metrics. The root mean squared error (RMSE) and the mean absolute error (MAE) measure residual errors, which show the difference between the modeled and observed values. The determination coefficient (R^2) is identical to the square of the correlation coefficient (R) in some cases, which evaluates the degree of variability that can be explained by the model. Nash–Sutcliffe efficiency (NSE) [40] is a popular likelihood function to define the goodness of fit between monitoring data and model outputs. Smaller RMSE and MAE values and larger R^2 and NSE values indicate better model performance.

3. Results

3.1. Alternative Input Variable Selection

The original input dataset was selected through the results of Spearman's correlation analysis between the CO_2 flux and other nine variables of these reservoirs. Spearman correlation coefficient was selected because the data of CO_2 flux was not normally distributed (Kolmogorov-Smirnov test, $p < 0.001$), even after several different attempts at transformation. Some reservoirs were measured many times in different years, which means these reservoirs have more than one set of data with the same parameters except age and CO_2 flux. Therefore, Spearman correlation analysis between CO_2 flux and age was carried out by the original values, while Spearman correlation analysis between CO_2 flux and other parameters (Lat, Chl-*a*, WT, MD, RT, DOC, TP, and NPP0) was made based on the arithmetic mean.

Based on the results of correlation analysis (Table 2), the variables that have the high absolute value of correlation coefficient and low value of significance coefficient were age, mean depth, and NPP0. Moreover, latitude was reported as a key factor in previous studies [8,11] and the available sample size is sufficient. Therefore, latitude is also considered as an input and tested the rationality by sensitive analysis. Since they are obviously independent, it is unnecessary to analyze the correlation among these parameters. Consequently, the four features were chosen as alternative input variables of the models for CO_2 prediction. Only the reservoir data where the four parameters' records are effective and available were used, otherwise, this set of data would be removed. After deleting the invalid data, a dataset containing 251 sets of data was selected for simulation.

Table 2. Spearman correlation coefficient between CO2 flux and parameters used in present study.

Variables	*n*	Correlation	Sig.	Variables	*n*	Correlation	Sig.
Lat	236	−0.025	0.69	RT	98	0.055	0.59
Age	266	**−0.307**	0.00	DOC	51	0.129	0.36
Chl-*a*	69	−0.115	0.35	TP	47	0.005	0.98
WT	158	−0.118	0.13	NPP0	234	**0.153**	0.02
MD	217	**−0.151**	0.02				

Note: (1) *n*, the number of samples; Correlation, Spearman Correlation; Sig., Significance coefficient. (2) Latitude was calculated by the absolute value. (3) Bold font: correlation is significant at the 0.05 level (two-tailed).

3.2. Model Parameters Selection

A three-layer BPNN model was used in this study. To determine the topology of the network, diverse numbers of hidden nodes that range from 1 to 15 were tried for the BPNN, and their performance was compared by RMSE. The BPNN with nine hidden neurons showed the best fit performance (RMSE = 399.01 mg C m^{-2} d^{-1}). As a result, the topology of BPNN used in this study has four input neurons, nine hidden neurons, and one output neuron.

For the GRNN model, the spread constant was attempted by comparing the RMSE values obtained for each model with different spread constant varies from 0.1 to 1. The RMSE is lowest at 0.1 spread constant (RMSE = 279.69 mg C m^{-2} d^{-1}).

3.3. Model Performances

Among 251 sets of data, 175 (about 70%) sets are selected randomly for training, and the rest (76 sets) are testing data. The statistical parameters of CO_2 flux for training and testing are given in Table 3.

Table 3. The statistical parameters for CO_2 flux in training and testing phase.

Statistical Parameters	Unit	Training Set	Testing Set
n		76	175
Min	mg C m^{-2} d^{-1}	−325.90	−356.00
Max	mg C m^{-2} d^{-1}	3776.00	3800.00
Mean	mg C m^{-2} d^{-1}	390.82	486.40
Median	mg C m^{-2} d^{-1}	243.29	312.03
SD	mg C m^{-2} d^{-1}	549.75	664.02

The statistical performance measures for MLR, MNLR, BPNN, and GRNN are presented in Table 4 for both training and testing data sets. The GRNN model achieved the best performance in both training and testing phase according to mean performance statistics.

Table 4. The performances of each model during training and testing phases.

Model	Training Data Set				Testing Data Set			
	RMSE	MAE	R^2	NSE	RMSE	MAE	R^2	NSE
MLR	476.42	313.22	0.25	0.25	625.36	429.96	0.12	0.11
MNLR	417.26	282.00	0.43	0.42	529.53	391.46	0.40	0.36
BPNN	396.59	268.53	0.52	0.48	505.43	395.33	0.47	0.42
GRNN	272.50	147.62	0.76	0.75	418.48	295.34	**0.61**	0.60

Note: (1) The unit of RMSE is mg C m^{-1} d^{-1}; the unit of MAE is mg C m^{-1} d^{-1}. RMSE, root mean squared error; MAE, mean absolute error; NSE, Nash–Sutcliffe efficiency; MLR, multiple linear regression; MNLR, multiple non-linear regression; BPNN, Back Propagation Neural Network; GRNN, Generalized Regression Neural Network.

The observed and predicted CO_2 flux values by MLR, MNLR, BPNN, and GRNN models in training and testing stages are plotted in Figure 3. Comparisons among three models above indicate that GRNN generally gives better accuracy than the BPNN, MNLR, and MLR models. This can also be clearly observed from the fit line equations.

The equations of MLR and MNLR are shown in Equations (8) and (9) respectively.

$$\text{CO2flux} = 678.70 - 6.34 \cdot |\text{Lat}| - 5.71 \cdot \text{Age} + 0.42 \cdot \text{NPP0} - 2.21 \cdot \text{MD}. \tag{8}$$

$$\text{CO2flux} = 471.21 - 3763 \cdot \left(\frac{1}{|\text{Lat}|}\right) - 168.97 \cdot \ln(\text{Age}) + 61.23 \cdot \ln(\text{NPP0}) - 2.04 \cdot \text{MD}. \tag{9}$$

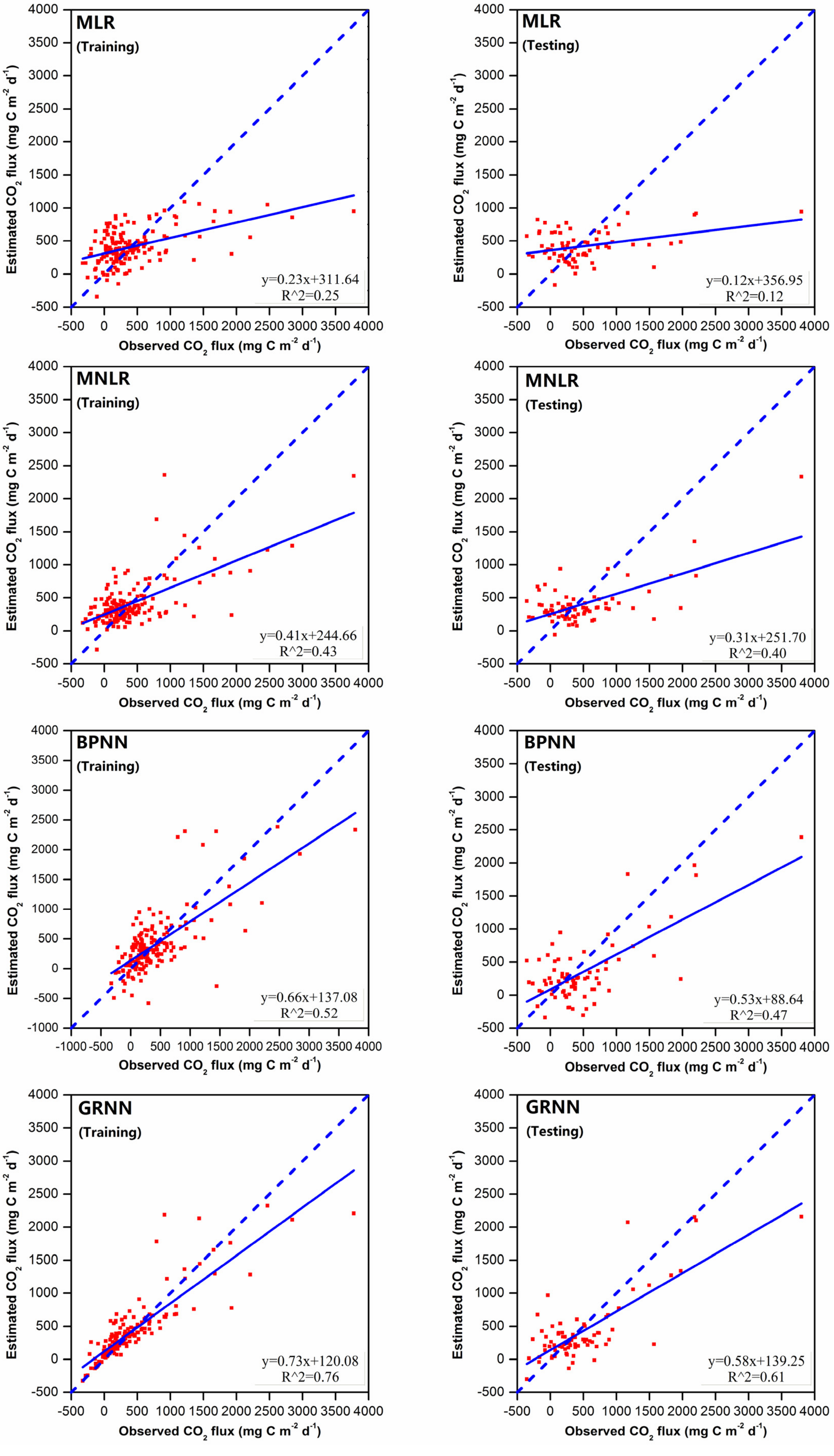

Figure 3. Comparison between CO_2 flux and predicted using MLR, MNLR, BPNN, and GRNN.

3.4. Sensitivity Analysis

Sensitivity analysis was applied to identify whether the selected set of inputs is suitable and which of them is the most important one in simulating CO_2 flux. We built new models based on different combination of input variables and compared their RMSE and R^2 values. These four sets of inputs were made up by omitting a parameter on each run. It was obvious that the omission of the most important parameter could have the highest influence on model performance, which was reflected in higher RMSE values and lower R^2 values [41]. Results of sensitivity analysis are presented in Table 5. The results demonstrated that the performance of the models with four input variables is better than other models', and the latitude parameter is an essential input for GRNN, BPNN, and MNLR.

Table 5. Results of sensitivity analysis for GRNN, BPNN, MNLR, and MLR in the testing phase.

Model	RMSE (mg C m^{-2} d^{-1})				R^2			
	GRNN	**BPNN**	**MNLR**	**MLR**	**GRNN**	**BPNN**	**MNLR**	**MLR**
All	418.48	505.43	529.53	625.36	0.61	0.47	0.40	0.12
Skip MD	432.14	552.55	530.95	629.10	0.59	0.35	0.39	0.11
Skip NPP0	463.63	567.08	532.26	635.25	0.52	0.38	0.38	0.09
Skip Age	462.65	519.62	535.68	633.44	0.57	0.45	0.39	0.10
Skip Lat	469.06	555.51	628.93	620.69	0.51	0.32	0.11	0.13

4. Discussion

4.1. Comparison of Results Obtained by Models

The results of the training and testing phase in models indicated that ANNs performed superior to the regression models. The GRNN model was found to have preferred accuracies over the BPNN model, while MNLR was superior to MLR in predicting CO_2 emission from reservoirs.

This outcome is consistent with other relevant literature, for example, MLR, MNLR, and GRNN models were used to forecast GHG emissions at national level and the results showed that the GRNN model gave the most preferable results [25]. When Firat et al. [42] predicted the scour depth around circular bridge piers based on the data from various studies, the GRNN model performed superior to BPNN and MLR. Among different types of ANNs, BPNN often shows a fair performance because the best model could be obtained by iterant parameters adjustment after calibration in the models [43]. However, the problem of overtraining often accompanies accuracy, because a large number of weights and biases is generated from many iterations [34]. In contrast, GRNN is a one-pass learning network, which does not require an iterative strategy as in BPNN. Therefore, GRNN can avoid this problem of overfitting to a large extent. Meanwhile, the problem of initial values determination and local minima often occurs in training stage of BPNN, while this problem does not exist in the GRNN procedure [44]. Thus, the GRNN can be preferred over the BPNN.

The regression lines in Figure 3 showed the same tendencies that the slopes were less than 1 and the slops in testing phases were less than ones in corresponding training phases. The first tendency reflects the problems of underestimation for high values of CO_2 flux in training and testing data sets, which is general in statistical prediction models [24,45]. Because the inputs cannot totally explain the outputs especially for extreme values, this tendency can also partly attribute to the non-homogeneous nature of data. Since CO_2 fluxes were measured from various reservoirs in different years, the lower slope in testing phases is due to the differences in training and testing data ranges. As listed in Table 3, the median and standard deviation of the testing data set were higher than those of the training data set.

In previous studies, St. Louis et al. [8] made a unary regression analysis between CO_2 and age based on datasets of 15 reservoirs (R^2 = 0.35), and Deemer et al. [12] built the relationship between CO_2 and mean annual precipitation based on datasets of 31 reservoirs (R^2 = 0.11). Barros et al. [11] used the multiple regression analysis to describe the relationship between three dependent variables (age, latitude, and DOC) and CO_2 flux based on datasets of 73 reservoirs (R^2 = 0.40). Compared with these regression models, our GRNN showed higher R^2 in both training (R^2 = 0.61) and testing (R^2 = 0.76) phases. The superior performance benefits from not only the advantages of GRNN, but also the larger database. Besides, without a testing process in the previous regression models, it is difficult to evaluate their generalization ability and apply in other reservoirs credibly. This study demonstrates that GRNN models could be an appropriate approach for prediction CO_2 emissions from reservoirs in other study systems.

4.2. Sensitivity Analysis

The results of sensitivity analysis demonstrated that the input variables of NPP0, age, and depth is the best set for MLR, which conformed to the results of Spearman correlation analysis. Moreover, the results also emphasized the importance of the latitude parameter in ANNs and MNLR that aim to fit non-linear functions, which proved the rationality of the selection of input variables. However, the significant relationship between latitude and CO_2 flux was not reflected adequately in the results of Spearman correlation analysis. The possible reason is that the relationship between latitude and CO_2 flux is non-linear. Consequently, the use of non-linear statistical dependence measures is more appropriate for determining inputs to ANN models [31].

4.3. Application of Established Model

4.3.1. Estimation of the Global Magnitude of the CO_2 Fluxes from Reservoirs

With the help of this GRNN, global magnitude of CO_2 emissions from documented reservoirs can be estimated. This estimation was based on the GRanD, which contains 6862 records of reservoirs updated in 2011 [19], and HANPP [32]. We selected latitude, the year of construction and the average depth from GRanD and extracted NPP0 from HANPP following the site of these reservoirs.

To predict CO_2 fluxes from global reservoirs by the tested GRNN, the confidence interval should be given together. However, ANNs are black box models that cannot be described as particular equations. Therefore, the potential predictive confidence interval was given based on the statistical analysis between observed and predicted CO_2 fluxes in testing phase. The methods to calculate confidence interval are as follows [46]: (a) the errors between observed and predicted values in testing phase were calculated; (b) the Bootstrap samples were created by resampling from the errors on 100,000 replicates; (c) the medians of Bootstrap error samples are computed; (d) the $1 - \alpha$ Bootstrap Pivotal Confidence Intervals (CIs) for median of errors are estimated by Equation (10):

$$C_n = \left(2\hat{\theta} - \widehat{\theta^*}\ ((1-\alpha/2)B),\ 2\hat{\theta} - \widehat{\theta^*}\ ((\alpha/2)B)\right) \tag{10}$$

where, $\hat{\theta}$ is an estimator of parameter θ; $\widehat{\theta^*}\ ((\alpha/2)B)$ is the $\alpha/2$ sample quantile of Bootstrap $\hat{\theta}$ samples. In this study, parameter θ is median of errors between observed and predicted values; α is 0.05, which means 95% confidence level. Bootstrap samples of medians were generated by Matlab R2013a.

After inputting the variables, the absolute values of latitude, the age of reservoirs in 2011, NPP0, and the average depth, we got the CO_2 emission fluxes from 6862 reservoirs in 2011. Because 95% CI of error from GRNN is (−68.11, 60.86), these fluxes were updated into intervals for subsequent estimations. Then we multiplied these fluxes, corresponding area, and the number of days in a year that CO_2 can diffuse on the surface of reservoirs. Considering the influence of seasonality, especially the ice cover, we made an assumption that the temperate reservoirs which located higher than 30° N or 30° S latitudes are ice-free for 200 days on average. This refers to the one of the study in St. Louis et al. [8], since only this study takes ice cover into account among pervious estimations listed in Table 6. As a result, we estimate that global reservoirs emit 40.03 Tg C yr^{-1} as CO_2 (5th and 95th confidence interval: 32.03–47.18 Tg C yr^{-1} as CO_2).

Compared to previous estimations with the same magnitude of area (Table 6), our estimation is moderate and more fairly accurate. However, the CO_2 flux estimated by St. Louis et al. [8] is larger, which might be caused by the overestimation of global reservoirs' area and the young age of sampled reservoirs. Only three tropical reservoirs with the average age of 7.70 years were used to estimate CO_2 fluxes from all tropical reservoirs. The estimation by Hertwich [16] is also a little larger, because CO_2 flux is multiplied by an uncertainty factor of 2. The previous estimations derived from the product of the average of CO_2 flux in database multiplied with the global surface area of reservoirs, while the estimation in this study took into account the annual-scale flux variability of a special reservoir and the difference in geographical position among global reservoirs. However, further refinement is still required for more precise estimates. Since there are no real data of time-series, this model cannot predict potential CO_2 fluxes in long temporal scale. The cause–effect relationships between water quality and CO_2 flux received little attention in this study because of the limited data. The direct and indirect influences from ice cover especially in the boreal region should be fully studied and accurately calculated in future estimations.

Table 6. The global CO_2 flux estimates and parameters of estimations.

Studies		Sample Size	Type of Dataset	Method	Area (10^5 km^2)	CO_2 (Tg C yr^{-1})
This study		251	All reservoirs	Individual [1]	4.47	40.03 [3]
Previous studies	Deemer et al. [12]	229	All reservoirs	Average [2]	3.1	36.8
	Hertwich [16]	142	Hydroelectric	Average [2]	3.3	76
	Barros et al. [11]	85	Hydroelectric	Average [2]	3.4	48
	St. Louis et al. [8]	19	All reservoirs	Average [2]	15.0	272.2

Note: (1) [1] Individual method, calculated by the sum of CO_2 flux from each individual reservoir; [2] Average method, calculated by the multiplication of average CO_2 flux of database and total reservoirs area. (2) [3] The 95% confidence interval is (32.03, 47.18) Tg C yr^{-1}.

4.3.2. Estimations of CO_2 Emissions from a Planned Reservoir

Considering the input variables of this GRNN, the possible CO_2 emission flux during a fixed period of time from a proposed reservoir with planned location and depth can be estimated by this model. Therefore, it is possible to give guidance for dam construction. A hypothetical case, which was not based on any actual events, was used to show this potential utilization of GRNN in this study. We assumed that a reservoir would be constructed, and the geographical coordinates could be selected from 23.5° N to 26.4° N with the same longitude (115° E), and the mean depth could vary from 5 m to 34 m. The possible CO_2 fluxes emission from this reservoirs in the first 20 years with different features were shown in Figure 4. Although the construction of reservoirs should consider many realistic questions, the possible carbon emissions from new reservoirs might also be used as an important index of reservoir construction in the future.

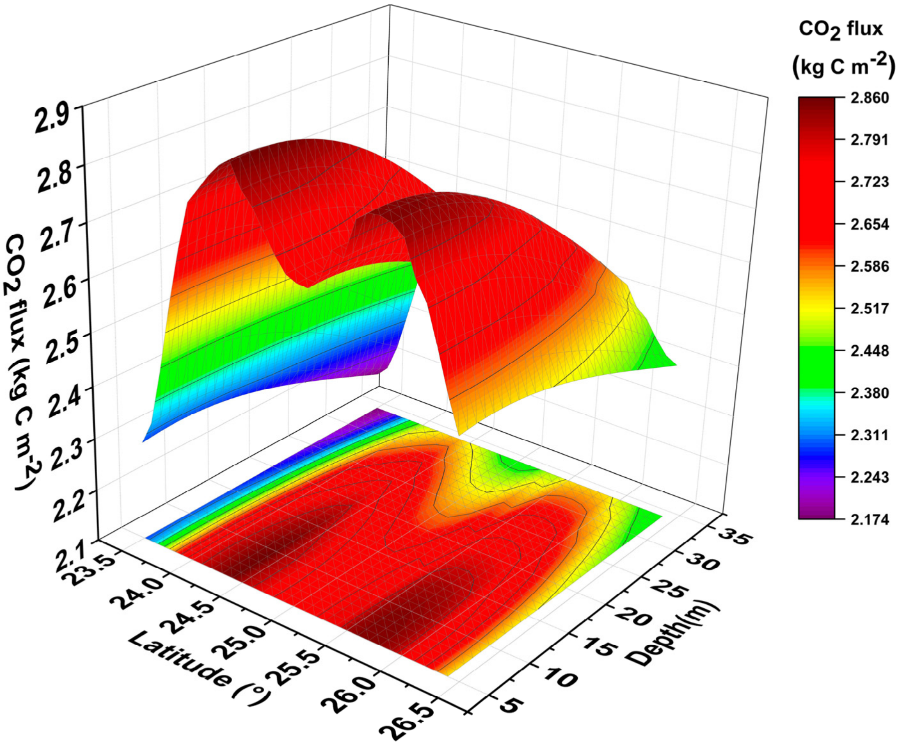

Figure 4. CO_2 emissions from reservoirs with different locations and depths in the first 20 years.

4.4. Future Research Directions

In this study, ANNs were built based on the data of gross CO_2 emissions from existing reservoirs, ignoring the potential CO_2 emissions from that land before impoundment, which might overestimate greenhouse effect from reservoirs. Recent study shows that the net carbon emissions from reservoirs are determined by different types of areas before flooding and provides a simple approach to quantify the net CO_2 emissions [47]. Future studies are therefore necessary to simulate and predict net CO_2 flux emits from reservoirs. Moreover, the combination of classification and regression machine learning can be a promising approach.

As the first attempt to apply ANNs to GHG emissions from reservoirs, CO_2 emission was chosen to simulate because of the quantity and quality of the monitoring data. However, CH_4 is a more powerful GHG than CO_2 [12]. Unlike CO_2 emissions, CH_4 emissions from reservoirs are new and anthropogenic [47]. Therefore, CH_4 footprint should be simulated and the magnitude needs to be estimated based on various released pathways in the future.

5. Conclusions

In this study, ANNs (GRNN and BPNN) and multiple regression models (MNLR and MLR) are applied to predict CO_2 emissions from reservoirs based on data records collected from published various field studies. Input variables used in models were selected by both Spearman correlation analysis and domain knowledge. The performance of models and observation was compared and evaluated by the indexes of RMSE, MAE, R^2, and NSE. It appears that the performance of ANNs is superior to the one of regression models. The GRNN's performance is better than BPNN's, while MNLR is superior to MLR. Sensitivity analysis of these four models confirmed that latitude-value is an important parameter in predicting CO_2 flux. The results demonstrate that GRNNs have great potential to estimate CO_2 emission from reservoirs when it is hard to acquire the monitoring data. The statistical models deserve more attention, because they are effective tools to assess global GHG emissions from reservoirs and provide new insights into the consideration of reservoir's construction during the planning stage. However, since the accuracy and generalization of statistical models largely depend on the measured data, more monitoring will be required in global reservoirs systematically. For example, the global CO_2 flux can be predicted in a longer time scale with the data of continuous monitoring on special reservoirs located in different latitude. Moreover, the mechanism models should be built to understand the relationship between CO_2 emission and other environmental factors clearly in the future.

Supplementary Materials: The following are available online at www.mdpi.com/2073-4441/10/1/26/s1, Table S1: CO_2 emission measurements and other data of reservoirs analyzed in the paper.

Author Contributions: Zhonghan Chen and Ping Huang conceived the idea; Zhonghan Chen and Xiaoqian Ye collected and checked the dataset; Zhonghan Chen built the model and analyzed the data; Zhonghan Chen and Xiaoqian Ye wrote the paper. All authors reviewed the manuscript.

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

References

1. Rudd, J.W.M.; Hecky, R.E. Are hydroelectric reservoirs significant sources of greenhouse gases? *Ambio* **1993**, *22*, 246–248.
2. Tremblay, A.; Varfalvy, L.; Roehm, C.; Garneau, M. *Greenhouse Gas Emissions-Fluxes and Processes: Hydroelectric Reservoirs and Natural Environments*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2005; ISBN 3-540-23455-1.
3. Teodoru, C.R.; Bastien, J.; Bonneville, M.C.; del Giorgio, P.A.; Demarty, M.; Garneau, M.; Hélie, J.F.; Pelletier, L.; Prairie, Y.T.; Roulet, N.T. The net carbon footprint of a newly created boreal hydroelectric reservoir. *Glob. Biogeochem. Cycles* **2012**, *26*. [CrossRef]
4. Demarty, M.; Bastien, J.; Tremblay, A.; Hesslein, R.H.; Gill, R. Greenhouse gas emissions from boreal reservoirs in Manitoba and Québec, Canada, measured with automated systems. *Environ. Sci. Technol.* **2009**, *43*, 8908–8915. [CrossRef] [PubMed]
5. Roland, F.; Vidal, L.O.; Pacheco, F.S.; Barros, N.O.; Assireu, A.; Ometto, J.P.H.B.; Cimbleris, A.C.P.; Cole, J.J. Variability of carbon dioxide flux from tropical (Cerrado) hydroelectric reservoirs. *Aquat. Sci.* **2010**, *72*, 283–293. [CrossRef]
6. Mosher, J.J.; Fortner, A.M.; Phillips, J.R.; Bevelhimer, M.S.; Stewart, A.J.; Troia, M.J. Spatial and temporal correlates of greenhouse gas diffusion from a hydropower reservoir in the Southern United States. *Water* **2015**, *7*, 5910–5927. [CrossRef]
7. Bevelhimer, M.S.; Stewart, A.J.; Fortner, A.M.; Phillips, J.R.; Mosher, J.J. CO_2 is dominant greenhouse gas emitted from six hydropower reservoirs in Southeastern United States during peak summer emissions. *Water* **2016**, *8*, 15. [CrossRef]
8. St. Louis, V.L.; Kelly, C.A.; Duchemin, É.; Rudd, J.W.M.; Rosenberg, D.M. Reservoir surfaces as sources of greenhouse gases to the atmosphere: A global estimate. *Bioscience* **2000**, *50*, 766–775. [CrossRef]
9. Saidi, H.; Koschorreck, M. CO_2 emissions from German drinking water reservoirs. *Sci. Total Environ.* **2017**, *581*, 10–18. [CrossRef] [PubMed]
10. Fearnside, P.M.; Pueyo, S. Greenhouse gas emissions from tropical dams. *Nat. Clim. Chang.* **2012**, *2*, 382–384. [CrossRef]
11. Barros, N.; Cole, J.J.; Tranvik, L.J.; Prairie, Y.T.; Bastviken, D.; Huszar, V.L.M.; del Giorgio, P.; Roland, F. Carbon emission from hydroelectric reservoirs linked to reservoir age and latitude. *Nat. Geosci.* **2011**, *4*, 593–596. [CrossRef]
12. Deemer, B.R.; Harrison, J.A.; Li, S.; Beaulieu, J.J.; Delsontro, T.; Barros, N.; Bezerra-Neto, J.F.; Powers, S.M.; dos Santos, M.A.; Vonk, J.A. Greenhouse gas emissions from reservoir water surfaces: A new global synthesis. *Bioscience* **2016**, *66*, 949–964. [CrossRef]
13. Raymond, P.A.; Hartmann, J.; Lauerwald, R.; Sobek, S.; McDonald, C.; Hoover, M.; Butman, D.; Striegl, R.; Mayorga, E.; Humborg, C. Global carbon dioxide emissions from inland waters. *Nature* **2013**, *503*, 355–359. [CrossRef] [PubMed]
14. Zhao, Y.; Wu, B.F.; Zeng, Y. Spatial and temporal patterns of greenhouse gas emissions from three gorges reservoir of china. *Biogeosciences* **2013**, *10*, 1219–1230. [CrossRef]

15. Teodoru, C.R.; Prairie, Y.T.; del Giorgio, P.A. Spatial heterogeneity of surface CO_2 fluxes in a newly created eastmain-1 reservoir in Northern Quebec, Canada. *Ecosystems* **2011**, *14*, 28–46. [CrossRef]
16. Hertwich, E.G. Addressing biogenic greenhouse gas emissions from hydropower in LCA. *Environ. Sci. Technol.* **2013**, *47*, 9604–9611. [CrossRef] [PubMed]
17. Soumis, N.; Duchemin, É.; Canuel, R.; Lucotte, M. Greenhouse gas emissions from reservoirs of the Western United States. *Glob. Biogeochem. Cycles* **2004**, *18*. [CrossRef]
18. Li, S.Y.; Zhang, Q.F.; Bush, R.T.; Sullivan, L.A. Methane and CO_2 emissions from China's hydroelectric reservoirs: A new quantitative synthesis. *Environ. Sci. Pollut. Res.* **2015**, *22*, 5325–5339. [CrossRef] [PubMed]
19. Lehner, B.; Liermann, C.R.; Revenga, C.; Vörösmarty, C.; Fekete, B.; Crouzet, P.; Döll, P.; Endejan, M.; Frenken, K.; Magome, J.; et al. High-resolution mapping of the world's reservoirs and dams for sustainable river-flow management. *Front. Ecol. Environ.* **2011**, *9*, 494–502. [CrossRef]
20. Musenze, R.S.; Grinham, A.; Werner, U.; Gale, D.; Sturm, K.; Udy, J.; Yuan, Z.G. Assessing the spatial and temporal variability of diffusive methane and nitrous oxide emissions from subtropical freshwater reservoirs. *Environ. Sci. Technol.* **2014**, *48*, 14499–14507. [CrossRef] [PubMed]
21. De Faria, F.A.M.; Jaramillo, P.; Sawakuchi, H.O.; Richey, J.E.; Barros, N. Estimating greenhouse gas emissions from future Amazonian hydroelectric reservoirs. *Environ. Res. Lett.* **2015**, *10*, 1–13. [CrossRef]
22. Rosa, L.P.; dos Santos, M.A.; Gesteira, C.; Xavier, A.E. A model for the data extrapolation of greenhouse gas emissions in the Brazilian hydroelectric system. *Environ. Res. Lett.* **2016**, *11*. [CrossRef]
23. Dogan, E.; Sengorur, B.; Koklu, R. Modeling biological oxygen demand of the Melen River in Turkey using an artificial neural network technique. *J. Environ. Manag.* **2009**, *90*, 1229–1235. [CrossRef] [PubMed]
24. Lu, F.; Chen, Z.; Liu, W.Q.; Shao, H.B. Modeling chlorophyll-a concentrations using an artificial neural network for precisely eco-restoring lake basin. *Ecol. Eng.* **2016**, *95*, 422–429. [CrossRef]
25. Antanasijević, D.; Pocajt, V.; Ristić, M.; Perić-Grujić, A. Modeling of energy consumption and related GHG (greenhouse gas) intensity and emissions in Europe using general regression neural networks. *Energy* **2015**, *84*, 816–824. [CrossRef]
26. Antanasijević, D.Z.; Ristić, M.Đ.; Perić-Grujić, A.A.; Pocajt, V.V. Forecasting GHG emissions using an optimized artificial neural network model based on correlation and principal component analysis. *Int. J. Greenh. Gas Control* **2014**, *20*, 244–253. [CrossRef]
27. Antonopoulos, V.Z.; Antonopoulos, A.V. Daily reference evapotranspiration estimates by artificial neural networks technique and empirical equations using limited input climate variables. *Comput. Electron. Agric.* **2017**, *132*, 86–96. [CrossRef]
28. Vakili, M.; Sabbagh-Yazdi, S.R.; Khosrojerdi, S.; Kalhor, K. Evaluating the effect of particulate matter pollution on estimation of daily global solar radiation using artificial neural network modeling based on meteorological data. *J. Clean. Prod.* **2017**, *141*, 1275–1285. [CrossRef]
29. Ding, W.F.; Zhang, J.S.; Leung, Y. Prediction of air pollutant concentration based on sparse response back-propagation training feedforward neural networks. *Environ. Sci. Pollut. Res.* **2016**, *23*, 19481–19494. [CrossRef] [PubMed]
30. Salami, E.S.; Salari, M.; Ehteshami, M.; Bidokhti, N.T.; Ghadimi, H. Application of artificial neural networks and mathematical modeling for the prediction of water quality variables (case study: Southwest of Iran). *Desalin. Water Treat.* **2016**, *57*, 27073–27084. [CrossRef]
31. 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]
32. Haberl, H.; Erb, K.H.; Krausmann, F.; Gaube, V.; Bondeau, A.; Plutzar, C.; Gingrich, S.; Lucht, W.; Fischer-Kowalski, M. Quantifying and mapping the human appropriation of net primary production in earth's terrestrial ecosystems. *Proc. Natl. Acad. Sci. USA* **2007**, *104*, 12942–12947. [CrossRef] [PubMed]
33. Antonopoulos, V.Z.; Gianniou, S.K.; Antonopoulos, A.V. Artificial neural networks and empirical equations to estimate daily evaporation: Application to Lake Vegoritis, Greece. *Hydrol. Sci. J.* **2016**, *61*, 2590–2599. [CrossRef]
34. Nielsen, M.A. *Neural Networks and Deep Learning*; Determination Press: USA, 2015. Available online: http://neuralnetworksanddeeplearning.com/ (accessed on 29 December 2017).

35. He, B.; Oki, T.; Sun, F.B.; Komori, D.; Kanae, S.; Wang, Y.; Kim, H.; Yamazaki, D. Estimating monthly total nitrogen concentration in streams by using artificial neural network. *J. Environ. Manag.* **2011**, *92*, 172–177. [CrossRef] [PubMed]
36. Specht, D.F. A general regression neural network. *IEEE Trans. Neural Netw.* **1991**, *2*, 568–576. [CrossRef] [PubMed]
37. 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]
38. Sammen, S.S.H.; Mohamed, T.A.; Ghazali, A.H.; El-Shafie, A.H.; Sidek, L.M. Generalized regression neural network for prediction of peak outflow from dam breach. *Water Resour. Manag.* **2017**, *31*, 549–562. [CrossRef]
39. Yaseen, Z.M.; El-Shafie, A.; Jaafar, O.; Afan, H.A.; Sayl, K.N. Artificial intelligence based models for stream-flow forecasting: 2000–2015. *J. Hydrol.* **2015**, *530*, 829–844. [CrossRef]
40. Zhao, D.Q.; Chen, J.N.; Wang, H.Z.; Tong, Q.Y.; Cao, S.B.; Sheng, Z. Gis-based urban rainfall-runoff modeling using an automatic catchment-discretization approach: A case study in Macau. *Environ. Earth Sci.* **2009**, *59*, 465–472. [CrossRef]
41. Csábrági, A.; Molnár, S.; Tanos, P.; Kovács, J. Application of artificial neural networks to the forecasting of dissolved oxygen content in the hungarian section of the river danube. *Ecol. Eng.* **2017**, *100*, 63–72. [CrossRef]
42. Firat, M.; Gungor, M. Generalized regression neural networks and feed forward neural networks for prediction of scour depth around bridge piers. *Adv. Eng. Softw.* **2009**, *40*, 731–737. [CrossRef]
43. Safari, M.-J.-S.; Aksoy, H.; Mohammadi, M. Artificial neural network and regression models for flow velocity at sediment incipient deposition. *J. Hydrol.* **2016**, *541*, 1420–1429. [CrossRef]
44. Stamenković, L.J.; Antanasijević, D.Z.; Ristić, M.Ð.; Perić-Grujić, A.A.; Pocajt, V.V. Modeling of methane emissions using artificial neural network approach. *J. Serbian Chem. Soc.* **2015**, *80*, 421–433. [CrossRef]
45. Wang, L.; Kisi, O.; Zounemat-Kermani, M.; Salazar, G.A.; Zhu, Z.; Gong, W. Solar radiation prediction using different techniques: Model evaluation and comparison. *Renew. Sustan. Enery Rev.* **2016**, *61*, 384–397. [CrossRef]
46. Wassermann, L. *All of Nonparametric Statistics*, 3rd ed.; Springer: New York, NY, USA, 2006; pp. 30–35, ISBN 978-0-387-30623-05.
47. Prairie, Y.T.; Alm, J.; Beaulieu, J.; Barros, N.; Battin, T.; Cole, J.; Giorgio, P.D.; DelSontro, T.; Guérin, F.; Harby, A.; et al. Greenhouse gas emissions from freshwater reservoirs: What does the atmosphere see? *Ecosystems* **2017**, 1–14. [CrossRef]