sustainability

MDPI

Article

Evaluation of Nitrate Load Estimations Using Neural Networks and Canonical Correlation Analysis with K-Fold Cross-Validation

Kichul Jung [1], Deg-Hyo Bae [2], Myoung-Jin Um [3], Siyeon Kim [1], Seol Jeon [1] and Daeryong Park [1,*]

1 Department of Civil and Environmental Engineering, Konkuk University, Seoul 05029, Korea
2 Department of Civil and Environmental Engineering, Sejong University, Seoul 05006, Korea
3 Department of Civil Engineering, Kyonggi University, Suwon 16227, Korea
* Correspondence: drpark@konkuk.ac.kr

Received: 6 November 2019; Accepted: 28 December 2019; Published: 3 January 2020

Abstract: The present work aimed to examine the feasibility of using artificial neural network (ANN) based models to obtain accurate estimates of nitrate loads in river basins, which is an important parameter for water quality management. Both Single ANN (SANN) and Ensemble ANN (EANN) models were used to obtain the load estimations for five river basins in the Midwest United States. These basins included the Cuyahoga, Raisin, Sandusky, Muskingum, and Vermilion basins in Michigan and Ohio. Further, canonical correlation analysis (CCA) was applied to the ANN models to improve the performance. The k-fold cross-validation method was then utilized to evaluate the proposed models based on two statistical indices, namely, the *rRMSE* and *rBAIS*, and the estimates were compared for four different k values (k = 3, 5, 7, and 10). According to the results, the EANN model seemed to produce better load estimations than the SANN model, and the CCA based EANN model tended to produce the best estimates among all of the proposed models in this study. The box plot data for the *rRMSE* index were also investigated, and the plot results indicated that increasing values of k tended to generate better estimates. Thus, the use of k = 10 is recommended for load estimations since this value was associated with better performances and less biased estimates.

Keywords: single artificial neural network; canonical correlation analysis; ensemble artificial neural network; k-fold cross-validation; load estimations; Midwest; nitrate

1. Introduction

Nutrient enrichment is a growing problem in rivers and streams, where excessive nutrients can cause degradation in water quality. Nutrients represent one of the most problematic water quality constituents in rivers in the Midwestern United States [1–3]. In particular, nutrient levels in streams and rivers in the state of Ohio are a critical issue and have been monitored for many years to obtain accurate nitrate load estimations. For designing suitable conservation measures or reduction strategies, it is necessary to accurately calculate the nutrient loads on a monthly, seasonal, and yearly basis at each monitoring station. However, evaluation results for nutrient loads typically contain many potential sources of uncertainties including those related to the models used and the data sets.

Various type of methods have been developed to resolve the data gaps encountered when estimating nutrient loads [2,4]. One well-known approach for estimating water quality constituents is a regression based method, in which water data are correlated with the constituents of concern such as the nutrient load. The United States Geological Survey (USGS) has developed several methods to calculate various water quality constituents. The most well-known USGS methods are based on

multiple regression techniques that relate observed concentrations with the daily discharge, time, and season. In obtaining water quality constituent data, the estimation of daily nutrient concentrations is often a critical issue.

Annual, seasonal, and monthly load estimations are important because these loads are the summation of the daily load multiplied by the daily discharge and daily nutrient concentration [4–7]. Nutrient concentrations are often not monitored every day and over long periods of time. Hence, many researchers have suggested different approaches for estimating the missing nutrient concentrations. The most representative methods for load estimations are based on regression analyses between the streamflow and nutrient concentrations. Cohn et al. [8] and Cohn [4] suggested the use of a regression model with seven parameters to estimate daily concentrations. This model estimates the logarithm-transformed concentrations through use of a second-order polynomial regression equation with data on logarithm-transformed daily flows and decimals of time. The algorithm of the regression method can be obtained by downloading the LOADEST or FLUXMASTER load-estimation software package from USGS. Hirsch et al. [9] also developed another load estimation method, which is referred to as the Weighted Regressions on Time Discharge and Season (WRTDS). This method estimates the logarithm of daily concentrations by using the sine and cosine transformations of decimal time, and the logarithm of daily discharge. The method is implemented with five or seven parameter equations. The main input parameters in WRTDS are decimal time and streamflow discharge. One of the important processes in WRTDS is the estimation of weights for each day in the sample depending on the differences in the values of the variables between the prediction and sample day [10]. In a more recent study, the algorithm of WRTDS was applied to Exploration and Graphics for RivEr Trend (EGRET) to enhance load estimations [11].

The prediction of nitrate concentrations in streamflow by using artificial intelligence algorithms has been studied. Markus et al. [12] and Markus et al. [13] used artificial neural network (ANN) based models to forecast weekly nitrate concentrations. They also compared those ANN model results with results from evolutionary polynomial regressions and naïve Bayes models for several watersheds in Illinois, USA. The authors demonstrated that the most outstanding models differed depending on the error evaluation method used, and they proposed a multi-tool approach for analysis. Besides, many other studies have applied ANN models to predict the monthly biological oxygen demand (BOD) [14,15], monthly total nitrogen content, total phosphorus content, and dissolved oxygen level [16,17] in various types of rivers located in different countries.

Several research projects have been conducted to estimate various hydrological variables based on an Ensemble ANN (EANN) approach. These studies used hydrological variables that have been spatially and temporally monitored for long periods of time, and the results seem to suggest that an EANN approach is appropriate for application in artificial intelligence algorithms. For example, the EANN approach has been applied to simulations and forecasts of the rainfall-runoff process [18], flood frequency [19,20], peak discharge [21], and monthly potential evapotranspiration [22]. These studies proposed that the EANN was more effective than a Single ANN (SANN) or other existing physical approaches. Moreover, cross-validation techniques have been widely used for different hydrologic variables to assess the estimates obtained from hydrological models [23–26].

Recently, the EANN approach has been applied for forecasting and simulations of water quality constituents to improve the estimation modeling. Kan et al. [27] used an EANN based on a hybrid function approximator, named the PEK model, to simulate runoff in three different catchments in China. They investigated the results for the runoff hydrograph and peak flow derived with the EANN by comparing the model performance with the performances of two physical runoff models, namely, the Xinanjiang model and the IHACRES (identification of unit hydrographs and component flows from rainfall, evaporation, and streamflow) model. The authors demonstrated that the performances of the EANN model for runoff and peak flow results were better than those of the other two physical watershed models tested. In addition, Huang and Gao [28] applied an EANN to simulate chlorophyll with other seven water quality parameters. They reported that the ensemble simulations were affected by the

ensemble size and that determination of the appropriate ensemble size was significant for ensemble simulations. However, little research has been carried out to investigate EANN applications with multivariate statistics for enhancing load estimations and to evaluate the load estimation performances based on both SANN and EANN approaches.

In the present study, we aimed to identify a better estimation model for obtaining load estimations, which can be used for nutrient concentration simulations in Midwest streams or rivers. The SANN and EANN models were applied to determine a proper model by investigating daily load estimations and by comparing the performances. Further, multivariate statistics, namely, canonical correlation analysis (CCA) results, were used to improve the load estimations by establishing a correlation structure between the data sets of two variables that were strongly related to nutrient concentrations. For the model validation, this study utilized the k-fold cross-validation technique, which involved identifying the appropriate value of folds and evaluating the model performance.

2. Data Sets

For the analysis of load estimations, we focused on Midwest river basins in the USA and used data from five stations located in the Cuyahoga, Raisin, Sandusky, Muskingum, and Vermilion basins. The five stations cover river basins characterized by various areas ranging from 697 km^2 to 19,208 km^2. The average for the nitrate concentration ranged from 1.389 to 3.957 mg/L, and the average for the discharge ranged from 491.013 to 8642.170 m^3/s. The monitoring durations were 35, 27, 35, 22, and 7 years for the Cuyahoga, Raisin, Sandusky, Muskingum, and Vermilion basins, respectively. The land in the river basins is basically agricultural, urban, and wooded land. The agricultural areas represent a large portion of the overall land use, and agricultural activities here have led to high nitrate concentrations downstream in Lake Erie and the Mississippi River basin. Around the Cuyahoga, Raisin, Sandusky, Muskingum, and Vermilion stations, the proportion of land used for agriculture amounts to 17%, 72%, 83%, 71%, and 52%, respectively. The drainage basins also contain many wetlands, lakes, and floodplain forests where ecological resources are plentiful. Figure 1 shows the five river basins studied in the present study.

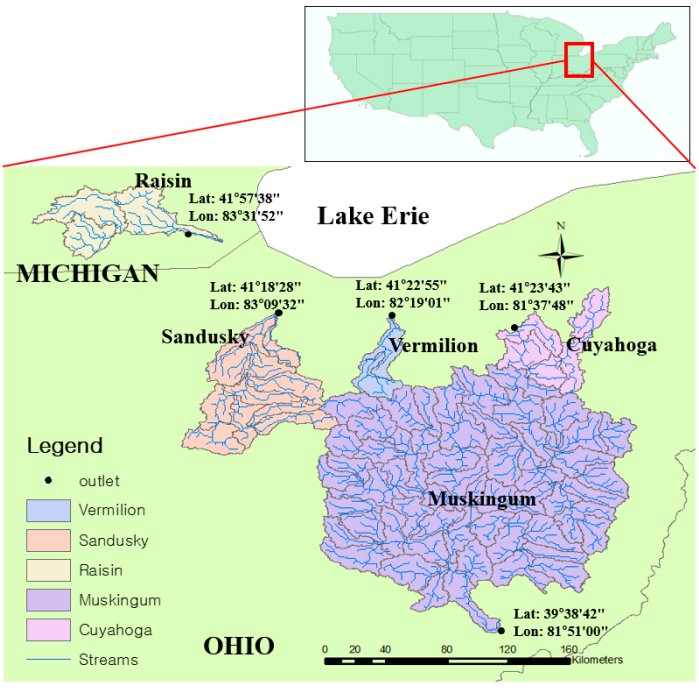
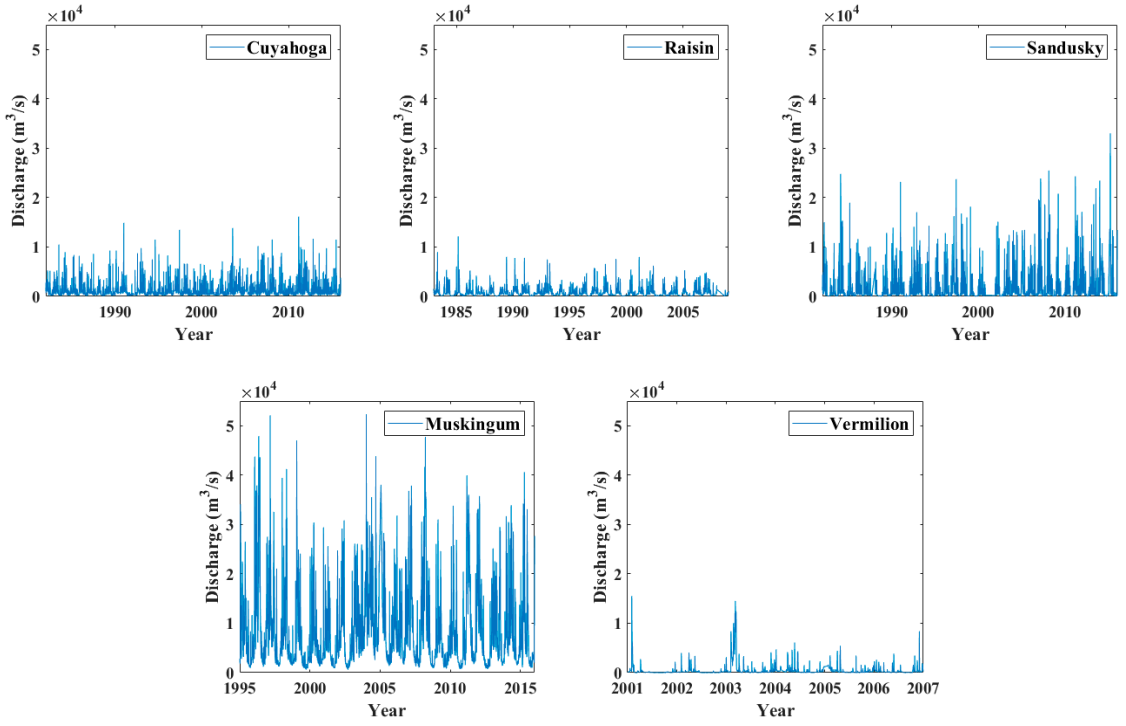


Figure 1. Five river basins and locations of stations used in this study. Gray lines within each basin indicate the Hydrologic Unit Code (HUC) 10, and blue lines indicate the streams in each basin.

To conduct the analysis of load estimations, we used daily discharge data, nitrate concentrations, and the day of year at the five stations. The daily discharge was denoted as Q, and the daily nitrite plus nitrate (NO_2-N+NO_3-N) data were represented as nitrate (NO_3) in this study. The loads we aimed to estimate can be calculated by $Q \times NO_3$. The data sets for the analysis of load estimations were obtained from USGS (https://waterdata.usgs.gov/nwis/sw) for the discharge data and the Water Quality Laboratory (WQL) of the National Center for Water Quality Research at Heidelberg University, Tiffin, Ohio (https://www.heidelberg.edu/tributary-data-download) for the nitrate data. The periods of the data sets were 1982–2016 for the Cuyahoga station, 1983–2009 for the Raisin station, 1982–2016 for the Sandusky station, 1995–2016 for the Muskingum station, and 2001–2007 for the Vermilion station. A description of each station is presented in Table 1. Figure 2 shows Q and nitrate concentration for the five stations.

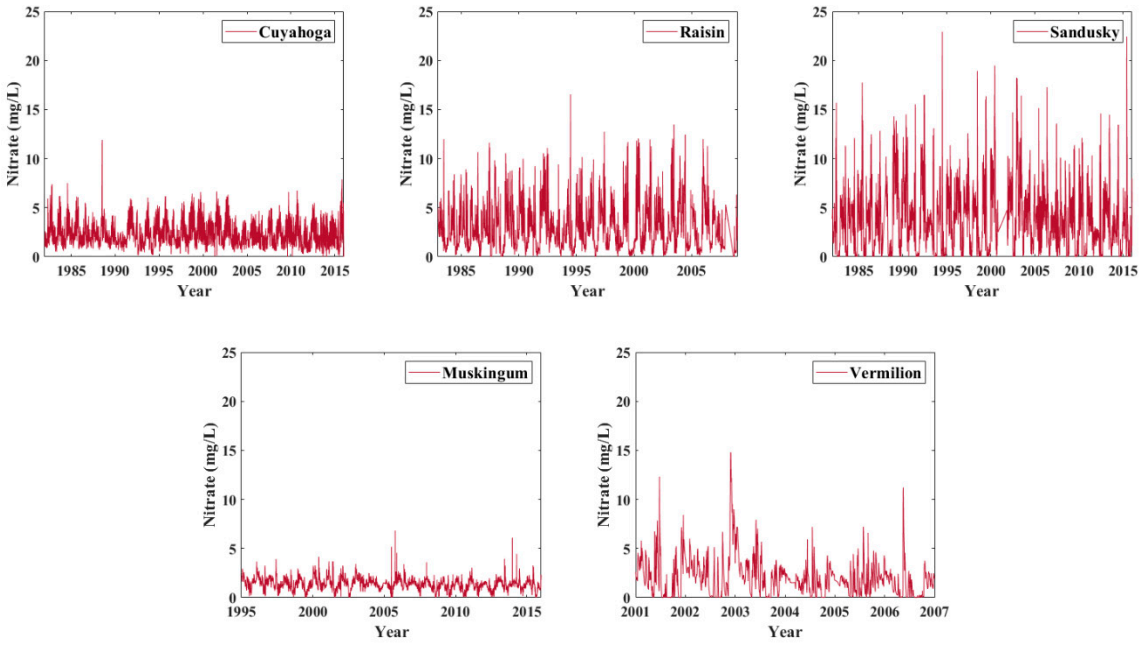
Table 1. Descriptive features for the five stations in the USA that were used for the estimation of nitrate loads.

Station Name	USGS Station Number	Year	Drainage Area (km^2)	Mean Discharge (m^3/s)	Mean Nitrate Concentration (mg/L)	Land Use (%)		
						Agriculture	Urban	Wooded
Cuyahoga	04208000	1982–2016	1843	1005.486	2.457	17	47	35
Raisin	04176500	1983–2009	2755	825.608	2.962	72	11	16
Sandusky	04198000	1982–2016	3285	1474.102	3.957	83	9	8
Muskingum	03150000	1995–2016	19,208	8642.170	1.389	52	2	43
Vermilion	04199500	2001–2007	697	491.013	2.191	71	1	26



(**a**)

Figure 2. *Cont.*



(**b**)

Figure 2. Plots of (**a**) daily discharge (m^3/s) and (**b**) daily nitrate concentrations (mg/L) for the five river basins.

3. Methods

3.1. ANN Models for Load Estimations

In the present study, we constructed several ANN models to evaluate the optimal ANN model for estimating daily loads at the sites of interest. Variables for discharge, nitrate concentration, and time were used to obtain the load estimations. An ANN is an information processing model designed for reproducing certain structures and for identifying the interconnections among a group of nodes. Such a model can be utilized to solve complex problems related to classification, pattern recognition, and estimation as a nonlinear mathematical model. With an ANN model, multilayer perceptrons (MLPs) have been commonly used to provide the ANN predictions, which represent the model output [29]. The MLPs consist of three types of layers including input, hidden, and output layers, and they can be characterized as a feed forward supervised model. During this process, the input layer receives information consisting of the input variables, and then, the hidden layer connects the input layer with the output layer based on weighted acyclic arcs. The number of hidden neurons of the hidden layer plays a crucial role in predictions of the model output with the ANN model. The model can be overfitted when many hidden neurons are applied due to the use of an insufficient training sample, while the model can underfitted when few neurons are utilized due to the difficulty of determining functional relationships among the variables. Here, five hidden neurons were selected for the hidden layer to build the ANN model by considering the model performance. This number was also used in previous studies to estimate hydrological variables [20,23].

For the training algorithm in the ANN process, the Levenberg–Marquardt (LM) algorithm was applied for the determination of optimal solutions to decrease errors of the model. This algorithm is relatively faster and more accurate than other algorithms including the gradient descent algorithm. The scalar parameter of the LM algorithm was selected based on the analysis of Demuth et al. [30]. If the value of the scalar parameter is large, the algorithm follows the features of the gradient descent method, whereas if the value of the parameter is small, it follows the properties of the Gauss–Newton method. Here, the error of a specific configuration was identified and compared with the target output

by running the training samples. Then, early stopping criteria were applied in this study to find the optimal network parameters that could minimize the estimation error.

3.2. Ensemble ANN Models

Once we set up the SANN model for load estimations, the ensemble technique was used for creating the EANN model, which was applied to improve the ability of generalization and stability of model performance. Based on the purposes of this study, we compared the results obtained from the SANN model and the EANN model. The EANN model was based on a number of ANNs that were trained and generated by individual networks [31,32]. As a result, the number of ANN models is 14 in this study. In the EANN processes, the bagging approach was used to produce unique predictions of the model [33]. With the bagging method, a number of ANNs were trained based on a subset of the training set by solving a given problem. Then, the results generated by the individual networks were combined to produce the unique output, which is the output of the ensemble.

The size of an ensemble plays a significant role in the design of the EANN model as it defines the amount of information and determines the degree of homogeneity within the training subsets. If the size for the ensemble is relatively large, the time of training will be increased because of the many sub-models. This affects the amount of information assigned in each ensemble, and it decreases the model performance. If the size for the ensemble is relatively small, the ability for generalization and stability will be not improved. Different ensemble sizes were examined for this study, and an ensemble size of 14 was selected. The estimation error was gradually reduced by the size of 11, while subsequent increases in the size seemed to result in very little change in the error. Notably, Shu and Ouarda [20] investigated the size of the ensemble and chose a size of 14 for the EANN model with the bagging method.

3.3. Integration of ANN Models and CCA

The CCA technique is basically used to establish a linear relationship between two groups of random variables. This multivariate approach provides a general theoretical framework for factorial discriminant analysis and multivariate regression. Previously, CCA has been applied for estimations of hydrological variables and recommended for CCA based ANN models to enhance the generalization and performance [20,23]. If X and Y are two random variables, CCA computes two sets for basis vectors that are canonical variables. Given that W and V are a linear combination of X and Y, respectively, we have

$$W = \alpha' X \tag{1}$$

$$V = \beta' Y \tag{2}$$

where α' denotes the transpose of the vector α and β' denotes the transpose of the vector β.

The correlation between W and V is calculated as

$$\rho = \frac{\alpha' \sum_{XY} \beta}{\sqrt{\alpha' \sum_X \alpha \beta' \sum_Y \beta}} \tag{3}$$

Based on the above equation, the vectors of α and β are determined by maximizing the correlation, ρ. If the first pair of canonical variables is calculated, other pairs of the canonical variables are estimated based on the correlation subject to the constraint of unit variance for normalization. Note that X implies a set of discharge and time variables and Y implies a set of load variables in this study. A time variable was used in this study because the water quality and discharge can be influenced by seasonal trends. The CCA constructs a transformed space called a canonical space, and then, a calibration with data is conducted by establishing the functional relationship between the two sets of variables in the space. Detailed information on the process of CCA is available in the literature [34].

Once the discharge and time variables were projected for the use of the ANN model in the canonical space, the projected variables were fed to the model for estimations of load variables. The ANN model

approximates the functional relationships among the canonical variables as the input variables and load variables as the output variables. Through the MLP process, the output layers generate the ANN predictions. The present study used an integration of the SANN and CCA (SANN-CCA) and an integration of the EANN and CCA (EANN-CCA) to achieve improvements in the load estimations. Figure 3 presents a diagram of the processes to estimate loads using the ANN based models.

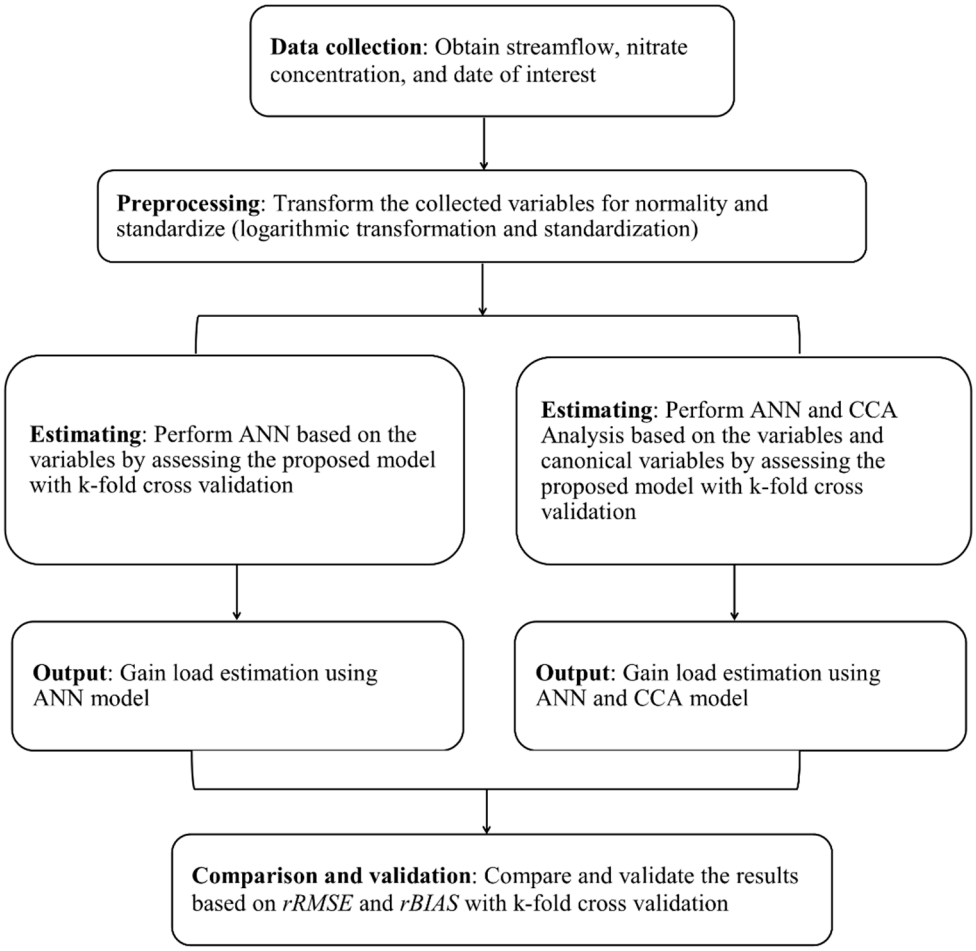


Figure 3. Diagram for processes used to obtain load estimations with the ANN and ANN-CCA.

3.4. Evaluation Approaches and Criteria

To assess the performance of models, cross-validation techniques have been commonly used as resampling methods [19,35,36]. This study used the k-fold cross-validation method to evaluate the relative performance of several ANN models during load estimations.

In the k-fold cross-validation procedure, the original sample is randomly grouped into k subsamples based on the same size. The subsamples are classified for a testing member and training members. The testing set as the validation data set represents an unknown data set, and the training sets represent known data sets. Then, the model conducts the analysis on the training sets and validates the analysis on the testing set. The cross-validation process is repeated k times to obtain a single estimate of the model output from the average value of the results for the different sets.

The ANN models can be assessed on the basis of two measures, namely, the relative root mean squared error (*rRMSE*) and the relative mean bias (*rBIAS*), which were used for flood quantile estimations [20]. The two measures can be computed as follows:

$$rRMSE = 100 \times \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{q_i - \hat{q}_i}{q_i}\right)^2} \quad (4)$$

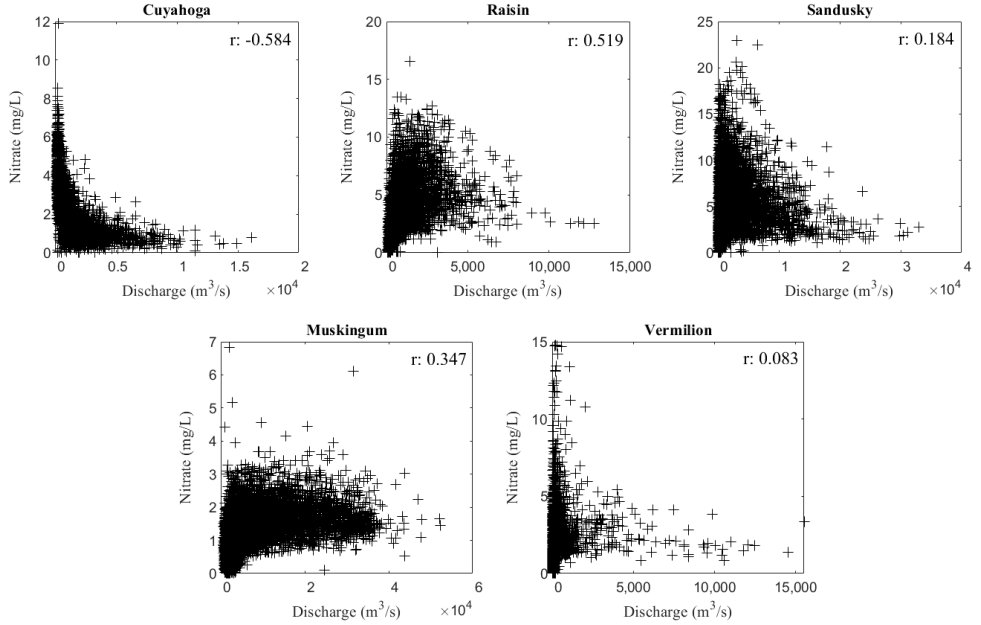
$$rBIAS = \frac{100}{n}\sum_{i=1}^{n}\left(\frac{q_i - \hat{q}_i}{q_i}\right) \quad (5)$$

where n is the total number of data points used for the analysis, q_i is the measured data for day i, and $\hat{q}_i$ indicates the load estimation derived from the ANN models for day i. The *rRMSE* ranges from zero to large positive numbers. The *rBIAS* ranges from large negative numbers to large positive numbers. The optimal value of both *rRMSE* and *rBIAS* is zero.

4. Results and Discussion

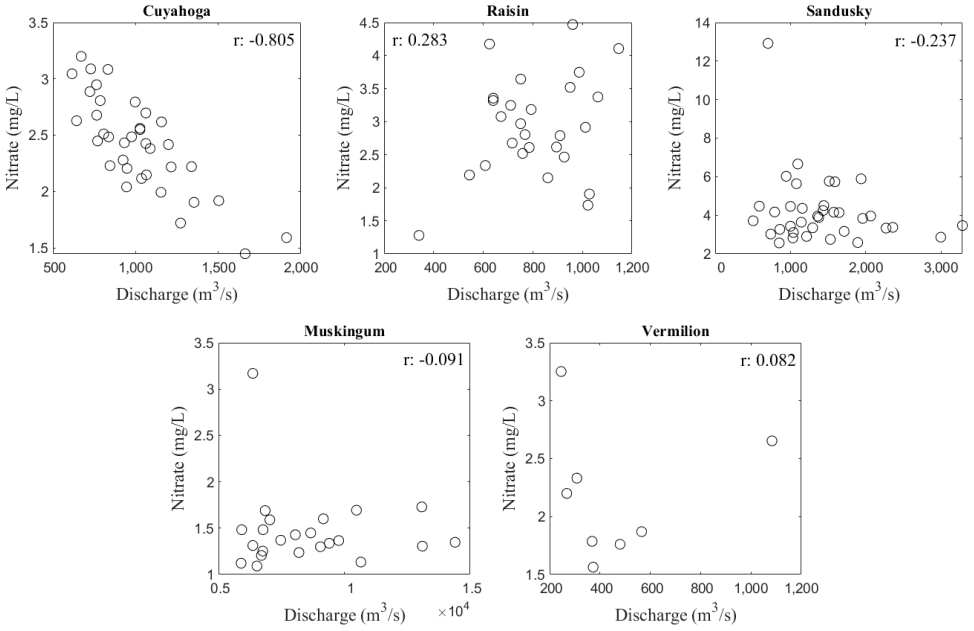
4.1. Single ANN and Ensemble ANN

In the analysis, the SANN and EANN models were structured for load estimations, which could be used to investigate nutrient concentrations and manage water quality. The k-fold cross-validation procedure was applied to the study areas, and 3-fold, 5-fold, 7-fold, and 10-fold cross-validation techniques were examined during the estimations of loads for five river basins. Rodriguez et al. [37] used various values of the folds including 2, 5, and 10 for the identification of optimal values of the folds. We also selected fold values less than 10 for the analysis of load estimations in this study. A correlation analysis for discharge and nitrate concentrations and for discharge and loads was conducted as shown in Figures 4 and 5. Figure 4a shows that correlation coefficients between the daily discharge and daily nitrate concentration ranged from −0.584 to 0.519. Figure 4b shows that correlation coefficients between the annual discharge and annual nitrate concentration ranged from −0.805 to 0.283. Additionally, Figure 5 presents that correlation coefficients between the daily discharge and daily load and between the annual discharge and annual load ranged from 0.777 to 0.919 and from 0.675 to 0.918, respectively.



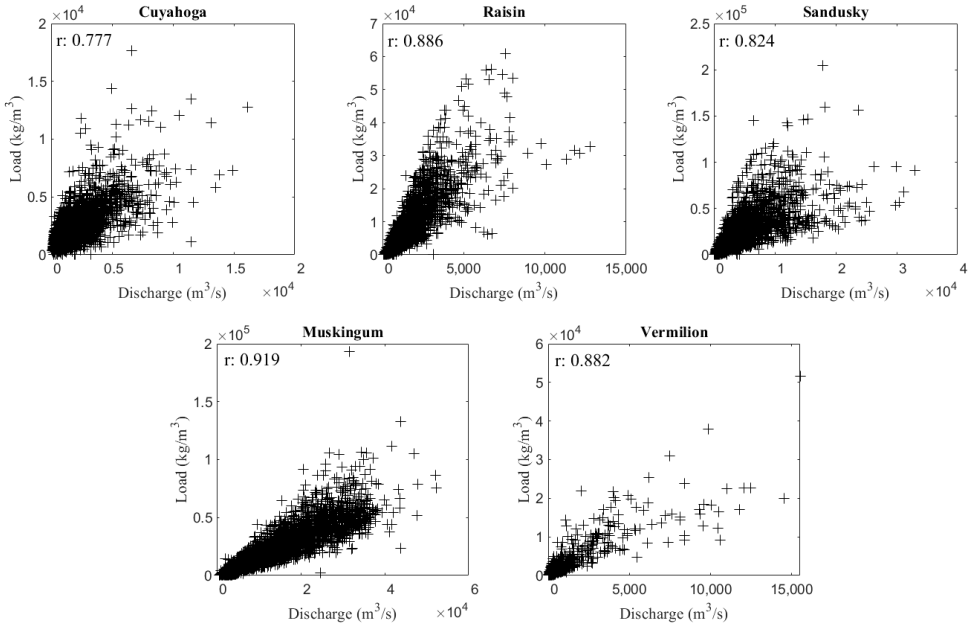
(**a**)

Figure 4. *Cont.*



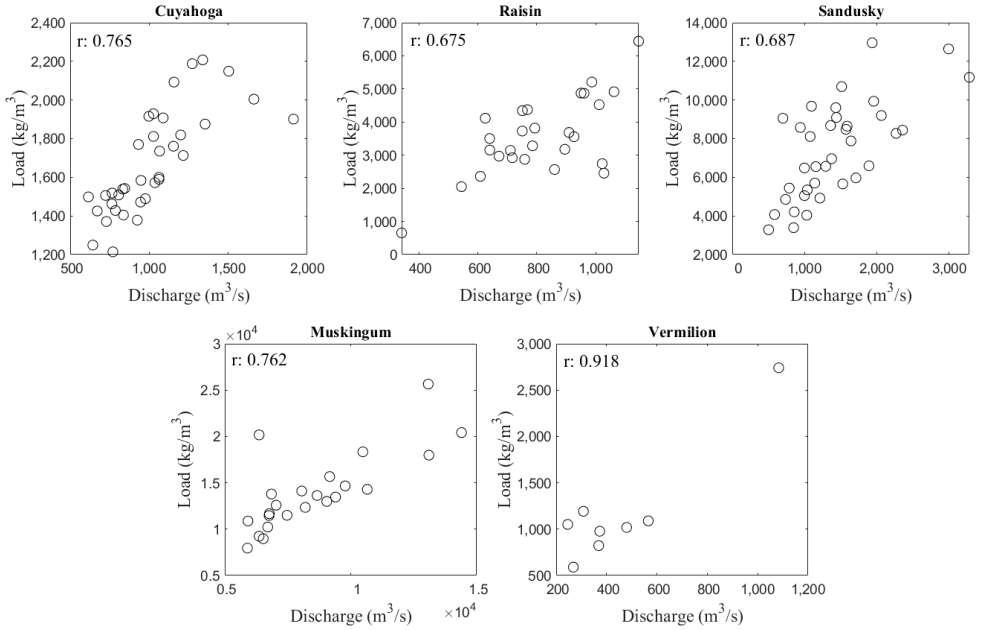
(**b**)

Figure 4. Plots of the (**a**) relationship between the daily discharge and daily nitrate concentration and (**b**) relationship between the annual discharge and annual nitrate concentration for each basin. r indicates the correlation coefficients.



(**a**)

Figure 5. *Cont.*



(**b**)

Figure 5. Plots of the (**a**) relationship between the daily discharge and daily nitrate load and (**b**) relationship between the annual discharge and annual nitrate load for each basin. r indicates the correlation coefficients.

With both the SANN and EANN models, we obtained load estimations for different cross-validations, and Table 2 presents the corresponding *rRMSE* and *rBIAS* indices for the 3-fold, 5-fold, 7-fold, and 10-fold cross-validations. The *rRMSE* index basically provides an assessment of accuracy for the ANN predictions, and the *rBIAS* index generally provides an indication of whether the proposed model seems to overestimate or underestimate. The results in Table 2 show that the *rRMSE* and *rBIAS* indices tended to improve for the five river basins when the number of folds in the cross-validation increased. The analysis using the *rBIAS* index shows that the models produce underestimated loads. Further, Figure 6 presents the *rRMSE* index derived from the SANN and EANN models with the different cross-validations. This figure indicates that the EANN model seemed to have a better performance than the SANN model according to the *rRMSE* criterion. The 10-fold cross-validation for each basin provided enhanced performances among the four different k-fold cross-validations tested.

Table 2. K-fold cross-validation results based on the SANN and EANN models for the five stations.

Stations		Single ANN				Ensemble ANN			
		3-fold	**5-fold**	**7-fold**	**10-fold**	**3-fold**	**5-fold**	**7-fold**	**10-fold**
Cuyahoga	***rRMSE* (%)**	352.831	311.046	306.131	237.211	321.558	297.889	241.552	214.550
	***rBIAS* (%)**	−13.800	−12.382	−13.469	−11.957	−11.798	−12.754	−11.202	−11.314
Rasin	***rRMSE* (%)**	280.892	274.257	270.516	258.239	280.725	270.368	260.635	257.411
	***rBIAS* (%)**	−31.303	−30.402	−30.380	−30.342	−32.069	−30.754	−30.942	−30.302
Sandusky	***rRMSE* (%)**	740.484	740.347	735.574	692.975	730.151	714.809	724.670	692.351
	***rBIAS* (%)**	−119.670	−116.277	−119.456	−114.721	−116.436	−114.695	−115.791	−114.841
Muskingum	***rRMSE* (%)**	303.115	299.618	295.394	285.894	309.235	300.883	295.583	280.123
	***rBIAS* (%)**	−35.836	−34.618	−34.795	−34.516	−35.810	−34.773	−35.031	−34.156
Vermilion	***rRMSE* (%)**	574.167	550.747	544.796	455.043	533.059	524.288	490.741	454.188
	***rBIAS* (%)**	−133.011	−128.781	−116.168	−109.642	−125.444	−128.916	−118.546	−111.249

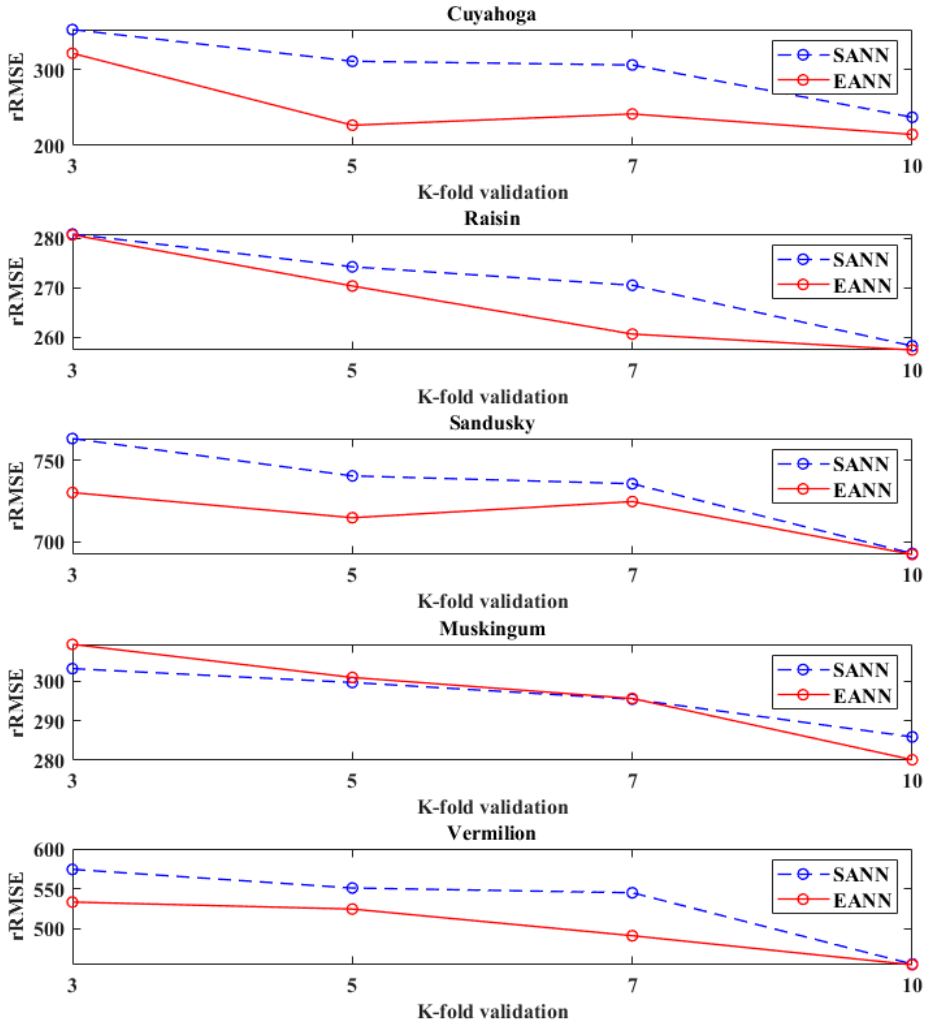


Figure 6. *rRMSE* of the SANN and EANN model results for the five stations based on different values of k in the cross-validation method.

Furthermore, we identified how the number of folds affects the model performance for load estimations based on the *rRMSE* of the SANN and EANN models. Figure 7 shows the results for the four types of fold cross-validation in the five river basins. In the figure, box plots are presented for each model. The center line in the box plot represents the median value of the load estimation, and the top and bottom of the plot show the 75th and 25th percentiles of the *rRMSE* of the estimation, respectively. The left box plots with the blue color indicate the results of the SANN model, while the right box plots with the red color indicate the results of the EANN model. The increasing number of k-folds showed a decreasing trend in both models. However, the model with the 10-fold cross validation for the five river basins presented a slightly decreasing tend or no noticeable trend compared to the model with the 7-fold cross validation. This phenomenon was observed in both the SANN and EANN models. To assess the sensitivity of the predictions based on k-fold cross-validation, Rodriguez et al. [37] conducted an experimental study in which they changed the training set with various values of k. In their analysis, the use of the low value of k = 2 seemed to produce the most biased result, and the use of k = 5 or 10 was recommended to obtain a less biased result on the basis of the experimental results. We also observed that a k value of 5 seemed to produce a better estimation than a k value of 7 in the Cuyahoga and Sandusky basins. Based on the load estimations examined in this study, we propose the application of a k value of 10, which tends to provide a less biased error estimator for the loads.

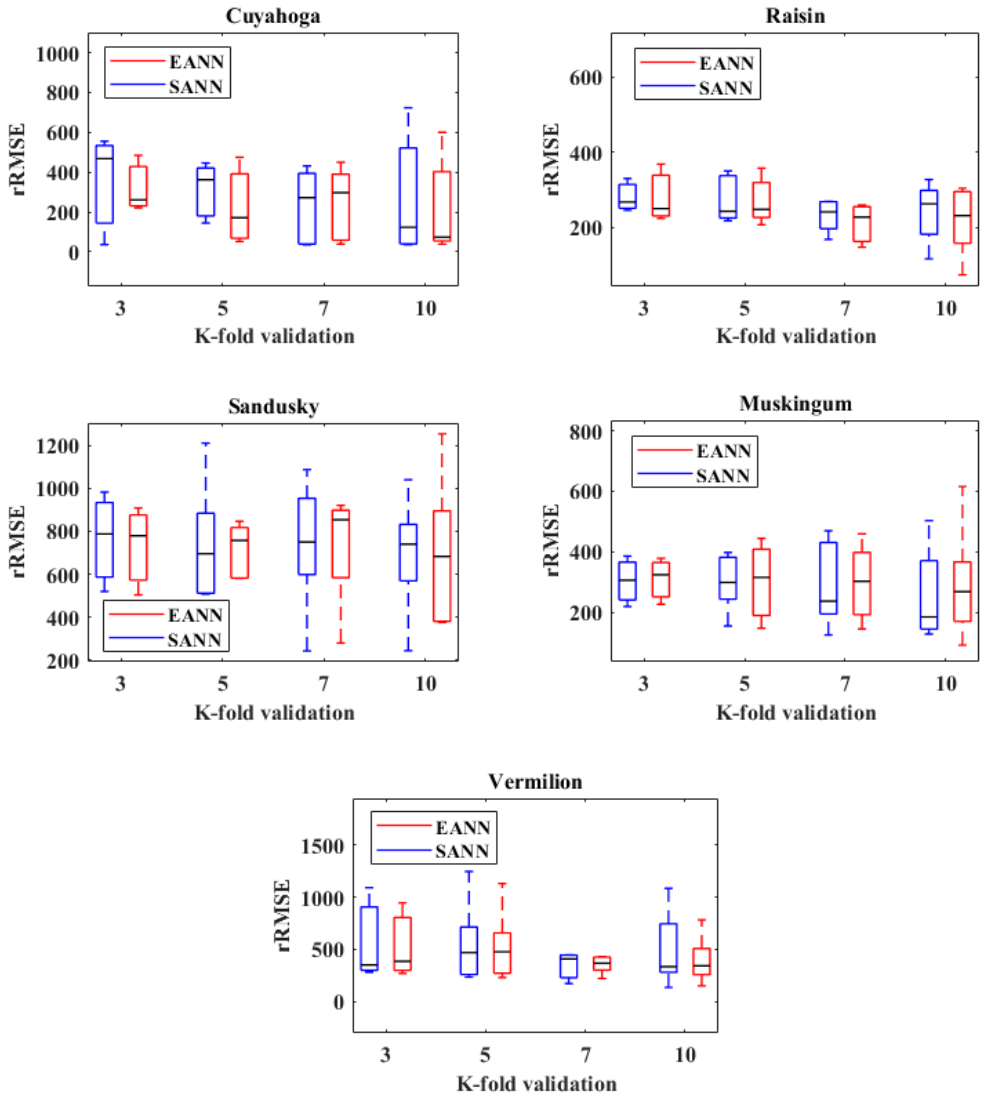


Figure 7. Box plots of the *rRMSE* of the SANN and EANN model results for the five stations; data were derived by examining different values of k in the cross-validation method.

4.2. Single ANN-CCA and Ensemble ANN-CCA

In order to identify whether the variable of time affects the ANN predictions, the CCA based SANN and EANN model results were examined by changing the number of folds in the cross-validation. In conducting the CCA, the variables of discharge and time were used to estimate loads for the five river basins. The combination with the variables can help the performance of ANNs. Table 3 shows the *rRMSE* and *rBIAS* indices obtained from the SANN-CCA and EANN-CCA models for assessment of the model performance. This table also includes the results for 3-fold, 5-fold, 7-fold, and 10-fold cross validations. The model performance based on the two indices seemed to be enhanced when the number of folds increased. Overall, the CCA based models outperformed the SANN and EANN models. The EANN-CCA model showed a better performance according to the *rRMSE* and *rBIAS* indices than the SANN-CCA model for the 10-fold cross-validation. These results indicate that the use of the ANN models based on the combination of the variables in the CCA space can improve the performance relative to the ANN models with one variable. Figure 8 presents the results derived by using different folds of cross-validations for the SANN-CCA and EANN-CCA models in the five river basins. From this figure, we can observe that the performance of ANNs seemed to be improved for all sites except for the Sandusky basin.

Table 3. K-fold cross-validation results based on the SANN-CCA and EANN-CCA models for the five stations.

Stations		Single ANN-CCA				Ensemble ANN-CCA			
		3-fold	5-fold	7-fold	10-fold	3-fold	5-fold	7-fold	10-fold
Cuyahoga	***rRMSE* (%)**	305.525	264.872	259.891	219.974	302.339	238.481	194.688	166.095
	***rBIAS* (%)**	−11.488	−11.576	−11.929	−11.458	−11.519	−10.830	−10.706	−9.798
Rasin	***rRMSE* (%)**	279.883	271.493	269.687	241.273	273.169	262.427	258.720	237.591
	***rBIAS* (%)**	−31.159	−31.469	−30.401	−30.166	−30.246	−30.679	−30.682	−29.921
Sandusky	***rRMSE* (%)**	716.695	696.797	665.316	665.208	719.670	692.090	666.620	662.612
	***rBIAS* (%)**	−118.735	−114.608	−116.932	−111.576	−117.194	−115.492	−113.638	−114.878
Muskingum	***rRMSE* (%)**	300.587	285.457	282.776	276.290	297.780	287.342	278.656	273.975
	***rBIAS* (%)**	−35.664	−33.865	−33.832	−33.378	−34.948	−35.097	−32.854	−34.721
Vermilion	***rRMSE* (%)**	560.439	495.932	449.676	441.531	462.424	457.477	443.022	397.644
	***rBIAS* (%)**	−129.024	−110.639	−109.929	−101.932	−105.677	−102.146	−100.717	−99.477

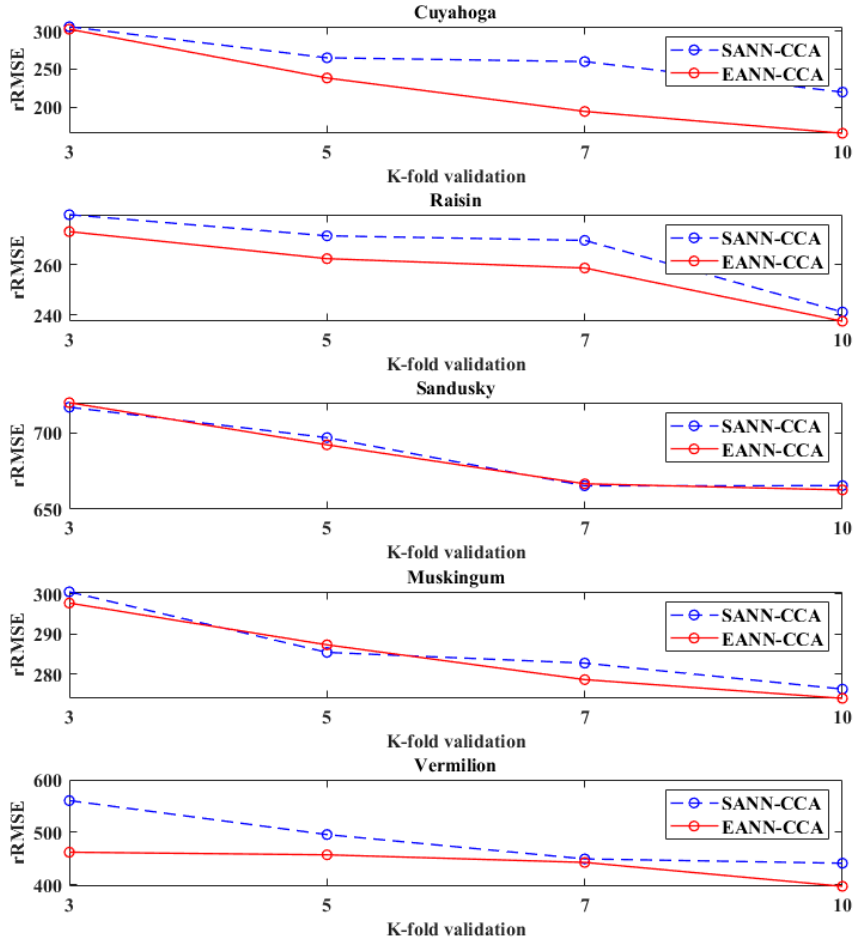


Figure 8. *rRMSE* of the SANN-CCA and EANN-CCA model results for the five stations based on different values of k in the cross-validation method.

The model performance based on the *rRMSE* index for load estimations was also analyzed with box plots for various folds in the cross-validation. Figure 9 shows the box plots according to the *rRMSE* index of the five river basins. As in Figure 7, the left box plots with the blue color represent the results of the SANN-CCA model, whereas the right box plots with the red color represent the results of the EANN-CCA model in Figure 9. This figure shows that there was a decreasing trend in the index for the ANN models when the number of folds increased. The results indicated that the variables that were significantly related to nutrient concentrations could improve the ANN predictions, and the model

with the 10-fold cross-validation seemed to provide a better performance. This examination shows that the proposed model can produce good estimates by adding in the important variables correlated with water quality, and it can be used for improved load estimation applications.

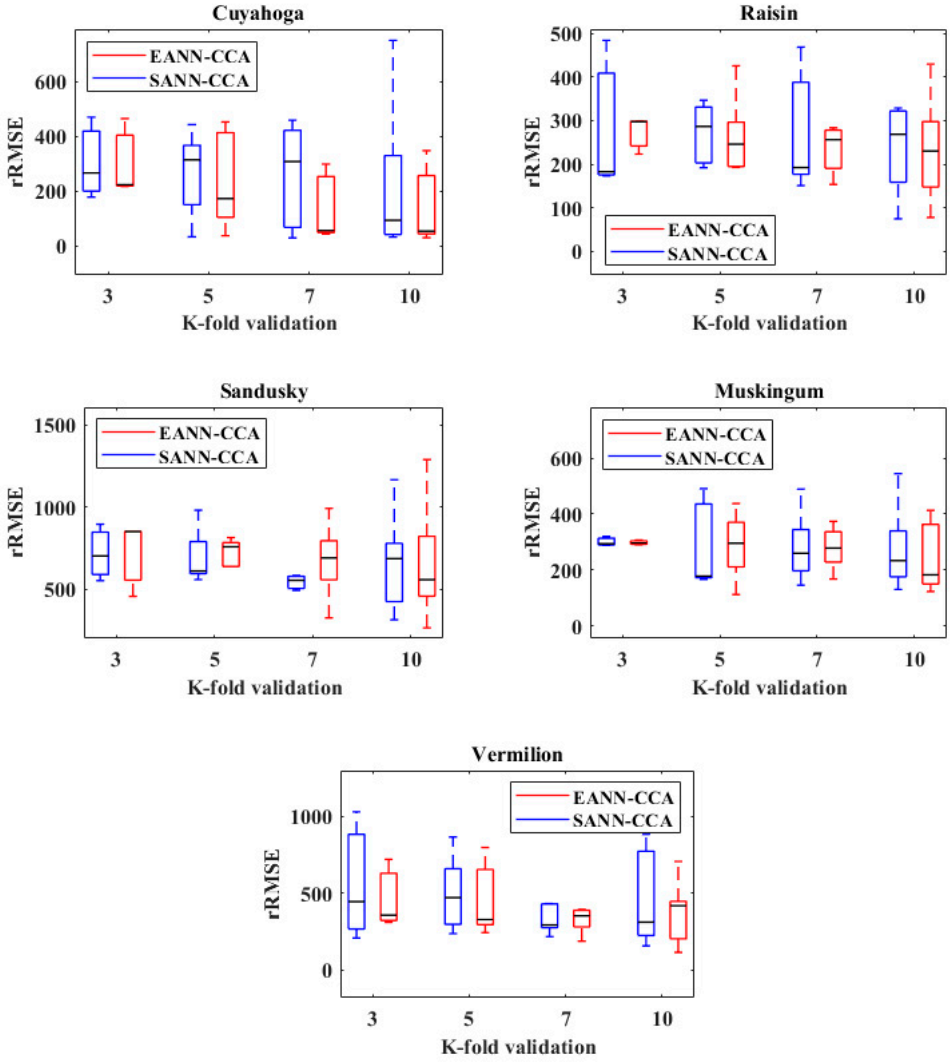


Figure 9. Box plots of the *rRMSE* of the SANN-CCA and EANN-CCA model results for the five stations; data were derived by examining different values of k in the cross-validation method.

5. Conclusions

A methodology based on ANN models to achieve nitrate load estimations was examined for river basin nutrient concentration assessments in this study. The SANN and EANN models were built for five river basins in the Midwest US, and these basins included the Cuyahoga, Raisin, Sandusky, Muskingum, and Vermilion in the states of Michigan and Ohio. To improve the model performance, CCA was also used with a combination of variables such as the discharge and time for load estimations. The proposed models were assessed by using 3-fold, 5-fold, 7-fold, and 10-fold cross validations. Two statistical indices, namely, the *rRMSE* and *rBIAS* were applied for the validation of the proposed models.

Application of the ANN based models in the study region showed that better performances can be obtained when the number of folds in the cross-validation is high. The best *rRMSE* and *rBIAS* indices seemed to be produced with the 10-fold validation for the five river basins and for the three river basins, respectively. Moreover, the EANN model tended to produce better estimations for loads than the SANN model. The box plots for the *rRMSE* index based on the two models were also analyzed, and the plots indicated that an increasing number of folds seemed to provide a better performance.

However, the station data for the Raisin and Vermilion basins only tended to show improvements up to the 7-fold validation and results were nearly constant at the 10-fold validation. The station data for Cuyahoga and Sandusky basins seemed to show that the results of the 5-fold validation were better than the results of the 7-fold validation in the ensemble model. Overall, the use of a k value of 10 can be recommended to estimate loads when other basins in a different region are investigated for load estimations. This is because the k value of 10 steadily provides good estimations derived from the single and ensemble models proposed in the present study within the study regions.

Moreover, the CCA based ANN models were proposed for the achievement of better estimations. The SANN-CCA and EANN-CCA models were applied to obtain load estimations, and the corresponding statistical indices were compared for different folds of cross-validation. Compared to the SANN and EANN models, the CCA based ANN models led to a better performance in the measures including the *rRMSE* and *rBIAS*. The 10-fold cross-validation tended to provide a better estimation than the other fold cross-validations for the five river basins. The EANN-CCA model improved the performance in terms of both the *rRMSE* and *rBIAS* indices compared to the SANN-CCA model. The box plots for the *rRMSE* index were also examined, and the index seemed to show a decreasing trend when the number of folds increased for the studied regions.

The ANN based models were analyzed in this study for load estimations to determine a better estimation model. Ultimately, the CCA based ANN models with significant variables related to nitrate loads were proposed to improve the model performance. Based on this work, the models can be applied for load estimations, especially to deal with missing monitoring data of interest when investigating nitrate loads in a river basin. Additionally, other common methods such as EGRET can be applied to achieve a better load estimation technique.

Author Contributions: Conceptualization, D.P. and K.J.; methodology, K.J., D.P., and D.-H.B.; investigation, K.J., D.P., M.-J.U., S.K., and S.J.; funding acquisition, D.P. and D.-H.B.; writing—review and editing, K.J., D.P., M.-J.U., S.K., and S.J. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Konkuk University, grant number 2017-A019-0567.

Acknowledgments: This paper was supported by Konkuk University in 2017.

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

References

1. Markus, M.; Demissie, M.; Short, M.B.; Verma, S.; Cooke, R.A. Sensitivity analysis of annual nitrate loads and the corresponding trends in the lower Illinois River. *J. Hydrol. Eng.* **2013**, *19*, 533–543. [CrossRef]
2. Pellerin, B.A.; Bergamaschi, B.A.; Gilliom, R.J.; Crawford, C.G.; Saraceno, J.; Paul Frederick, C.; Downing, B.D.; Murphy, J.C. Mississippi River nitrate loads from high frequency sensor measurements and regression-based load estimation. *Environ. Sci. Technol.* **2014**, *48*, 12612–12619. [CrossRef] [PubMed]
3. Verma, S.; Markus, M.; Cooke, R.A. Development of error correction techniques for nitrate-N load estimation methods. *J. Hydrol.* **2012**, *432*, 12–25. [CrossRef]
4. Cohn, T.A. Estimating contaminant loads in rivers: An application of adjusted maximum likelihood to type 1 censored data. *Water Resour. Res.* **2005**, *41*. [CrossRef]
5. Robertson, D.M.; Roerish, E.D. Influence of various water quality sampling strategies on load estimates for small streams. *Water Resour. Res.* **1999**, *35*, 3747–3759. [CrossRef]
6. Robertson, D.M.; Saad, D.A. Nutrient inputs to the Laurentian Great Lakes by source and watershed estimated using SPARROW watershed models 1. *J. Am. Water Resour.* **2011**, *47*, 1011–1033. [CrossRef]
7. Runkel, R.L.; Crawford, C.G.; Cohn, T.A.; US GS, L.E. A FORTRAN program for estimating constituent loads in streams and rivers. *US Geol. Surv.* **2004**. [CrossRef]
8. Cohn, T.A.; Delong, L.L.; Gilroy, E.J.; Hirsch, R.M.; Wells, D.K. Estimating constituent loads. *Water Resour. Res.* **1989**, *25*, 937–942. [CrossRef]
9. Hirsch, R.M.; Moyer, D.L.; Archfield, S.A. Weighted regressions on time, discharge, and season (WRTDS), with an application to Chesapeake Bay river inputs 1. *J. Am. Water Resour.* **2010**, *46*, 857–880. [CrossRef]

10. Moyer, D.; Hirsch, R.M.; Hyer, K. *Comparison of Two Regression-Based Approaches for Determining Nutrient and Sediment Fluxes and Trends in the Chesapeake Bay Watershed*; Scientific Investigations Report 2012–5244; United States Geological Survey: Reston, VA, USA, 2012; 118p.
11. Hirsch, R.M.; De Cicco, L.A. User guide to Exploration and Graphics for RivEr Trends (EGRET) and dataRetrieval: R packages for hydrologic data. *US Geol. Surv.* **2015**. [CrossRef]
12. Markus, M.; Tsai, C.W.S.; Demissie, M. Uncertainty of weekly nitrate-nitrogen forecasts using artificial neural networks. *J. Environ. Eng.* **2003**, *129*, 267–274. [CrossRef]
13. Markus, M.; Hejazi, M.I.; Bajcsy, P.; Giustolisi, O.; Savic, D.A. Prediction of weekly nitrate-N fluctuations in a small agricultural watershed in Illinois. *J. Hydroinform.* **2010**, *12*, 251–261. [CrossRef]
14. 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]
15. Singh, K.P.; Basant, A.; Malik, A.; Jain, G. Artificial neural network modeling of the river water quality—A case study. *Ecol. Model.* **2009**, *220*, 888–895. [CrossRef]
16. Chen, D.; Lu, J.; Shen, Y. Artificial neural network modelling of concentrations of nitrogen, phosphorus and dissolved oxygen in a non-point source polluted river in Zhejiang Province, southeast China. *Hydrol. Process.* **2010**, *24*, 290–299. [CrossRef]
17. He, B.; Oki, T.; Sun, F.; 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]
18. Jeong, D.I.; Kim, Y.O. Rainfall-runoff models using artificial neural networks for ensemble streamflow prediction. *Hydrol. Process.* **2005**, *19*, 3819–3835. [CrossRef]
19. Ouarda, T.B.M.J.; Shu, C. Regional low-flow frequency analysis using single and ensemble artificial neural networks. *Water Resour. Res.* **2009**, *45*. [CrossRef]
20. Shu, C.; Ouarda, T.B.M.J. Flood frequency analysis at ungauged sites using artificial neural networks in canonical correlation analysis physiographic space. *Water Resour. Res.* **2007**, *43*. [CrossRef]
21. Araghinejad, S.; Azmi, M.; Kholghi, M. Application of artificial neural network ensembles in probabilistic hydrological forecasting. *J. Hydrol.* **2011**, *407*, 94–104. [CrossRef]
22. El-Shafie, A.; Najah, A.; Alsulami, H.M.; Jahanbani, H. Optimized neural network prediction model for potential evapotranspiration utilizing ensemble procedure. *Water Resour. Manag.* **2014**, *28*, 947–967. [CrossRef]
23. Alobaidi, M.H.; Marpu, P.R.; Ouarda, T.B.M.J.; Chebana, F. Regional frequency analysis at ungauged sites using a two-stage resampling generalized ensemble framework. *Adv. Water Resour.* **2015**, *84*, 103–111. [CrossRef]
24. Chokmani, K.; Ouarda, T.B.M.J. Physiographical space-based kriging for regional flood frequency estimation at ungauged sites. *Water Resour. Res.* **2004**, *40*. [CrossRef]
25. Leclerc, M.; Ouarda, T.B.M.J. Non-stationary regional flood frequency analysis at ungauged sites. *J. Hydrol.* **2007**, *343*, 254–265. [CrossRef]
26. Raman, H.; Sunilkumar, N. Multivariate modelling of water resources time series using artificial neural networks. *Hydrol. Sci. J.* **1995**, *40*, 145–163. [CrossRef]
27. Kan, G.; Yao, C.; Li, Q.; Li, Z.; Yu, Z.; Ding, L.; He, X.; Liang, K. Improving event-based rainfall-runoff simulation using an ensemble artificial neural network based hybrid data-driven model. *Stoch. Environ. Res. Risk A.* **2015**, *29*, 1345–1370. [CrossRef]
28. Huang, J.; Gao, J. An ensemble simulation approach for artificial neural network: An example from chlorophyll a simulation in Lake Poyang, China. *Ecol. Inform.* **2017**, *37*, 52–58. [CrossRef]
29. McClelland, J.L.; Rumelhart, D.E.; Group, P.R. *Parallel Distributed Processing*; MIT Press: Cambridge, UK, 1987.
30. Demuth, H.B.; Beale, M.H.; De Jess, O.; Hagan, M.T. *Neural Network Design*; Martin Hagan, Oklahoma State University: Stillwater, OK, USA, 2014.
31. Cannon, A.J.; Whitfield, P.H. Downscaling recent streamflow conditions in British Columbia, Canada using ensemble neural network models. *J. Hydrol.* **2002**, *259*, 136–151. [CrossRef]
32. Dietterich, T.G. Machine-learning research. *AI Mag.* **1997**, *18*, 97–134.
33. Shu, C.; Burn, D.H. Artificial neural network ensembles and their application in pooled flood frequency analysis. *Water Resour. Res.* **2004**, *40*. [CrossRef]

34. Ouarda, T.B.M.J.; Girard, C.; Cavadias, G.S.; Bobee, B. Regional flood frequency estimation with canonical correlation analysis. *J. Hydrol.* **2001**, *254*, 157–173. [CrossRef]
35. Miller, R.G. A trustworthy jackknife. *Ann. Math. Stat.* **1964**, *35*, 1594–1605. [CrossRef]
36. Shao, J.; Tu, D. *The Jackknife and Bootstrap*; Springer Science & Business Media: Berlin, Germany, 2012.
37. Rodriguez, J.D.; Perez, A.; Lozano, J.A. Sensitivity analysis of k-fold cross validation in prediction error estimation. *IEEE T. Pattern Anal.* **2009**, *32*, 569–575. [CrossRef] [PubMed]