


Article

Application of Artificial Neural Networks to Rainfall Forecasting in the Geum River Basin, Korea

Jeongwoo Lee *, Chul-Gyum Kim, Jeong Eun Lee, Nam Won Kim and Hyeonjun Kim 

Department of Land, Water and Environment Research, Korea Institute of Civil Engineering and Building Technology (KICT), 283 Goyang-daero, Ilsanseo-Gu, Goyang-Si 10223, Korea; cgkim@kict.re.kr (C.-G.K.); jeus22@kict.re.kr (J.E.L.); nwkim@kict.re.kr (N.W.K.); hjkim@kict.re.kr (H.K.)

* Correspondence: ljw2961@kict.re.kr; Tel.: +82-2-910-0529; Fax: +82-2-910-0251

Received: 24 August 2018; Accepted: 12 October 2018; Published: 14 October 2018



Abstract: This study develops a late spring-early summer rainfall forecasting model using an artificial neural network (ANN) for the Geum River Basin in South Korea. After identifying the lagged correlation between climate indices and the rainfall amount in May and June, 11 significant input variables were selected for the preliminary ANN structure. From quantification of the relative importance of the input variables, the lagged climate indices of East Atlantic Pattern (EA), North Atlantic Oscillation (NAO), Pacific Decadal Oscillation (PDO), East Pacific/North Pacific Oscillation (EP/NP), and Tropical Northern Atlantic Index (TNA) were identified as significant predictors and were used to construct a much simpler ANN model. The final best ANN model, with five input variables, showed acceptable performance with relative root mean square errors of 25.84%, 32.72%, and 34.75% for training, validation, and testing data sets, respectively. The hit score, which is the number of hit years divided by the total number of years, was more than 60%, which indicates that the ANN model successfully predicts rainfall in the study area. The developed ANN model, incorporated with lagged global climate indices, could allow for more timely and flexible management of water resources and better preparation against potential droughts in the study region.

Keywords: artificial neural network; rainfall forecasting; climate indices; relative importance of input variables

1. Introduction

Rainfall prediction is of great importance to prevent flooding and manage water resources, saving lives and property and securing economic activities. Insufficient rainfall has a strong adverse influence on water supply, water quality, and the aquatic ecosystem. If possible, the ability to forecast rainfall several months in advance would enable effective water use. Therefore, accurate rainfall forecasting is a challenging task in operational water resources management [1].

Several methods are available for rainfall forecasting, such as numerical weather prediction (NWP) models, statistical methods, and machine learning techniques. Among these, machine learning techniques, such as artificial neural network (ANN), k-nearest neighbor, support vector machine, and random forest model, are more suitable for rainfall forecasting because physical processes affecting rainfall occurrence are highly complex and non-linear [2]. The ANN is a form of machine learning technique that has been widely used in rainfall prediction given its ability to identify highly complex non-linear relationships between input and output variables without the need to understand the nature of the physical processes.

Various studies on rainfall prediction have been published using ANNs. Bodri and Cermak [3] developed the ANN model for predicting the time series of monthly precipitation for two Czech meteorological stations using actual precipitation data in the previous months of the current year and a given month in the two previous years. Bodri and Cermak [4] further predicted the next month's

precipitation for six Czech and four Hungarian meteorological stations using ANNs. Wu et al. [5] developed ANN models to forecast monsoon rainfall over the Yangtze delta region in China 1, 5, and 10 years in advance using only historical data of the total amount of summer rainfall. Philip and Joseph [6] predicted monthly rainfall in Kerala State, the southern part of the Indian Peninsula, using the adaptive basis function neural network with historical data of the previous four years and three months of rainfall. Chakraverty and Gupta [7] predicted southwestern monsoon rainfall over India six years in advance using only historical data as inputs for ANN models. Chattopadhyay and Chattopadhyay [8] generated forecasts using ANNs for Indian average summer monsoon rainfall with the previous year's rainfall amount in the months of June, July, and August. Gholizadeh and Darand [9] forecasted monthly precipitation for Tehran, Iran a year in advance using ANNs with a genetic algorithm. Aksoy and Dahamsheh [10] used feed-forward back-propagation (FFBP), radial basis function (RBF), and generalized regression-type ANNs to forecast precipitation one month ahead. Bilgili and Sahin [11] applied ANNs to predict the long-term monthly temperature and rainfall for stations in Turkey with geographical variables and neighboring measuring stations data. The ANN models in these studies used the characteristics of previously observed rainfall data, which were rainfall-rainfall models. Those studies revealed that ANNs are a useful tool to forecast rainfall amounts on various time scales in advance in arid to humid areas.

Global teleconnections are statistical associations between climate variables separated by large distances [12]. Teleconnections are the result of large-scale dynamics between the ocean and atmosphere, linking different regional climates into a unified global climatic system [13,14]. Many attempts have been made to forecast precipitation using various climate indices data representing teleconnection patterns. Silverman and Dracup [15] applied ANNs to forecast the total water year precipitation of California's seven zones using the monthly 700-hPa teleconnection indices and El Niño Southern Oscillation (ENSO) indicators. They emphasized the possibility of long-term precipitation predictions using ANNs and large scale climate variables. Kumar et al. [16] developed summer monsoon rainfall forecasting ANN models with current and lagged climate indices of ENSO, Indian Ocean Oscillation, and local ocean-land temperature contrast as inputs. Iseri et al. [17] developed ANN models for August rainfall forecasting in Fukuoka, Japan using sea surface temperature anomalies in the Pacific Ocean and the lagged climate indices of the Southern Oscillation Index (SOI), Pacific Decadal Oscillation (PDO), and North Pacific Index (NPI). Hartman et al. [18] predicted summer rainfall in the Yangtze River basin using a set of climate indices including the SOI, the East Atlantic/Western Russia (EA/WR) pattern, the Scandinavia (SCA) pattern, the Polar/Eurasia (POL) pattern, and several indices calculated from sea surface temperatures (SST), sea level pressures (SLP), and snow data. Yuan et al. [12] predicted summer precipitation in the source region of the Yellow River, China using ANN models with inputs of North Atlantic Oscillation (NAO), West Pacific (WP), ENSO, and POL patterns. Most studies showed the applicability of the combined use of ANNs and large-scale climate teleconnections for regional summer rainfall forecasting in monsoon areas.

Abbot and Marohasy [19] applied ANNs to forecast monthly and seasonal rainfall three months in advance in Queensland, Australia by inputting climate indices such as SOI, PDO, and Niño 3.4, as well as historical rainfall and temperature data. Abbot and Marohasy [20] further applied ANNs to forecast the one-year-ahead monthly rainfall for locations in the Murray-Darling basin, Australia. They used the SOI, Dipole Mode Index (DMI), Niño 4, Niño 3.4, Niño 3, Niño 1.2, and the Inter-Decadal Pacific Oscillation (IPO) climate indices as inputs. Badr et al. [21] employed ANNs to forecast summer rainfall anomalies in the Sahel region of Africa using springtime surface air temperature (SAT) anomalies and sea surface temperature (SST). Rasel et al. [22] predicted spring rainfall for South Australia using ANNs with the lagged climate indices of the ENSO-DMI-SAM (Southern Annual Mode). Badr et al. [21] and Rasel et al. [22] showed the predictive superiority of ANNs by comparing with alternative statistical methods. As these comprehensive studies of the combined use of ANNs and teleconnection features have been completed to improve the predictability of rainfall by considering other local meteorological parameters together with large-scale climate indices, the predictive accuracy of ANN models has been shown to be superior to that

of different forecasting methods, and have enabled a longer lead-time. As reviewed in previous research, a variety of ANN-based regional rainfall forecast models using teleconnection climate indices have been created extensively around the world, but no models exist for South Korea.

Extreme droughts are occurring more frequently, resulting in a serious reduction in the water supply in the central region of South Korea from 2013 to 2016 and in the southern region in 2017. As a result of droughts, a large number of regions have experienced inaccessibility to safe water. Droughts have caused damage, such as restrictive water rationing, restrictions of instream flow, and reduced agricultural water supply for dams located in the Han River and Geum River Basin [23]. Drought damage continued until the following year due to insufficient rainfall during the summer season. In 2017, the cumulative precipitation in spring (March to May) was low, at 50% compared with the historically normal level of the season, 30% of a normal year in May, and 38% of a normal year in June. The rainfall amount in late spring and early summer, in which a large amount of irrigation water is required for farming, greatly impacts rice production. Therefore, it is essential to forecast seasonal rainfall in advance for more timely and efficient management of water resources. Appropriate rainfall forecasts can be achieved using global teleconnection patterns.

In this study, a simple ANN model with inputs from several lagged climate indices is developed to predict late spring and early summer rainfall during May and June in the Geum River Basin. A preliminary ANN structure is constructed after identifying the lagged correlation between climate indices and rainfall. Then, a compact optimal ANN model is established through a process of determining the contribution of input variables, and the model performance is evaluated.

2. Data

2.1. Rainfall Data

The Geum River Basin is located in the west central region of South Korea (Figure 1), having an area of 9912.15 km² and a main channel length of 360.70 km. The areal average monthly rainfall data for the period from January 1966 to December 2017 for the Geum River Basin were obtained from the Water Resources Management Information System (WAMIS) and the hydrological survey reports of the Geum River Flood Control Office (GRFCO) in Korea. Figure 1 shows the study area and the locations of the rainfall stations. This study attempts to predict the late spring-early summer rainfall, which presents the total amount of rainfall in May and June (M-J).

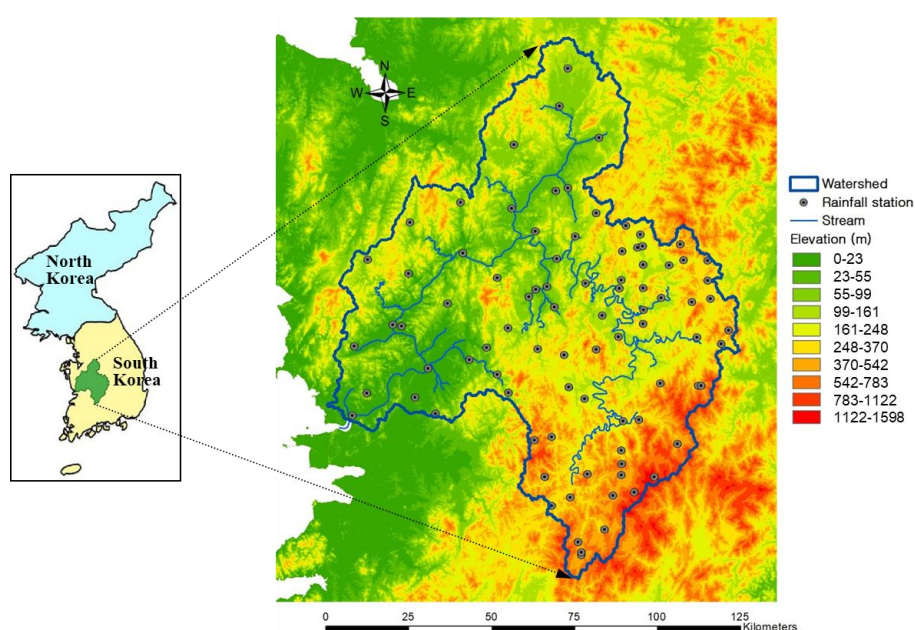


Figure 1. Map of the study area and rainfall stations.

2.2. Climate Indices

The global climate indices and historical rainfall data were used as predictors of the ANN-based forecasting model in this study. The monthly values of climate indices were collected from the Climate Prediction Center under National Oceanic and Atmospheric Administration (<https://www.esrl.noaa.gov/psd/data/climateindices/list/>). To select the predictor variables and identify the months that could be used as input to the ANNs, cross-correlation analyses were carried out for delayed climate indices. Many previous studies have validated that lagged climate variables are good predictors [19–22,24] for ANN models. The highly correlated 10 climate indices, including Arctic Oscillation (AO), East Pacific/North Pacific Oscillation (EP/NP), East Atlantic Pattern (EA), North Atlantic Oscillation (NAO), North Tropical Atlantic (NTA), Tropical Northern Atlantic Index (TNA), Western Pacific Index (WP), Pacific Decadal Oscillation (PDO), Southern Oscillation (SOI), and Sea Level Pressure of Darwin (SLP_D), were selected as inputs for the ANN models. The October rainfall amount of the previous year was also found to be highly correlated with a predictand. Table 1 provides a list of inputs with time lags and cross-correlation values that were used for the forecasting model. A maximum negative correlation of -0.375 was achieved for NAO, and the month with the highest correlation was the month of December of the previous year. A positive maximum correlation of 0.326 is obtained for the six-month-lagged EP/NP. For different climate indices, different months had a significant correlation with the M-J rainfall.

Table 1. Selected climate indices and lagged months with the highest cross-correlation value.

Climate Index	Description	Time Lag (months)	Correlation Coefficient
AO	The first leading mode from the Empirical Orthogonal Function (EOF) analysis of monthly mean height anomalies at 1000-hPa	10	0.196
EP/NP	A Spring–Summer–Fall pattern with three main anomaly centers	6	0.326
EA	The second prominent mode of low-frequency variability over the North Atlantic, and appears as a leading mode in all months	7	-0.294
NAO	One of the most prominent teleconnection patterns in all seasons is the North Atlantic Oscillation	5	-0.375
NTA	The timeseries of SST anomalies averaged over 60W to 20W, 6N to 18N and 20W to 10W, 6N to 10N	4	0.255
TNA	Anomaly of the average of the monthly SST from 5.5N to 23.5N and 15W to 57.5W	10	0.225
WP	A primary mode of low-frequency variability over the North Pacific in all months	4	-0.202
PDO	Pacific Decadal Oscillation is the leading principal component (PC) of monthly SST anomalies in the North Pacific Ocean	11	0.200
SOI	The development and intensity of El Niño or La Niña events in the Pacific Ocean, calculated using the pressure differences between Tahiti and Darwin	10	0.183
SLP_D	Sea Level Pressure at Darwin	10	-0.209
GEUM	Areal average monthly precipitation for the Geum River basin	7	-0.195

3. Methodology

3.1. Artificial Neural Network

ANN is a data-driven mathematical model that was developed to imitate the structure of a human brain neural network and has been widely applied to solve problems such as prediction and discrimination. The ANN is based on the perceptron—a compound word combining the role of neurons and recognition. The perceptron consists of one input layer and an output layer, and each layer contains nodes for data operations corresponding to a cell body. By adding a hidden layer and nodes inside the input layer and the output layer of the perceptron, the network expands to a multilayer perceptron structure. In general, an artificial neural network refers to a multilayer perceptron structure.

The three-layered feed-forward neural network has been widely used in hydrologic forecasting models [25–28]. The input data in the input layer is transferred to each neuron in the hidden layer through a linear sum operation, and the result of inputting the linear sum to the activation function is the result of the hidden layer neuron. The same procedure is followed from the hidden layer to

the output layer. A neural network with three layers can be expressed mathematically by a linear combination of the transferred input values as:

$$\hat{y}_k = f_0 \left[\sum_{j=1}^n w_{kj} f_h \left(\sum_{i=1}^m w_{ji} x_i + w_{jb} \right) + w_{kb} \right] \quad (1)$$

where \hat{y}_k is the forecasted k th output value, f_0 is the activation function for the output neuron, n is the number of output neurons, w_{kj} is the weight connecting the j th neuron in the hidden layer and k th neuron in the output layer, f_h is the activation function for the hidden neuron, m is the number of hidden neurons, w_{ji} is the weight connecting the i th neuron in the input layer and j th neuron in the hidden layer, x_i is the i th input variable, w_{jb} is the bias for the j th hidden neuron, and w_{kb} is the bias for the k th output neuron [29,30].

Learning the ANN model is a training process entailing the search for the optimal weight vector used in Equation (1). In this study, the weights that minimize the sum of errors of the network in Equation (2) were calculated using the back-propagation algorithm [31]:

$$E = \sum_{p=1}^P E_p = \sum_{p=1}^P \sum_{k=1}^n (y_{pk} - \hat{y}_{pk})^2 \quad (2)$$

where E is the error for all input patterns and E_p is the error based on the squared difference between the true outputs y_{pk} and the forecasted outputs \hat{y}_{pk} for pattern p [32].

3.2. ANN Model Development

The identification of significant input variables and the optimization of the network structure are important steps in building an optimal ANN model. As described in Table 1, the input variables were determined by investigating cross-correlations between the lagged climate indices and the total rainfall in May and June. The number of input nodes is equivalent to the number of input variables. The number of hidden layers was selected as one, and the number of nodes in the hidden layer was experimentally determined by trial and error with a learning rate of 0.01 and a hyperbolic tangent sigmoid activation function for the hidden layer. The number of hidden nodes was changed for the trial networks from 2 to 10, and a decision was made regarding the relative root mean square error (RRMSE) of the forecasted M-J rainfall against the true observed rainfall during the training and validation stages. Since the artificial neural network randomly sets the initial weight value at the beginning of the training, a different neural network model is created for each training process, yielding different performance. Therefore, the optimal prediction model was selected based on the average accuracy obtained by repeating the ANN model generation process 100 times.

In this study, the K-fold cross-validation (CV) procedure [33,34], which is one of the most widely used re-sampling methods, was used to evaluate the model performance. The reasons for using the K-fold CV method were to allow the selection of the best model architecture and to avoid overfitting the specific training data set. Data for the ANN model development were firstly divided into five equally sized subsets (10 patterns per each subset), avoiding duplication. Four subsets were used for training and validation, and the one remaining subset was used for testing. Therefore, the four-fold CV procedure on the calibration data (i.e., training and validation data) was performed to determine which model is best. After the training and validation were repeated four times using the four subsets used in the cross-validation process, the average of the RRMSEs on each fold was used to obtain an aggregate measure. In order to avoid overtraining, an early stopping technique was also applied with continuous monitoring of the errors in both the training set and the validation set during training. The number of hidden nodes that showed the smallest CV error was chosen for the M-J rainfall forecasts. After the selection of the optimum number of hidden nodes, the ANN model was trained with the aggregate

data of the training and validation sets, and the trained model was finally tested using the unseen data set to evaluate the model performance.

In order to quantify the contributions of each input variable to the prediction of the output variable in ANNs, we applied Garson's connection weight method [35] and Olden's connection weight method [36]. Garson's method uses the absolute values of the connection weights between nodes to determine the relative importance (RI) of each input variable. The computation procedure of Garson's algorithm is as follows:

- (1) The products P_{ij} are obtained by multiplying the input-hidden connection weight and the hidden-output connection weight for each hidden neuron i and repeating this for each input neuron j ;
- (2) Scaled products Q_{ij} are obtained by dividing the absolute values of P_{ij} by the sum for all input variables $\sum_{j=1}^J \text{abs}(P_{ij})$ for each hidden neuron i ;
- (3) The product S_j is obtained by summing Q_{ij} for each input neuron; and
- (4) Relative importance values RI_j (%) are obtained by dividing S_j by the sum for all the input variables $\sum_{j=1}^J S_j$ and expressing the figure as a percentage.

Olden's method calculates the product of the raw input-hidden and hidden-output connection weights between each input node and output node, and sums the products across all hidden nodes [36].

After determining the contribution of the input variables, a simpler ANN model was constructed. The best ANN model was obtained from the four-fold CV procedure, as described previously, and the model performance was evaluated for the unseen testing data set.

4. Results and Discussion

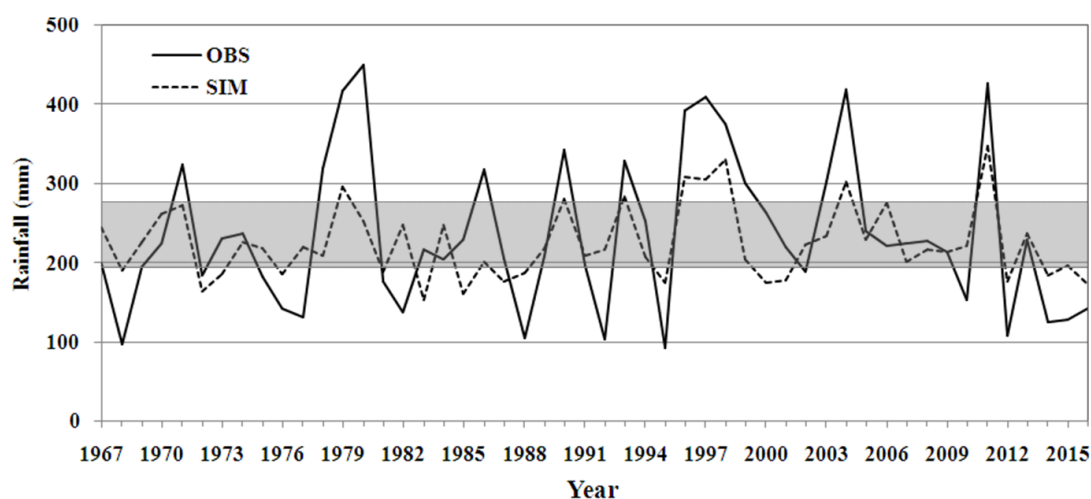
4.1. Preliminary ANN Model for Rainfall Forecasting

The optimal ANN model structure with 11 input variables was determined using the four-fold CV procedure by varying the number of hidden neurons from 2 to 10. For each hidden neuron, four iterations of the training and validation were performed. The average RRMSE between the actual rainfall and the predicted rainfall measured using the ANN models is presented in Table 2. The results show that the network performance with different numbers of hidden neurons was not significantly different. The results also show that a large number of hidden neurons did not always lead to better performance. For the training part, the RRMSE values ranged from 29.75% to 31.54% (RMSE: 68.76 mm to 72.90 mm), and for the validation part, the RRMSE values ranged from 35.52% to 35.73% (RMSE: 82.04 mm to 82.50 mm), and for testing part, the RRMSEs ranged from 34.28% to 34.88% (RMSE: 85.82 mm to 87.15 mm). The Pearson correlation coefficient (CC) values ranged from 0.731 to 0.757 for training, from 0.453 to 0.460 for validation, and from 0.723 to 0.743 for testing. The resultant accuracy was within the acceptable range. Based on the minimum error in CV, the optimal ANN structure with four hidden neurons was considered the best, which is denoted as ANN (11,4,1) hereafter. The selected ANN (11,4,1) was trained using the whole data set of the training and validation parts. The weights for this trained structure were saved and the network was evaluated for the testing part. The RRMSE values for the testing part for other structures that were not selected as the best are described in Table 2 for comparison. As evident from the results shown in Table 2, the RRMSE values for the test data were acceptable, with little difference from the training and validation results.

Table 2. Preliminary artificial neural network (ANN) model performance for training, validation, and testing parts.

Number of Hidden Nodes	RRMSE (%)			CC		
	Training	Validation	Testing	Training	Validation	Testing
2	29.75	35.60	34.46	0.757	0.459	0.723
3	30.51	35.62	34.51	0.747	0.457	0.735
4	30.43	35.52	34.41	0.742	0.460	0.733
5	30.46	35.67	34.28	0.740	0.455	0.743
6	30.89	35.68	34.49	0.736	0.454	0.740
7	31.32	35.73	34.75	0.733	0.455	0.743
8	30.95	35.72	34.47	0.734	0.453	0.743
9	31.54	35.71	34.88	0.731	0.454	0.743
10	31.23	35.72	34.66	0.731	0.453	0.742

The observed rainfall data were divided into three categories based on $\mu + 0.43\sigma$, under the assumption of a normal distribution with a mean of μ and standard deviation of σ . Then, the observed and the predicted rainfall were classified into one of the three categories of below, near, and above-normal rainfall conditions. Figure 2 compares the ANN (11,4,1)-forecasted rainfall with observed rainfall for May-June. The shaded area in the figure means near-normal rainfall. Even though there are some deviations from the observed rainfall, the model results show reasonable accuracy. However, the model has under-forecasted for high rainfall years (1979, 1980, 1986, 1999, and 2004), and over-forecasted for low rainfall years (1968 and 1992). That is, the results indicated that the prediction performance for the below- and above-normal rainfall conditions is not as good as for near normal condition. These characteristics of the scatter plot of the observed and the predicted rainfall values are more clearly shown in Figure 3. The hit years are represented as filled symbols in the figure.

**Figure 2.** Comparison of the observed and predicted M-J rainfalls using ANN (11,4,1).

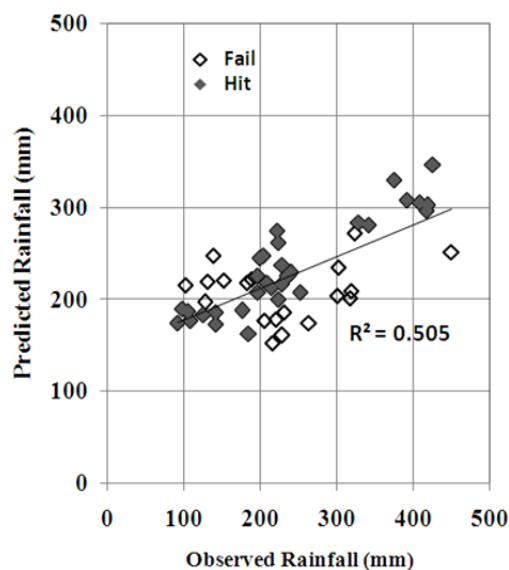


Figure 3. Plot of the observed and predicted M-J rainfalls using ANN (11,4,1).

Table 3 presents the number of years for each category and the hit score which is the number of true years divided by the number of total years. As shown in Table 3, the overall hit score from the ANN model was 62.0%, and for the near-normal rainfall condition, the hit score was particularly high at 70%. However, in the case of below- and above-normal rainfall, the hit scores were not as high as those of near-normal rainfall.

Table 3. Number of hit/fails and hit scores for three categories.

Category	Below Normal	Near Normal	Above Normal	Total
Hit	9	14	8	31
Fail	7	6	6	19
Total	16	20	14	50
Hit Score (%)	56.3	70.0	57.1	62.0

4.2. Quantification of Relative Importance of Input Variables

Each input variable's contribution to the output value was evaluated using Garson's method and Olden's method. Figure 4 shows the relative importance of the 11 independent variables obtained from the Garson's connection weights method in the form of a box plot. A wide variability of the importance values was observed depending on the different random initial weights used. Results show that the differences between the minimum and maximum values were about 6.0–12.8% for the input variables. The greatest difference was observed for the variable EA, and the lowest difference was observed for AO. Predictor contributions of median values ranged from 4.0% to 18.8%, with EA showing the strongest relationship with predicted rainfall, and SLP_D and WP exhibiting the weakest relationships. The EA index followed by NAO, PDO, EPNP, and TNA was identified as the most important predictor of late spring and early summer rainfall in the Geum River Basin. These top five important predictors, which always have a relative importance above 5%, were selected for constructing a more concise ANN model.

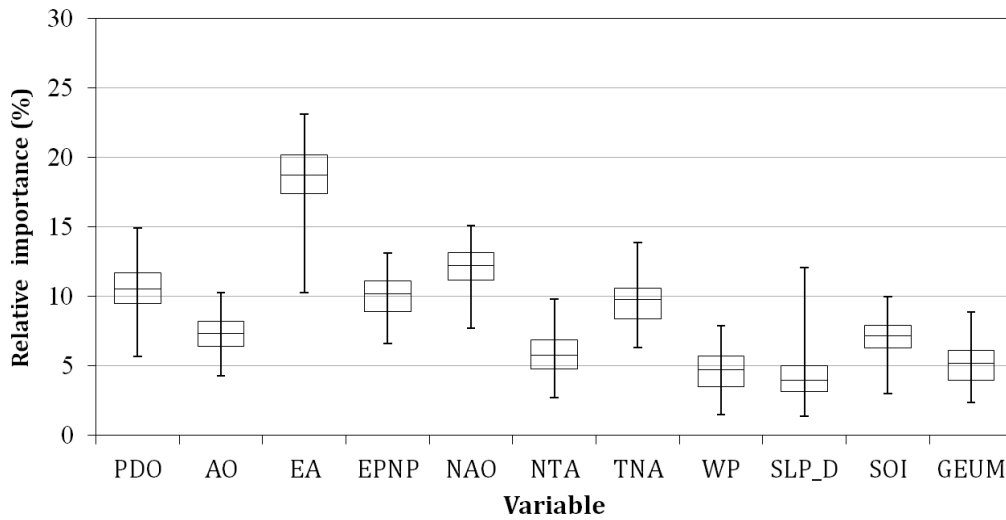


Figure 4. Relative importance of input variables.

Olden and Jackson [36] stated that Garson’s method may be potentially misleading for the interpretation of the contribution of input variables because the method does not consider the direction of the input-output interaction. In some cases, the influence of an input variable on the output response can be negligible when a positive influence through a hidden node is counteracted by a negative influence through another hidden node. In order to compensate for the drawback of Garson’s method, Olden’s method was additionally applied for the quantification of the relative importance of input variables in the present study.

Figure 5 shows the overall connection weights for each input variable obtained from Olden’s method. As expected, a higher variability in the overall connection weights was observed. The predictors’ contributions ranged from -0.41 to 0.23 of the median values. Most variables affected this positively, except for EA, NAO, and WP, which showed a negative influence; in other words, as those values increased, the output rainfall decreased. It was apparent that the most influential variables were EA and NAO, and the least influential variables were SLP_D and GEUM. Therefore, both methods performed similarly in terms of determining the variable importance. Ranking produced the order of EA, NAO, PDO, EPNP, and TNA according to the magnitude of median values. Thus, these top five influential variables were selected as predictors, which agree with the results obtained from Garson’s method.

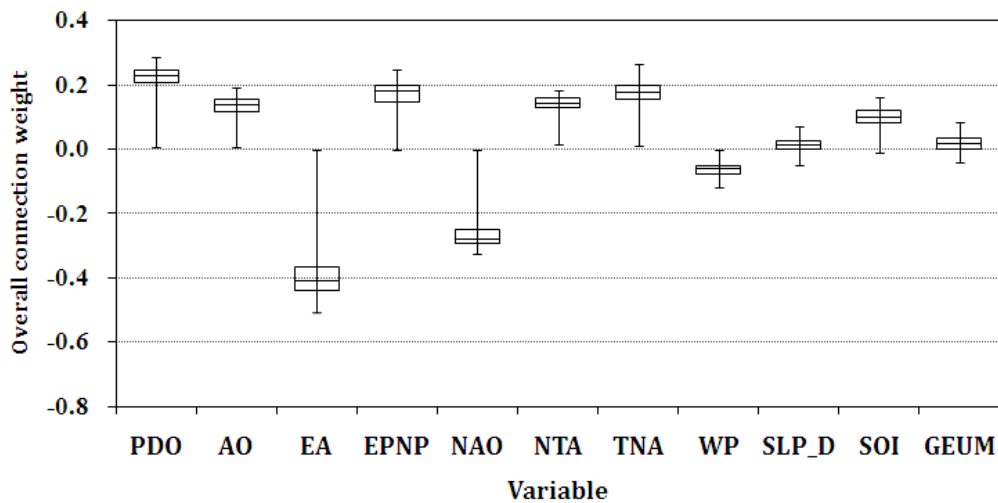


Figure 5. Overall connection weights of input variables.

As pointed out previously [37], the use of a single ANN structure can cause misunderstanding when extracting the contributed input variables because the relative importance can be highly different from that averaged from the group of ANNs. The wider ranges of relative importance are depicted in Figures 4 and 5. Therefore, a set of ANNs with different initial weights and two or more methods of variable importance quantification, such as Garson's and Olden's algorithms, should be used to select significant predictors and to produce a reliable output response.

4.3. Best ANN Model for Rainfall Forecasting

The optimal ANN model structure with five input variables (EA, NAO, PDO, EPNP, and TNA) was determined by varying the number of hidden neurons from 2 to 10. Table 4 shows the training, validation, and testing results for ANN structures. As the number of hidden neurons increased, the RRMSE for the training part decreased and reached a minimum value after a certain number of hidden neurons, but increased for the validation part. For the training part, the RRMSE values ranged from 25.84% to 26.70% (RMSE: 59.31 mm to 60.63 mm), and for the validation part, the values of RRMSE range from 32.72% to 34.79% (RMSE: 75.73 mm to 80.55mm), which shows more accurate performance than the results of the preliminary ANN model. CC values ranged from 0.763 to 0.774 for training, from 0.584 to 0.614 for validation, and from 0.623 to 0.656 for testing. As indicated in Table 4, the ANN (5,2,1) with two hidden neurons had the best prediction accuracy in the validation part. After the ANN (5,2,1) was re-trained using the whole data set of the training and validation parts, the performance was evaluated for the testing part, which shows acceptable results with an RRMSE value of 34.75% (RMSE: 86.84 mm).

Table 4. ANN models performance for training, validation, and testing parts.

Number of Hidden Nodes	RRMSE (%)			CC		
	Training	Validation	Testing	Training	Validation	Testing
2	25.84	32.72	34.75	0.771	0.614	0.623
3	25.87	33.10	34.57	0.770	0.613	0.630
4	26.70	33.46	34.04	0.764	0.610	0.655
5	26.24	33.51	34.23	0.763	0.608	0.646
6	26.18	34.79	34.11	0.764	0.584	0.656
7	25.66	33.83	34.78	0.774	0.609	0.624
8	25.80	33.92	34.66	0.771	0.607	0.630
9	25.76	34.03	34.72	0.772	0.606	0.625
10	25.77	34.11	34.76	0.772	0.604	0.627

Figure 6 shows the actual rainfall versus the predicted data using ANN (5,2,1). The model results showed reasonable accuracy with observed rainfall for most of the yearly M-J rainfall values, except for some deviation for the years of 1980, 1986, 1988, 1992, and 1999. The model significantly under-forecasted the high rainfall years (1980, 1986, and 1999), over-forecasted the low rainfall years (1988 and 1992), and forecasted for the remaining years with reasonable accuracy. Figure 7 shows the scatter plot of the observed and forecasted M-J rainfall values. From the comparison of Figures 3 and 7, ANN (5,2,1) model performed better than ANN (11,4,1) model for the predictions of higher or lower rainfall values.

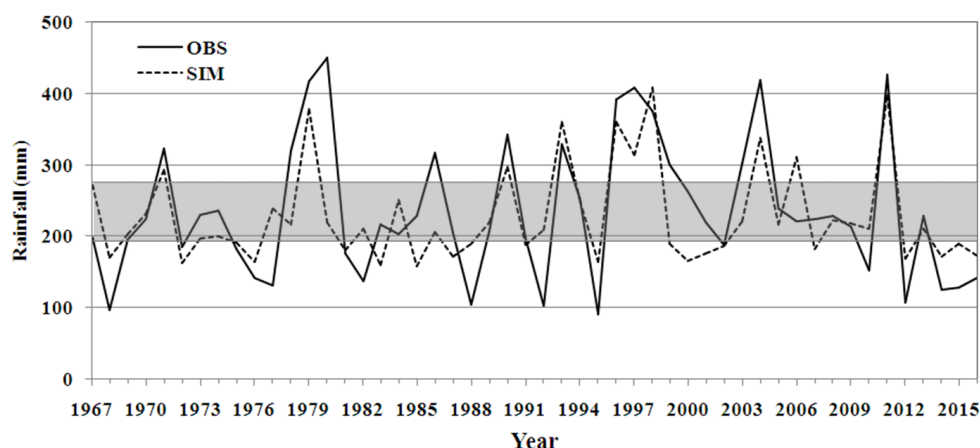


Figure 6. Comparison of the observed and predicted M-J rainfalls using ANN (5,2,1).

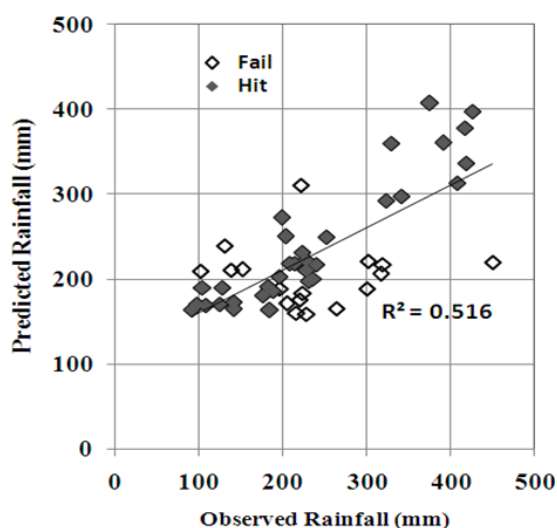


Figure 7. Plot of the observed and predicted M-J rainfalls using ANN (5,2,1).

Table 5 includes the number of hits and fails for each category and the hit score. The overall hit score for the ANN (5,2,1) model was 66.0%, which was higher than the result for ANN (11,4,1) model. The ANN (5,2,1) hit score for the below-normal rainfall was very high at 75%, which is much higher than that of the ANN (11,4,1) model. In terms of the prediction performance of rainfall being below normal, the ANN (5,2,1) model can be more useful for low rainfall forecasting.

Table 5. Number of hit/fails and hit scores for three categories.

Category	Below Normal	Near Normal	Above Normal	Total
Hit	12	12	9	33
Fail	4	8	5	17
Total	16	20	14	50
Hit Score (%)	75.0	60.0	64.3	66.0

This study was the first attempt to construct ANN models with predictors of global teleconnection patterns to forecast basin scale rainfall amounts several months in advance in South Korea. The optimal ANN model had fairly acceptable predictive performance with an RRMSE of about 30% and a Pearson correlation coefficient of more than 0.6. This performance is as good as that of other studies [12,17,18]

on forecasting monsoon summer rainfall for East Asian regions, despite the fact that the ANN model developed in the present study had a longer lead time.

The present study intended to reduce the time-consuming work needed to construct ANN architecture, such as the determination of significant input variables. There could be a large number of climate variables affecting output response; thus, it is difficult to find optimal input variables by a trial and error procedure. To overcome this problem, firstly, possible candidates of lagged climate indices were determined from the correlation analysis, a preliminary ANN model was constructed using the candidates, and then the final optimal ANN model was developed using a few significant inputs, which were selected by evaluating the contribution of each variable. With the help of the correlation analysis and the quantification of variable importance, the time-consuming laborious trial and error procedure could be greatly reduced. To the best of our knowledge, the approach using teleconnection climate indices and quantifying variable importance has not been applied to seasonal rainfall forecasting ANN models. We think that this approach can be useful to enable quick forecasting.

5. Conclusions

We constructed an artificial neural network model to predict rainfall in late spring and early summer for the Geum River Basin, South Korea. For this purpose, several delayed global climate indices and the areal average rainfall of the basin were used as predictors and a predictand of the ANN model, respectively.

After identifying the lagged correlation between climate indices and rainfall amount in May and June, a preliminary ANN model with 11 input variables including the global climate indices of AO, EP/NP, EA, NAO, NTA, TNA, WP, SOI, PDO, SLP_D, and the areal rainfall of the Geum River Basin with different lag times for each was constructed. The optimal hidden neuron number of the ANN model with 11 input variables was selected as four based on the four-fold CV procedure. The preliminary ANN (11,4,1) model showed satisfactory prediction performance with RRMSE values of 30.43%, 35.52%, and 34.41% for the training, validation, and testing data sets, respectively. The hit score, which is the number of hit years divided by the number of total years, was 62.0%. However, ANN (11,4,1) has a tendency to under-forecast for high rainfall years while over-forecasting for low rainfall years.

We quantified the relative importance of input variables using Garson's and Olden's connection weight methods to identify highly significant predictors and to construct a simple ANN model with a few input variables. The five lagged climate indices—EA, NAO, PDO, EP/NP, and TNA—were selected as predictors and the optimal structure of the ANN with two hidden neurons was determined based on the four-fold CV results. The final best ANN (5,2,1) model showed acceptable performance with RRMSE values of 25.84%, 32.72%, and 34.75% for training, validation, and testing parts, respectively. The hit score was found to be 66% for total years, and 75.0%, 60.0%, 64.3% for below, near, and above-normal historical conditions, respectively. The results revealed that ANN (5,2,1) was more successful than ANN (11,4,1) in predicting early spring and late summer rainfall of the basin of interest, particularly showing good performance in below-normal condition. The results also indicated that the quantification of the contribution of the variable relative importance was able to improve the accuracy of forecasting rainfall forecasts by removing some input variables that show a weak correlation.

The optimal model predicted higher values of rainfall to be acceptable, but the prediction of the lower values was relatively insufficient. Future studies need to be carried out to improve the prediction of the extreme lower rainfall amount with the additional consideration of new climatic indices, as well as weather data such as temperature, humidity, and wind.

In conclusion, this study revealed the possibility of seasonal rainfall forecasting using ANNs and lagged climate indices four months in advance for the study region. Good prediction of the late spring-early summer rainfall amount could allow for the more flexible operation of multi-purpose dams in the Geum River and provide sufficient time to prepare strategies against potential drought damage. The developed ANN model can be considered an alternative tool to the existing physically-based forecasting models.

Author Contributions: All authors substantially contributed in conceiving and designing of the approach and realizing this manuscript. J.L. implemented the artificial neural network models and analyzed the results. C.-G.K. and J.E.L. worked on the analysis and presentation of the results. N.W.K. and H.K. analyzed the results and supervised the entire research. All five authors jointly wrote the paper. All authors have read and approved the final manuscript.

Funding: This research was funded by the Korea Institute of Civil Engineering and Building Technology (grant number 20180101-001) and the APC was funded by the Korea Institute of Civil Engineering and Building Technology.

Acknowledgments: This research was supported by a grant from a Strategic Research Project (Developing technology for water scarcity risk assessment and securing water resources of small and medium sized catchments against abnormal climate and extreme drought) funded by the Korea Institute of Civil Engineering and Building Technology. Authors appreciate the editors of the journal and the reviewers for their valuable comments and suggestions for improvements.

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

References

- Nayak, R.D.; Mahapatra, A.; Mishra, P. A survey on rainfall prediction using artificial neural network. *Int. J. Comput. Appl.* **2013**, *72*, 32–40.
- Adamowski, J.; Sun, K. Development of a coupled wavelet transform and neural network method for flow forecasting of non-perennial rivers in semi-arid watersheds. *J. Hydrol.* **2010**, *390*, 85–91. [[CrossRef](#)]
- Bodri, L.; Cermak, V. Prediction of extreme precipitation using a neural network: Application to summer flood occurrence in Moravia. *Adv. Eng. Softw.* **2000**, *31*, 311–321. [[CrossRef](#)]
- Bodri, L.; Cermak, V. Neural network prediction of monthly precipitation: Application to summer flood occurrence in two regions of central Europe. *Stud. Geophys. Geod.* **2001**, *45*, 155–167. [[CrossRef](#)]
- Wu, X.; Hongxing, C. Forecasting monsoon precipitation using artificial neural networks. *Adv. Atmos. Sci.* **2001**, *18*, 950–958.
- Philip, N.S.; Joseph, K.B. A neural network tool for analyzing trends in rainfall. *Comput. Geosci.* **2003**, *29*, 215–223. [[CrossRef](#)]
- Chakraverty, S.; Gupta, P. Comparison of neural network configurations in the long-range forecast of southwest monsoon rainfall over India. *Neural Comput. Appl.* **2008**, *17*, 187–192. [[CrossRef](#)]
- Chattopadhyay, S.; Chattopadhyay, G. Comparative study among different neural net learning algorithms applied to rainfall time series. *Meteorol. Appl.* **2008**, *15*, 273–280. [[CrossRef](#)]
- Gholizadeh, M.H.; Darand, M. Forecasting precipitation with artificial neural networks (Case Study: Tehran). *J. Appl. Sci.* **2009**, *9*, 1786–1790. [[CrossRef](#)]
- Aksoy, H.; Dahamsheh, A. Artificial neural network models for forecasting monthly precipitation in Jordan. *Stoch. Environ. Res. Risk Assess.* **2009**, *23*, 917–931. [[CrossRef](#)]
- Bilgili, M.; Sahin, B. Prediction of long-term monthly temperature and rainfall in Turkey. *Energy Sour. Part A* **2010**, *32*, 60–71. [[CrossRef](#)]
- Yuan, F.; Berndtsson, R.; Uvo, C.B.; Zhang, L.; Jiang, P. Summer precipitation prediction in the source region of the Yellow River using climate indices. *Hydrol. Res.* **2016**, *47*, 847–856. [[CrossRef](#)]
- Jiang, P.; Gautam, M.R.; Zhu, J.; Yu, Z. How well do the GCMs/RCMs capture the multi-scale temporal variability of precipitation in the Southwestern United States? *J. Hydrol.* **2013**, *479*, 75–85. [[CrossRef](#)]
- Leathers, D.J.; Yarnal, B.; Palecki, M.A. The Pacific North-American teleconnection pattern and United-States climate 1. Regional temperature and precipitation associations. *J. Clim.* **1991**, *4*, 517–528. [[CrossRef](#)]
- Silverman, D.; Dracup, J.A. Artificial neural networks and long-range precipitation in California. *J. Appl. Meteorol.* **2000**, *39*, 57–66. [[CrossRef](#)]
- Kumar, D.N.; Reddy, M.J.; Maity, R. Regional rainfall forecasting using large scale climate teleconnections and artificial intelligence techniques. *J. Intell. Syst.* **2007**, *16*, 307–322.
- Iseri, Y.; Dandy, G.C.; Maier, H.R.; Kawamura, A.; Jinno, K. Medium term forecasting of rainfall using artificial neural networks. In Proceedings of the International Congress on Modelling and Simulation, Melbourne, Australia, 12–15 December 2005; pp. 1834–1840.
- Hartmann, H.; Becker, S.; King, L. Predicting summer rainfall in the Yangtze River basin with neural networks. *Int. J. Climatol.* **2008**, *28*, 925–936. [[CrossRef](#)]

19. Abbot, J.; Marohasy, J. Application of artificial neural networks to rainfall forecasting in Queensland, Australia. *Adv. Atmos. Sci.* **2012**, *29*, 717–730. [[CrossRef](#)]
20. Abbot, J.; Marohasy, J. Application of artificial neural networks to forecasting monthly rainfall one year in advance for locations within the Murray Darling basin, Australia. *Int. J. Sustain. Dev. Plan.* **2017**, *12*, 1282–1298. [[CrossRef](#)]
21. Badr, H.S.; Zaitchik, B.F.; Guikema, S.D. Application of statistical models to the prediction of seasonal rainfall anomalies over the Sahel. *J. Appl. Meteorol. Climatol.* **2013**, *53*, 614–636. [[CrossRef](#)]
22. Rasel, H.M.; Imteaz, M.A.; Hossain, I.; Mekanki, F. Comparative study between linear and non-linear modelling techniques in rainfall forecasting for South Australia. In Proceedings of the International Congress on Modelling and Simulation, Gold Coast, Australia, 29 November–4 December 2015; pp. 2012–2018.
23. Hong, I.; Lee, J.H.; Cho, H.S. National drought management framework for drought preparedness in Korea (lessons from the 2014–2015 drought). *Water Policy* **2016**, *18*, 89–106. [[CrossRef](#)]
24. Peres, D.J.; Iuppa, C.; Cavallaro, L.; Cancelliere, A.; Foti, E. Significant wave height record extension by neural networks and reanalysis wind data. *Ocean Model.* **2015**, *94*, 128–140. [[CrossRef](#)]
25. Dawson, C.W.; Wilby, R.L. Hydrological modelling using artificial neural networks. *Prog. Phys. Geogr.* **2001**, *25*, 80–108. [[CrossRef](#)]
26. De Vos, N.J.; Rientjes, T.H.M. Constraints of artificial neural networks for rainfall-runoff modelling: Trade-offs in hydrological state representation and model evaluation. *Hydrol. Earth Syst. Sci.* **2005**, *9*, 111–126. [[CrossRef](#)]
27. Sumi, S.M.; Zaman, M.F.; Hirose, H. A rainfall forecasting method using machine learning models and its application to the Fukuoka city case. *Int. J. Appl. Math. Comput. Sci.* **2012**, *22*, 841–854. [[CrossRef](#)]
28. Singh, S.K.; Jain, S.K.; Bárdossy, A. Training of artificial neural networks using information-rich data. *Hydrology* **2014**, *1*, 40–62. [[CrossRef](#)]
29. Kim, T.W.; Valdes, J.B. Nonlinear model for drought forecasting based on a conjunction of wavelet transforms and neural networks. *J. Hydrol. Eng.* **2003**, *8*, 319–328. [[CrossRef](#)]
30. Sung, J.Y.; Lee, J.; Chung, I.M.; Heo, J.H. Hourly water level forecasting at tributary affected by main river condition. *Water* **2017**, *9*, 644. [[CrossRef](#)]
31. Rumelhart, D.E.; McClelland, J.L. Foundations. In *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*; MIT Press: Cambridge, MA, USA, 1986; Volume 1.
32. Supharatid, S. Application of a neural network model in establishing a stage-discharge relationship for a tidal river. *Hydrol. Process.* **2003**, *17*, 3085–3099. [[CrossRef](#)]
33. Stone, M. Cross-validatory choice and assessment of statistical predictions. *J. R. Stat. Soc. Ser. B* **1974**, *36*, 111–147.
34. Rodriguez, J.D.; Perez, A.; Lozano, J.A. Sensitivity analysis of k-fold cross validation in prediction error estimation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2010**, *32*, 569–575. [[CrossRef](#)] [[PubMed](#)]
35. Garson, G.D. Interpreting neural-network connection weights. *AI Expert* **1991**, *6*, 47–51.
36. Olden, J.D.; Jackson, D.A. Illuminating the “black box”: A randomization approach for understanding variable contributions in artificial neural networks. *Ecol. Model.* **2002**, *154*, 135–150. [[CrossRef](#)]
37. Pentos, K. The methods of extracting the contribution of variables in artificial neural network models-Comparison of inherent instability. *Comput. Electron. Agric.* **2016**, *127*, 141–146. [[CrossRef](#)]

