Modeling surface water quality using the adaptive neuro-fuzzy inference system aided by input optimization

sustainability

MDPI

Article

Modeling Surface Water Quality Using the Adaptive Neuro-Fuzzy Inference System Aided by Input Optimization

Muhammad Izhar Shah [1,*], Taher Abunama [2], Muhammad Faisal Javed [1], Faizal Bux [2], Ali Aldrees [3], Muhammad Atiq Ur Rehman Tariq [4,5] and Amir Mosavi [6,7,8,*]

1 Department of Civil Engineering, COMSATS University Islamabad, Abbottabad Campus, Abbottabad 22060, Pakistan; arbabfaisal@cuiatd.edu.pk
2 Institute for Water and Wastewater Technology, Durban University of Technology, Durban 4001, South Africa; tahera@dut.ac.za (T.A.); faizalb@dut.ac.za (F.B.)
3 Department of Civil Engineering, College of Engineering in Al-Kharj, Prince Sattam Bin Abdulaziz University, Al-Kharj 16273, Saudi Arabia; a.aldrees@psau.edu.sa
4 Institute for Sustainable Industries & Liveable Cities, Victoria University, P.O. Box 14428, Melbourne, VIC 8001, Australia; muhammadatiqurrehman.tariq@vu.edu.au
5 College of Engineering and Science, Victoria University, Melbourne, VIC 8001, Australia
6 Faculty of Civil Engineering, Technische Universität Dresden, 01069 Dresden, Germany
7 Department of Ecology, Technische Universität Kaiserslautern, 67663 Kaiserslautern, Germany
8 John von Neumann Faculty of Informatics, Obuda University, 1034 Budapest, Hungary
* Correspondence: mizharshah.civ@uetpeshawar.edu.pk (M.I.S.); amir.mosavi@mailbox.tu-dresden.de (A.M.)

Citation: Shah, M.I.; Abunama, T.; Javed, M.F.; Bux, F.; Aldrees, A.; Tariq, M.A.U.R.; Mosavi, A. Modeling Surface Water Quality Using the Adaptive Neuro-Fuzzy Inference System Aided by Input Optimization. *Sustainability* **2021**, *13*, 4576. https://doi.org/10.3390/su13084576

Academic Editors: Daeryong Park, Momcilo Markus and Myoung-Jin Um

Received: 4 March 2021
Accepted: 13 April 2021
Published: 20 April 2021

Abstract: Modeling surface water quality using soft computing techniques is essential for the effective management of scarce water resources and environmental protection. The development of accurate predictive models with significant input parameters and inconsistent datasets is still a challenge. Therefore, further research is needed to improve the performance of the predictive models. This study presents a methodology for dataset pre-processing and input optimization for reducing the modeling complexity. The objective of this study was achieved by employing a two-sided detection approach for outlier removal and an exhaustive search method for selecting essential modeling inputs. Thereafter, the adaptive neuro-fuzzy inference system (ANFIS) was applied for modeling electrical conductivity (EC) and total dissolved solids (TDS) in the upper Indus River. A larger dataset of a 30-year historical period, measured monthly, was utilized in the modeling process. The prediction capacity of the developed models was estimated by statistical assessment indicators. Moreover, the 10-fold cross-validation method was carried out to address the modeling overfitting issue. The results of the input optimization indicate that Ca^{2+}, Na^{+}, and Cl^{-} are the most relevant inputs to be used for EC. Meanwhile, Mg^{2+}, $HCO_3{}^{-}$, and $SO_4{}^{2-}$ were selected to model TDS levels. The optimum ANFIS models for the EC and TDS data showed R values of 0.91 and 0.92, and the root mean squared error (RMSE) results of 30.6 µS/cm and 16.7 ppm, respectively. The optimum ANFIS structure comprises a hybrid training algorithm with 27 fuzzy rules of triangular fuzzy membership functions for EC and a Gaussian curve for TDS modeling, respectively. Evidently, the outcome of the present study reveals that the ANFIS modeling, aided with data pre-processing and input optimization, is a suitable technique for simulating the quality of surface water. It could be an effective approach in minimizing modeling complexity and elaborating proper management and mitigation measures.

Keywords: data-driven; outlier detection; machine learning; surface water quality; input optimization; neuro-fuzzy; water quality management; hydrology; artificial intelligence; big data

1. Introduction

Surface water bodies are naturally available resources and have always been considered essential for the persistence of the ecosystem. The quality of these water resources is adversely affected due to anthropogenic activities, including industrialization and population growth. The term "water quality" refers to the physical, chemical, and biochemical

properties of water [1]. Rivers are most vulnerable to environmental pollution from various sources due to their dynamic natures, acting as a carrier for waste loads [2]. In developing countries, a huge volume of liquid waste is discarded into surface water bodies, which raises environmental issues and water quality concerns [3,4]. The major responsible factors for water quality deterioration are atmospheric processes, climatic factors, pollution from agricultural areas, and anthropogenic factors [5–7]. The poor quality of water is a serious problem threatening human health, agriculture, and ecosystems [8].

The salinity of surface water bodies has increased steadily over the years, negatively affecting the quality of irrigation and drinking water [9,10]. The deposition of salts due to salinity causes an unfavorable hydrologic environment that restricts the use of water for domestic and agricultural purposes. Moreover, the management of water resources and salinity has become imperative because the balance between water availability and demand has reached the critical limit [11]. The considerable parameters for water salinity are electrical conductivity (EC) and total dissolved solids (TDS). Both of these parameters are accepted indicators for the assessment of irrigation and drinking water. TDS comprises a variety of inorganic salts, that is, sodium (Na^+), calcium (Ca^{2+}), magnesium (Mg^{2+}), nitrates (NO_3), chloride (Cl^-), sulfate ($SO_4{}^{2-}$), and many kinds of dissolved organic matter [12]. Increased salts and organic content in water indicate poor water quality [13]. According to WHO guidelines, the permissible range of TDS in drinking water is 300–600 mg/L, while the allowable limit for agricultural water is 450–2000 mg/L [14].

Laboratory analysis and experimental approaches have also been reported in the literature for TDS and EC measurements [12,15]. The manual calculations and laboratory testing are time-consuming and require specialized equipment. Consequently, the desired outcome cannot be attained through such testing methods. Therefore, modeling techniques can be used to estimate water quality, as models are capable of accurately predicting the essential water quality parameters [16,17]. A number of research studies have been devoted to evaluating a variety of water quality parameters, employing stochastic, deterministic, and numerical models. These conventional modeling techniques can only deliver forecasts for stationary and linear sets of data, which makes them unsuitable for reliable predictions [18]. On the contrary, the artificial intelligence (AI) models are reported to have the ability to manage large, nonlinear datasets and the complex phenomena of environmental and hydrological processes, and therefore, overcoming the drawbacks of conventional models.

AI techniques, that is, an artificial neural network (ANN), an adaptive neuro-fuzzy inference system (ANFIS), gene expression programming (GEP), and a support vector machine (SVM) [17,19–24], have been employed in various research studies for modeling the water quality parameters. These parameters comprise dissolved oxygen (DO), biochemical oxygen demand (BOD), nitrate (NO_3), electrical conductivity (EC), pH, and the sodium absorption ratio (SAR) [15]. Ghavidel et al. (2014) [25] employed ANN, ANFIS, and GEP for modeling TDS concentration in a catchment in Iran. They concluded a more reliable prediction rate with GEP as compared to the ANN and ANFIS. Asadollahfardi et al. (2018) [26] employed time series and multilayer perception (MLP) for TDS prediction. Various water quality parameters, including bicarbonates, chlorides, and calcium were used as modeling inputs. The results of the study revealed the better performance of MLP than the time series. Zounemat-Kermani et al. (2019) [27] used an ANN in predicting the nitrogen content. The findings of the study indicated excellent performance of the MLP model. Sattari et al. (2016) [13] considered support vector regression (SVR) and the k-nearest neighbor (k-NN) method to predict the salinity in the Lighvan Chay River in Iran, by using a different combination of input variables. The authors reported improved performance of the SVR method over the k-NN method. Maroufpoor et al. (2019) [28] used ANN and ANFIS models for modeling the spatial distribution of EC in groundwater. The output of the study demonstrated reliable predictions of the models with the lowest mean absolute error (MAE) and root mean squared error (RMSE) values. Aryafar et al. (2019) [29] employed genetic programming, ANFIS, and ANN in predicting the EC and

TDS concentrations. The authors reported excellent performance and accurate predictions by the ANFIS and ANN. A study by Shah et al. (2020) [17] compared the performances of ANN, GEP, and regression techniques, in terms of predicting the TDS and EC concentrations. A sensitivity and parametric study was carried out for evaluating the connection between the inputs and the output. The authors reported improved performance of the GEP as compared with the ANN and regression techniques. Sensitive parameters were identified, which had a direct impact on the modeling output.

AI and soft computing techniques have been successfully applied in the abovementioned studies for water quality prediction but with some shortcomings. Algorithms such as ANN, SVM, and GEP have some unknown parameters [12,30]. These parameters have a significant effect on the accuracy of the model output. A common and reoccurring issue in the use of these algorithms is that these techniques may be trapped in a local optimum [30]. Moreover, the water quality data is highly chaotic, stochastic, and nonlinear, and the development of a standalone AI-based model has limitations in water quality modeling. Therefore, integrating the data pre-processing and optimization approaches with AI models are likely to enhance their accuracy and predicting capabilities.

In this study, the data pre-processing, followed by the exhaustive search method for input optimization and ANFIS modeling, are proposed to solve the complexity of modeling surface water quality. A two-sided outlier detection approach was used for data preprocessing, with the threshold outlier values set to $\pm 3\sigma$ (sigma rule). Various water quality parameters, recorded monthly over a period of 30 years (1975–2005), were used in the modeling process. An optimization routine was developed to select the most correlated and significant input variables. Afterward, an efficient ANFIS model structure was developed, which was efficient in predicting the surface water quality indicated by the EC and TDS concentrations in the upper Indus Basin (UIB). The best ANFIS structure, which yielded the lowest modeling error with a minimum rules number, was selected for reducing the modeling complexity. Furthermore, cross-validation was employed to evaluate the final outcome from the ANFIS model. The methodology adopted in this study will help in minimizing the data sampling and processing efforts in surface water quality assessments.

2. Materials and Methods

2.1. Case Study and Modeling Dataset

The Indus River is 2880 km long and is considered a major river in Asia, with a drainage area of almost 912,000 km^2 [31]. The portion of the Indus River upstream of the Tarbela reservoir is the upper Indus basin (UIB). It has a total length of 1150 km and drains a large area of 165,400 km^2 [32,33]. The elevation varies from 455 m to 8611 m, and the climate differs significantly inside the basin. The annual precipitation range is 100–200 mm and occurs due to the turbulences in the western mid-latitude [34–37]. The study area is shown in Figure 1.

The data employed in this study for ANFIS model development was collected from water & power development authority (WAPDA), Pakistan. The dataset contained 321 monthly data points collected over a period of 30 years (1975 to 2005), measured at the Bisham Qilla outlet. The acquired data have the information of nine variables, which are calcium (Ca^{2+}), magnesium (Mg^{2+}), sodium (Na^{+}), chloride (Cl^{-}), sulphate ($SO_4{}^{2-}$), bicarbonates ($HCO_3{}^{-}$), pH, EC, and TDS. The descriptive statistics, that is, the mean, skewness, standard deviation, and kurtosis of the data, are given in Table 1. The normal probability curves in Figure 2 show the distribution of the target EC and TDS concentrations. A symmetrical curve of the mean of the dataset depicts a normal distribution [19,38]. Moreover, the literature demonstrates that data preprocessing is an essential process in any data mining process, aiming to eliminate the effect of missing or outlier measurements, which may occur during the collecting of the data [19,39]. In this study, a two-sided outlier detection approach was used for data preprocessing and outlier elimination, with the threshold outlier values set to $\pm 3\sigma$ (sigma rule).

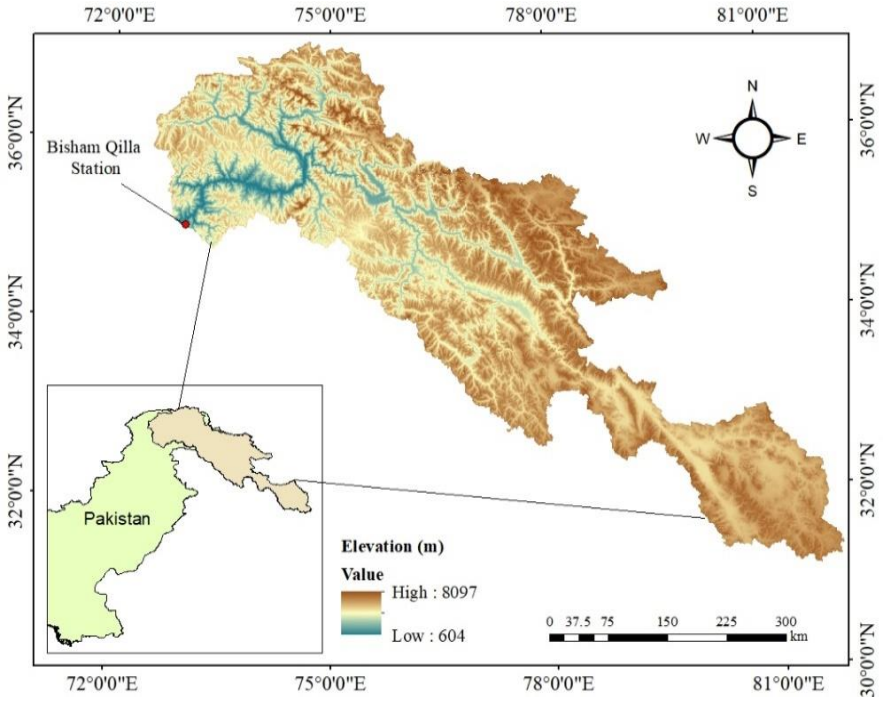


Figure 1. Detailed depiction of the study area.

Table 1. Descriptive statistics of the model variables.

	Ca^{2+}	Mg^{2+}	Na^{+}	HCO_3^{-}	Cl^{-}	SO_4^{2-}	PH	EC	TDS
Mean	1.50	0.58	0.50	1.72	0.27	0.56	7.88	244.32	143.75
Minimum	0.61	0.03	0.05	0.11	0	0.1	7.08	88	60
Maximum	3.15	1.5	2.1	3.5	0.78	1.6	8.4	510	308
SD	0.38	0.26	0.43	0.60	0.12	0.33	0.23	73.43	41.56
Kurtosis	3.01	0.51	3.92	0.89	3.16	0.55	0.23	0.91	1.19
Skewness	0.94	0.44	2.10	0.76	1.43	0.83	−0.47	0.71	0.86

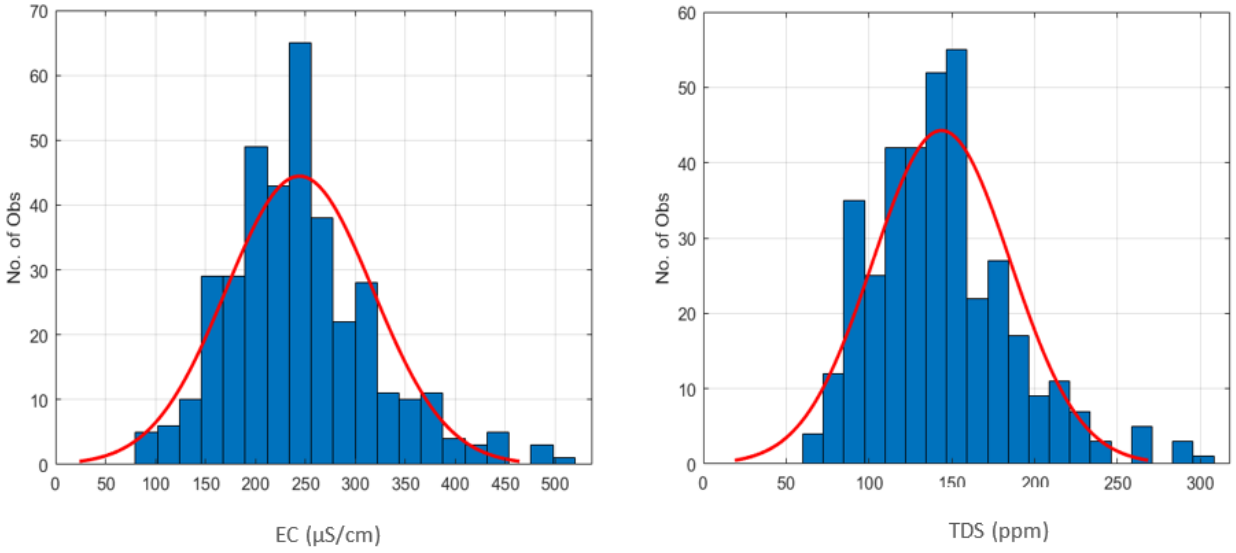


Figure 2. Normal probability curves of the electrical conductivity (EC) and total dissolved solids (TDS) data.

2.2. Input Optimization and Model Development

The Adaptive Neuro-Fuzzy Inference System (ANFIS) model is a kind of ANN, which is based on implementing the Takagi–Sugeno (TS) fuzzy approach, as shown in Figure 3. ANFIS implements fuzzy logic (FL) in the framework of an ANN [40]. The development process of ANFIS modeling involves identifying the most relevant inputs that correlate with a targeted output. The defining optimum rules, types, and the number of the associated membership functions (MFs) need to be evaluated, aiming at selecting the optimum ANFIS model structure with the lowest yielded errors. As an example, two TS fuzzy sets of "if–then" rules in a typical ANFIS structure are the following:

- Rule 1: If x_1 is A_1 and x_2 is B_1, then $f_1 = p_1x_1 + q_1x_2 + r_1$
- Rule 2: If x_1 is A_2 and x_2 is B_2, then $f_2 = p_2x_1 + q_2x_2 + r_2$

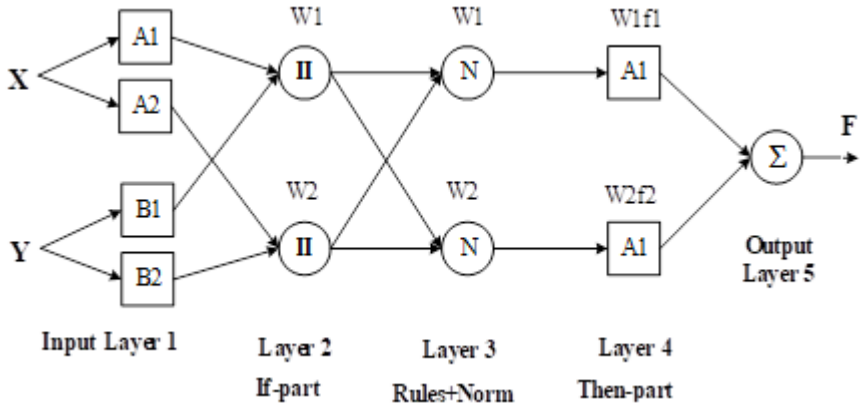


Figure 3. A typical adaptive neuro-fuzzy inference system (ANFIS) architecture

where p_1, q_1, p_2, and q_2 are ANFIS parameters, while Ai and Bj are the linguistic labels or grades. According to Ying et al. (1995) [41], ANFIS architecture consists of five layers. See Figure 3 for ANFIS architecture adapted from [42]. A brief description of the role of these layers are described as follows.

- Layer 1, or fuzzification layer, receives the input values and identifies the MFs.
- Layer 2, or rule layer, generates the firing strengths for the rules.
- Layer 3, or normalization layer, normalizes the computed strengths.
- Layer 4 receives the normalized values and the consequence parameter sets.
- Layer 5, or defuzzification layer, returns the values to the final output.

In the present study, an ANFIS edit toolbox and coding in the MATLAB 2019b environment were used to train and develop the proposed ANFIS models. As mentioned earlier, seven input variables and two outputs were involved in the modeling process. All selected input parameters were related to the targeted surface water quality (EC and TDS). Moreover, the model training phase was conducted using the odd records (data points), while the even records (data points) were used in the model testing phase. The ANFIS learning process was repeated for many epochs, with an aim at reducing the errors between the actual and the ANFIS modeling output. A flowchart of the data pre-processing, input optimization, and the ANFIS structure development is presented in Figure 4.

2.3. Model Assessment Criteria

The obtained results from the model were evaluated using numerous statistical checks. R-squared (R^2) was used to evaluate the relationship between the observed values and the predicted values. The equation for calculating R^2 is denoted by Equation (1), as follows:

$$R^2 = \left(\frac{\sum_{i=1}^{n}(t_i - \bar{t})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(t_i - \bar{t})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}\right)^2 \quad (1)$$

In Equation (1), n, t, and y are the numbers of the observed data, observed values, and predicted values, respectively, whereas, $\bar{t}$ and $\bar{y}$ are the average observed and predicted values. The range of R^2 values 0–1, with 1 being the highest accurate relationship possible. However, values of R^2 greater than 0.7 are considered highly reliable in engineering models.

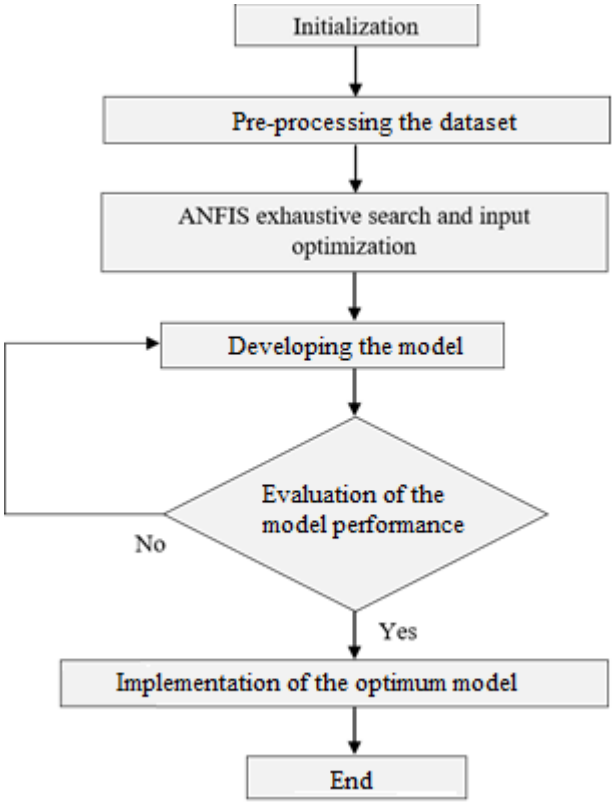


Figure 4. Flowchart of the study methodology.

Other proven statistical checks, like the mean absolute error (MAE) and root mean squared error (RMSE), were also applied to evaluate the accuracy of the developed model. One of the main advantages of using RMSE is to assign higher weightage (as it contains a square) to larger errors. Equations (2) and (3) show the mathematical expressions for the calculations of MAE and RMSE, respectively.

$$\mathrm{MAE} = \sum_{i=1}^{n} \frac{|t_i - y_i|}{n} \tag{2}$$

$$\mathrm{RMSE} = \sqrt{\sum_{i=1}^{n} \frac{(t_i - y_i)^2}{n}} \tag{3}$$

In addition to R, MAE, and RMSE, the percentage relative error (RE%) and the Nash–Sutcliffe coefficient (NSC) were also used to assess the accuracy of the developed model. The NSC is recommended by various researchers, including the American Society of Civil Engineers (ASCE) Code. The range of values for the NSC is $-\infty$ to 1, with values greater than 0.8 showing an accurate model. Equations (4) and (5) show the mathematical expressions for the calculations of RE% and NSC, respectively.

$$\mathrm{RE\%} = \left(\frac{(t_i - y_i)}{t_i}\right) \cdot 100 \tag{4}$$

$$\mathrm{NSC} = 1 - \left(\frac{\sum_{i=1}^{n_L} (t_i - y_i)^2}{\sum_{i=1}^{n_L} (\bar{t}_i - y_i)^2} \right) \quad (5)$$

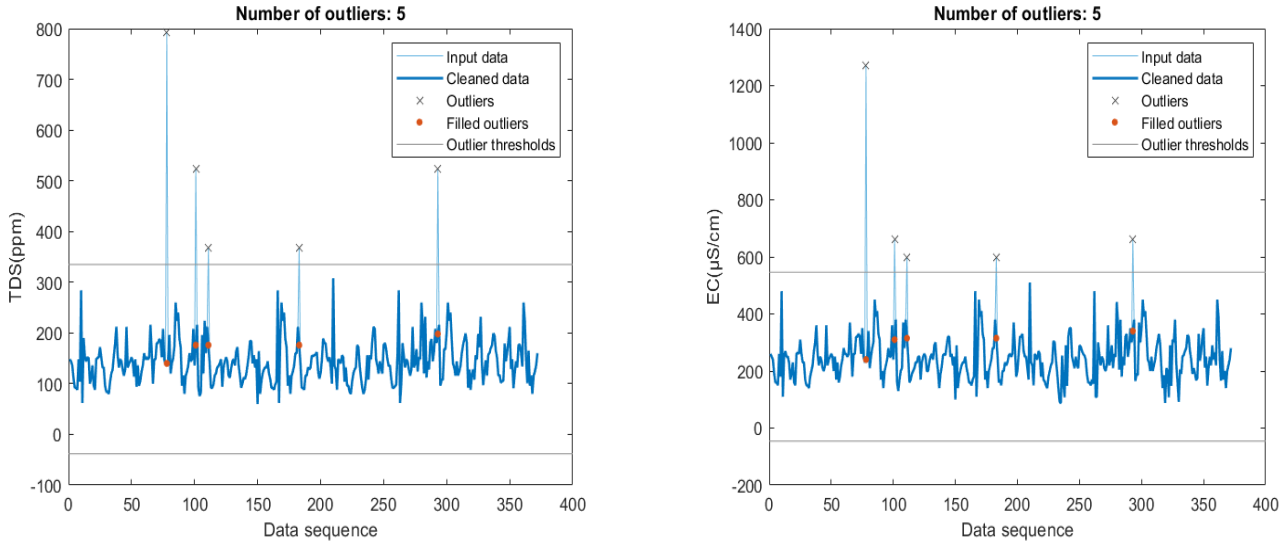
2.4. 10-Fold Cross-Validation

The performance assessment of a developed machine learning model can be a difficult task, as the model cannot provide the required output based on the data that has not been used for training. A dataset is usually divided into training and testing phases, and then the outcome is evaluated based on statistical criteria, but this method is not applicable in all scenarios [43]. Therefore, to verify the generalized capability and reduce the overfitting problem of a learning model, the 10-fold cross-validation method has been recommended in the literature [17,44,45]. This method divides the whole dataset into 10 subclasses. Among all 10 subclasses, the first 9 are used for model training and the remaining subclass is used for validation. The same process is carried out for all the subsets and the output was expressed, employing the mean accuracy obtained in the 10 rounds. In our study, the same cross-validation method has been applied to validate the ANFIS model.

3. Results and Discussion

3.1. Data Pre-Processing

The collected data were statistically analyzed in order to check the consistency and reliability. The final dataset included the information of nine variables, that is, Ca^{2+}, Mg^{2+}, Na^{+}, Cl^{-}, $SO_4{}^{2-}$, $HCO_3{}^{-}$, pH, EC, and TDS. The data-refining process was performed using MATLAB 2019b. Figure 5 shows the data pre-processing and outlier elimination from the targeted EC and TDS data. A two-sided outlier detection approach was adopted and the threshold outlier values were set to $\pm 3\sigma$ (sigma rule). Any values outside this threshold were identified as outliers and removed, as illustrated in Figure 5, for EC and TDS. Similarly, the data cleaning and outlier removal procedure was applied to the other parameters as well.



(**a**) (**b**)

Figure 5. Data pre-processing for the target (**a**) EC and (**b**) TDS data.

3.2. Input Optimization

The selection of the best input combination serves as a base for the accurate prediction of the desired output. If the number of variables is high, the computational time will be high, as will the number of combinations [46]. For the input optimization process, firstly the dataset was separated into two subsets, that is, the odd values were selected

for the model training phase, while the even values were selected for the testing phase. Secondly, the stepwise ANFIS exhaustive search function for input optimization was applied to identify the most relevant inputs for modeling the EC and TDS levels. The input optimization methods have been successfully employed in many research studies for robust model development [47–49].

The following Figure 6a,b shows the exhaustive error results, represented by the RMSE values for the training and testing datasets. The first three input variables (Ca^{2+}, Na^{+}, and Cl^{-}) are the most correlated and relevant variables to the targeted output. This combination was selected because it showed the lowest RMSE values in the training data. It can be noticed that some of the other combinations have lower errors in the testing data; however, they showed high training errors. Therefore, based on the results of the input optimization using the ANFIS exhaustive search method, the three more relevant inputs were Ca^{2+}, Na^{+}, and Cl^{-} in modeling the EC concentrations, while Mg^{2+}, HCO_3^{-}, and SO_4^{2-} were selected to model the TDS levels. These selected input parameters are the most correlated with the variation in the target output concentrations in each case. Various research studies reported that Ca^{2+} and Cl^{-} are important parameters for EC, while Mg^{2+}, total hardness, and SO_4^{2-}, along with other parameters, are the effective inputs to model TDS [17,20,50].

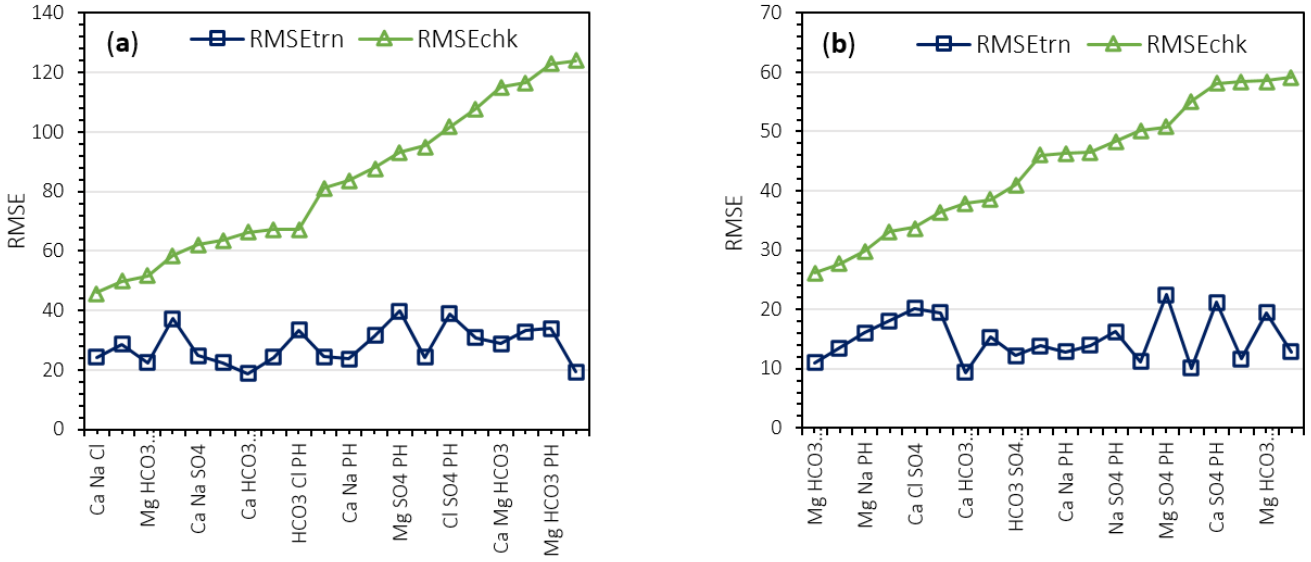


Figure 6. The results of RMSE of the possible input combinations for (**a**) EC and (**b**) TDS data.

3.3. ANFIS Model Development

Upon defining the best input combination, the development of the best ANFIS structure was conducted by applying various types and numbers of membership functions (MFs), and different rules and epoch numbers. This was performed to test all the possibilities of the ANFIS parameters and compare their abilities in modeling the surface water quality (EC and TDS). In the ANFIS modeling, 70% of the data points were used for training, while the remaining data were used for testing. Table 2 presents the performances of various ANFIS models using 2–5 MFs. The optimum MF number was 3 for both EC and TDS modeling, which gave the lowest modeling errors and the highest R values. Each optimum MF was assigned to handle each input parameter.

There were 8 different types of MFs used to select the optimum MF type, as shown in Table 3. These types were: triangular MF (Trimf), trapezoidal MF (Trapmf), generalized bell curve MF (Gbellmf), Gaussian curve MF (Gaussmf), two-sided Gaussian MF (Gauss2mf), pi-shaped curve MF (Pimf), the composed difference between two sigmoidal MFs (Dsigmf), and the product of two sigmoid MFs (Psigmf). Table 3 compares the resulted errors (RMSE and MAE) and R value of applying the previously mentioned MF types for training,

testing, and overall datasets. In modeling the EC concentrations, the triangular MF gave the lowest errors in all the datasets and the best performance, and outperformed the other MF types. Meanwhile, for TDS concentrations, the Gaussian curve MF showed the lowest modeling errors for training, testing, and overall datasets.

Table 2. Performance of various models with different membership functions (MFs) for EC and TDS, respectively.

MF No.	Training Data			Testing Data			Overall Data		
	RMSE	MAE	R	RMSE	MAE	R	RMSE	MAE	R
2	30.6	21.88	0.899	37.31	26.76	0.884	32.77	23.35	0.895
3 *	24.59	17.41	0.936	41.34	28.71	0.861	30.61	20.81	0.909
4	22.26	14.66	0.948	357.3	70.02	0.179	196.9	31.33	0.301
5	19.74	12.2	0.959	789.8	185.4	0.105	433.7	64.36	0.170
2	19.91	13.08	0.877	14.11	10.83	0.940	18.9	12.63	0.890
3 **	16.88	11.6	0.913	16.08	12.31	0.924	16.72	11.74	0.915
4	13.57	8.436	0.945	48.95	22.98	0.723	24.98	11.33	0.848
5	11.18	6.271	0.963	184.1	56.03	0.149	82.7	16.17	0.395

* The best function numbers for EC. ** The best function numbers for TDS.

Table 3. Performance results for various MF types for EC and TDS, respectively.

MF Type	Training Data			Testing Data			Overall Data		
	RMSE	MAE	R	RMSE	MAE	R	RMSE	MAE	R
* Trimf	24.59	17.41	0.936	41.34	28.71	0.861	30.61	20.81	0.909
Trapmf	29.08	21.17	0.91	179.4	52.11	0.52	101.4	30.49	0.595
Gbellmf	25.75	18.81	0.93	183.8	50.67	0.312	103.1	28.4	0.503
Gaussmf	24.93	18.01	0.934	142.7	44.29	0.406	81	25.92	0.606
Gauss2mf	28.03	20.24	0.916	65.33	38.25	0.729	42.83	25.66	0.833
Pimf	31.39	23.07	0.894	879.8	125.5	0.306	483.5	53.92	0.265
Dsigmf	28.75	20.9	0.912	92.56	44.36	0.644	56.19	27.96	0.761
Psigmf	28.75	20.9	0.912	92.56	44.36	0.644	56.19	27.96	0.761
Trimf	16.64	11.42	0.916	22.24	14.4	0.868	17.9	12.01	0.904
Trapmf	19.74	14.71	0.879	22.67	16.56	0.858	20.36	15.08	0.873
Gbellmf	17.51	12.33	0.906	18.03	13.25	0.909	17.62	12.52	0.906
** Gaussmf	16.60	11.6	0.913	16.08	12.31	0.924	16.72	11.74	0.915
Gauss2mf	18.86	13.9	0.891	24.15	16.22	0.856	20.02	14.36	0.878
Pimf	20.51	15.22	0.869	36.37	21.11	0.725	24.5	16.39	0.819
Dsigmf	19.31	14.34	0.885	26.03	16.98	0.839	20.82	14.86	0.868
Psigmf	19.31	14.34	0.885	26.03	16.98	0.839	20.82	14.86	0.868

* The best MF type performance for EC.; ** The best MF type performance for TDS.

The selection of the optimum epoch number is a very significant factor in ANFIS modeling. Increasing the epoch number does not always mean enhancing the performance of ANFIS modeling. Usually, the modeling errors decrease by increasing the epoch number to a point, and then the errors increase afterward. Identifying this point is a necessity in ANFIS modeling. The previously selected ANFIS parameters and varying epoch numbers are displayed in Figure 7. From the plots, RMSE values vary for both the training and testing datasets with increases in the epoch number (until 50 epochs). The optimum epoch number for EC modeling was 3, while for TDS it was 3 or 10 epochs. Using these epoch numbers, they give the lowest modeling errors and also avoid the model's overfitting problem.

The full descriptions of the final ANFIS models in modeling the EC and TDS concentrations are listed in Table 4. In EC modeling, the optimum ANFIS structure consists of three triangular MFs (each MF presents one input), and 27 rules, and being trained for 3 epochs. While, for TDS modeling, the optimum ANFIS structure consists of three Gaussmf MFs (each MF presents one input), and 27 rules, and being trained for 3 or

10 epochs to prevent overfitting. The RMSE results were 30.61 and 16.72 for the EC and TDS modeling, respectively. However, the fitting results for R and R^2 were 0.909 and 0.827 for EC modeling and 0.915 and 0.838 for TDS modeling, respectively. The NSC results were very close to R^2, which demonstrates the accurate performance of the resulted models in simulating the desired parameters. Figure 8 demonstrates the assigned rules in the optimum ANFIS model structure for modeling (a) EC and (b) TDS level, respectively.

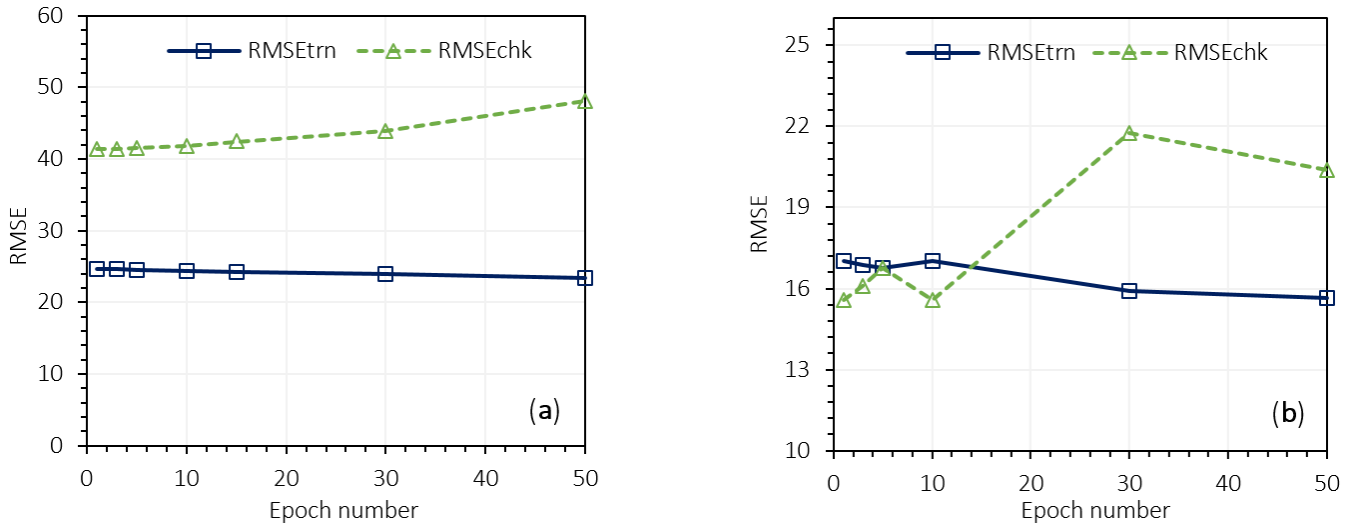


Figure 7. The optimum epoch number offered the lowest modeling error for (**a**) EC and (**b**) TDS modeling.

Table 4. Full description of the optimum ANFIS models for EC and TDS modeling.

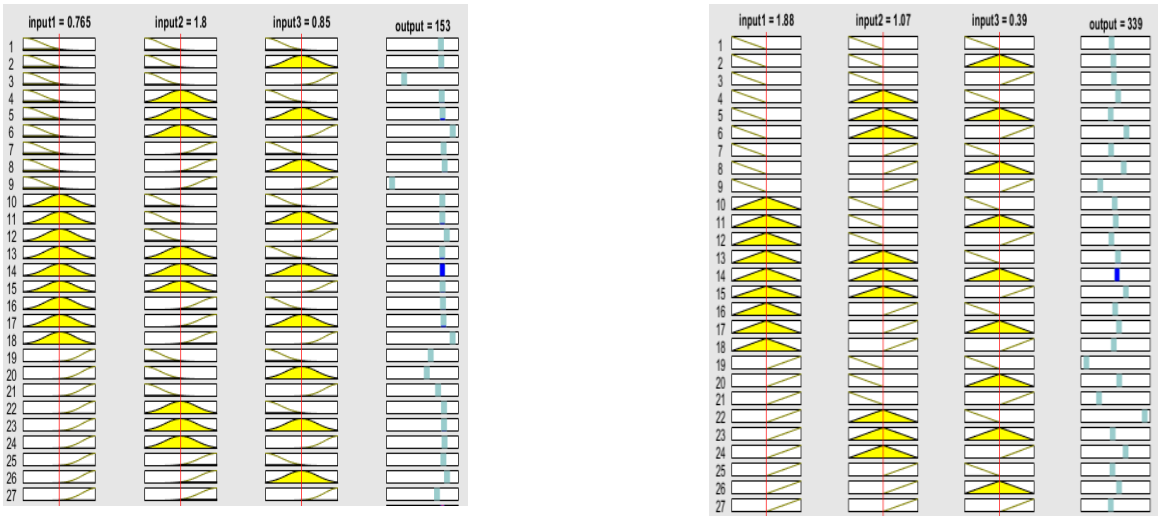
Description	EC	TDS
MF No.	3	3
MF Type	Trimf	Gaussmf
Rule No.	27	27
Optimum Epoch No.	3	3 or 10
R	0.909	0.915
R^2	0.827	0.838
RMSE	30.61	16.72
MAE	20.81	11.74
NSC	0.825	0.837

3.4. Model Statistical and Error Assessment

As discussed earlier, the efficiency of the developed ANFIS models was assessed in terms of statistical analysis and error assessment tests. Figure 9 illustrates the comparative evaluation of the observed and modeled simulated data. Moreover, Figure 10 shows the regression analysis between the observed and model-predicted results of EC and TDS concentrations by the selected structure of the ANFIS model. The modeling output shows an excellent correlation between the two datasets. An R^2 value above 0.9 was observed for both EC and TDS concentrations. Similarly, RMSE values below 20 µS/cm and 17 ppm were achieved by EC and TDS models, respectively. The modeling results show that the developed models are very efficient in modeling the surface water quality parameters, given the set of initial input parameters.

Besides statistical evaluation, the percent relative error (RE%) test was also conducted to check and demonstrate the accuracy of the proposed models. RE% plots are shown in Figures 11 and 12 for the optimum ANFIS model developed for EC and TDS, respectively. The results for both EC and TDS models show that the residual error of the data lies between +20% and −20%, describing the capacity of the developed ANFIS models for predicting the target output. Moreover, in both the modeling outputs, the max RE% results are under

60%, which indicates that, up to a limited extent, the ANFIS model underestimated the observed EC and TDS concentrations. However, the min RE% values are above −100% and −60% in the EC and TDS concentrations, respectively. The negative RE% results mean that the models overestimated the targeted salinity levels to a certain extent. Overall, the results exposed excellent accuracy of the models in predicting the EC and TDS.



(**a**) (**b**)

Figure 8. The 27 rules of the optimum ANFIS structure for (**a**) EC and (**b**) TDS modeling.

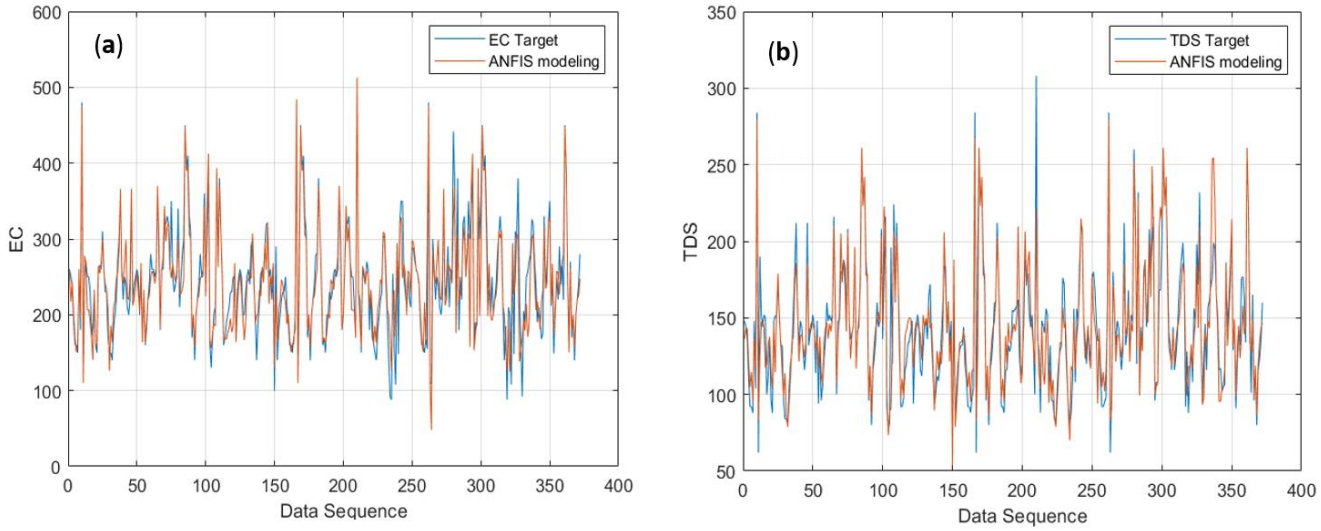


Figure 9. Comparison between the observed and modeled values for the (**a**) EC and (**b**) TDS data.

3.5. Variation of Model Output with Selected Inputs

As discussed previously (Section 3.2), the input optimization was applied to select the best and optimum combination of input parameters. The optimization process revealed that Combination 1 (Ca^{2+}, Na^{+}, and Cl^{-}) is the most optimum to model EC concentrations, while Combination 2 (Mg^{2+}, HCO_3^{-}, and SO_4^{2-}) are the most correlated to model the TDS concentrations. Figures 13 and 14, respectively, demonstrate the 3D surface plots of the selected input combinations and the target outputs. An increasing and fluctuating trend in Figure 13 can be observed for the EC concentrations, with the variation in input variables, that is, Ca^{2+}, Na^{+}, and Cl^{-}. Furthermore, a similar result can also be seen in

the TDS modeling (Figure 14), where the result shows a linearly increasing trend of TDS with all the input variables. The increasing tendency of the EC and TDS with all the input combinations of parameters may be attributed to the fact that both outputs, that is, EC and TDS, are directly related to the salt concentration of the water. Therefore, any change in the salts or ion concentrations in the water can directly affect levels of both the EC and TDS. The same trend of output was reported in many studies where the concentration of dissolved solids and conductivity were closely associated with ions and salt concentrations in water [17,20,50].

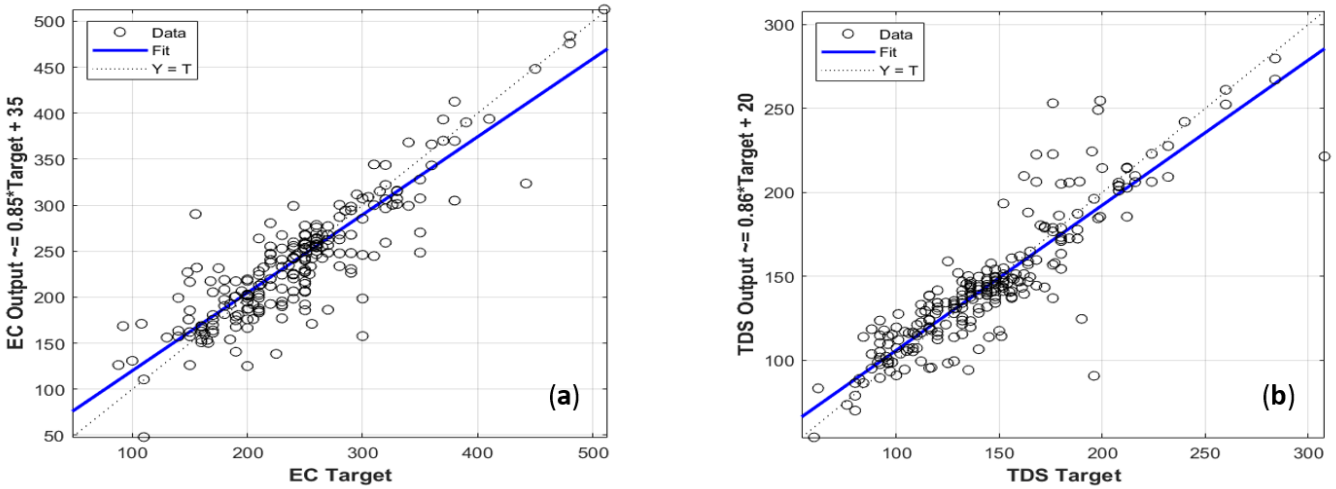


Figure 10. Regression curve between the observed and modeled datasets for (**a**) EC and (**b**) TDS.

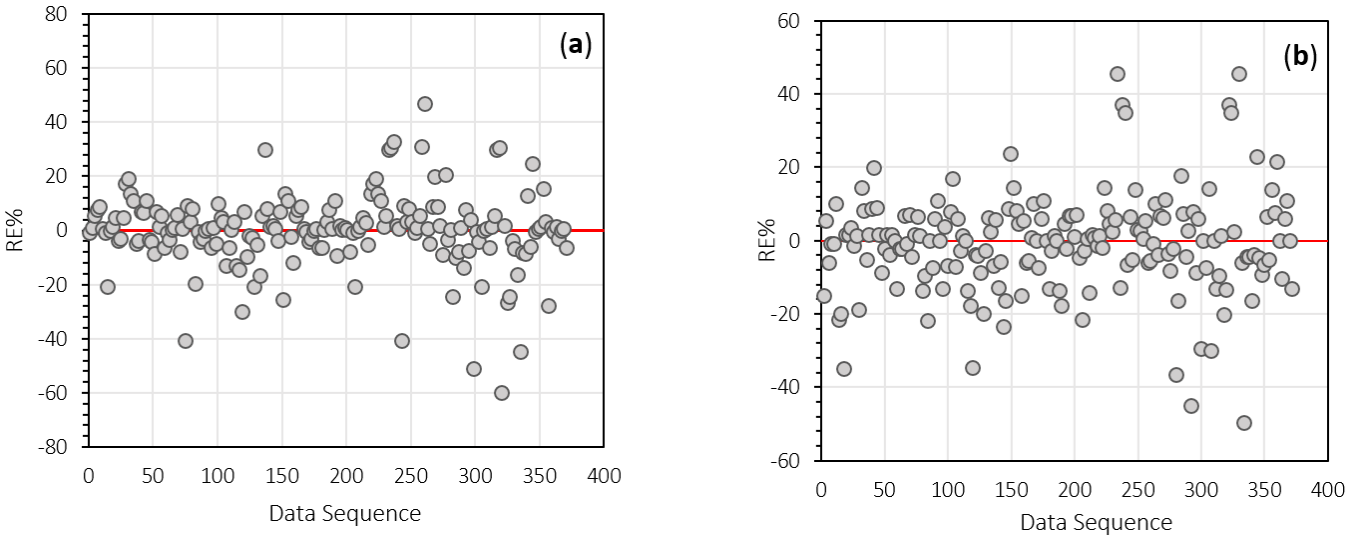


Figure 11. The percent relative error (RE%) of the final ANFIS model for the EC data: (**a**) training (**b**) testing.

3.6. Cross-Validation Output

The cross-validation method was used in accessing the ANFIS models for EC and TDS. The output is graphically illustrated in Figure 15 using RMSE and R as assessment criteria. A deviation in the validation output can be observed for the single-fold subclass. Nevertheless, the results demonstrate a good mean accuracy in the 10-folds. The average RMSE values obtained by the EC and TDS models were 3.8 µS/cm and 4.2 ppm, respectively. The mean R values accomplished by the EC and TDS models were 0.81 and 0.77, respectively. In the 10-folds, the minimum and maximum R values of 0.48 and 0.83 were respectively obtained during the model validation for TDS. The lowest RMSE value,

2.56 ppm, was accomplished for TDS in the second-fold subclass. Evidently, the output of the cross-validation method demonstrated efficient performance and generalized results of the ANFIS models, indicating that the model can accomplish good results on unseen data as well.

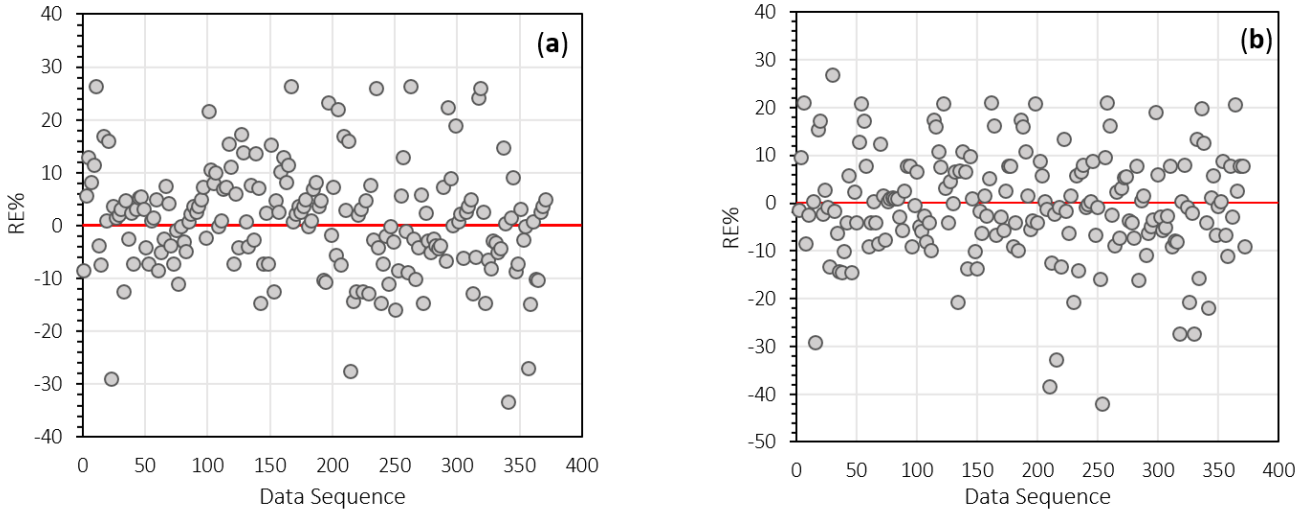


Figure 12. The percent relative error (RE%) of the final ANFIS model for the TDS data: (**a**) training (**b**) testing.

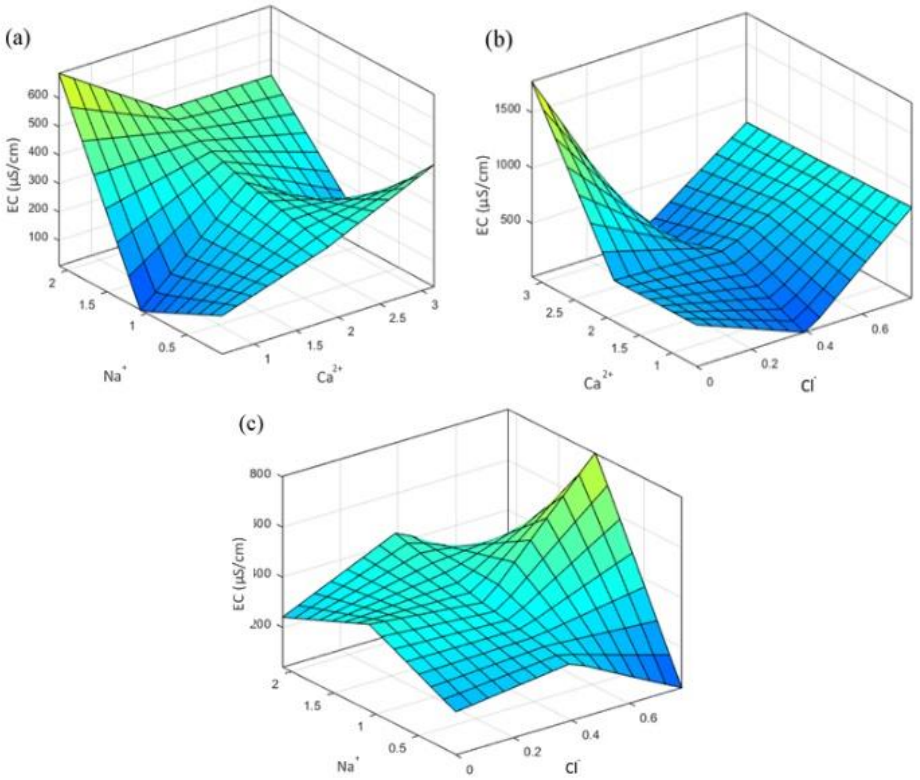


Figure 13. 3D surface plots and variation of EC with (**a**) Na^+ and Ca^{2+} (**b**) Ca^{2+} and Cl^- (**c**) Na^+ and Cl^-.

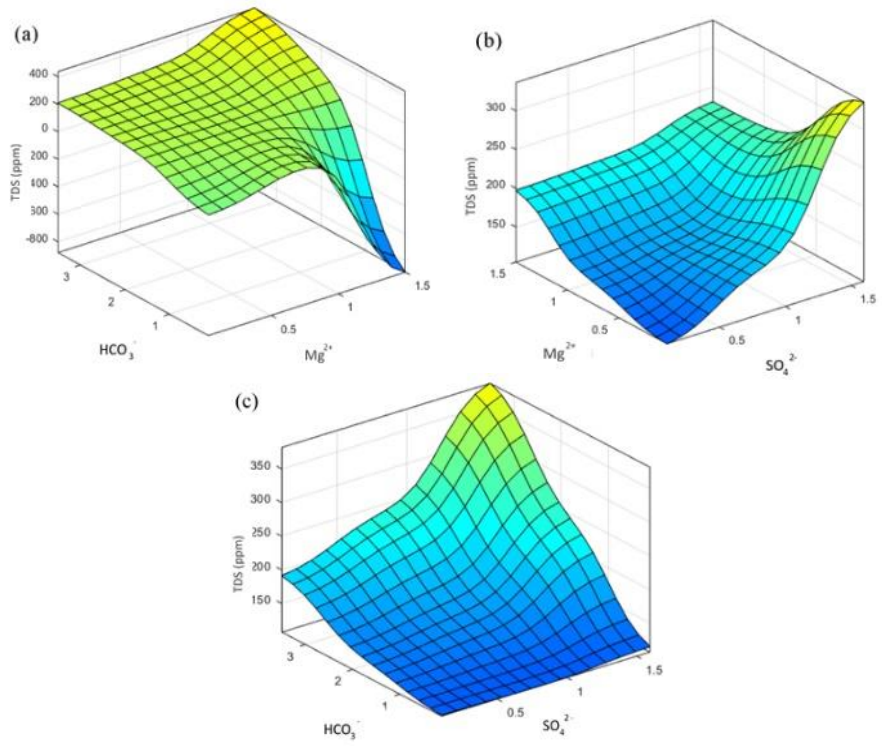


Figure 14. 3D surface plots and variation of TDS with (**a**) HCO_3^- and Mg^{2+} (**b**) Mg^{2+} and SO_4^{2-} (**c**) HCO_3^- and SO_4^{2-}.

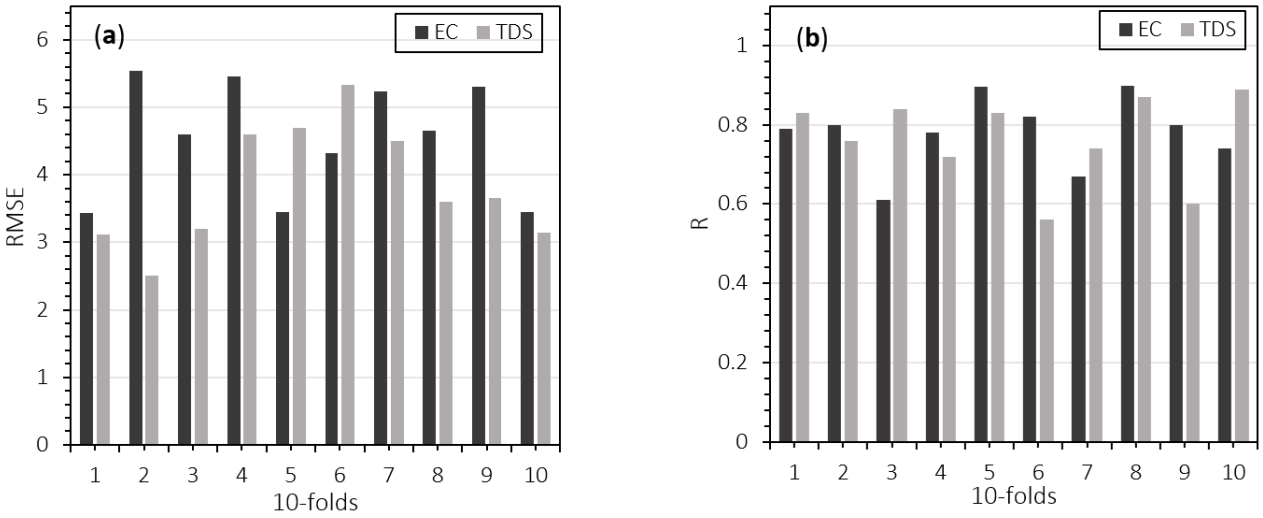


Figure 15. ANFIS models cross-validation outcomes in terms of (**a**) RMSE and (**b**) R.

4. Discussion

The focus of the present study, that is, data pre-processing, input optimization, and optimum ANFIS model development, has been presented in the previous sections. The majority of published work in modeling surface water quality parameters have used the standalone ANN, GEP, SVM, DT, RF, and regression-based models. Modeling and predicting water quality parameters with classical AI techniques cannot provide the desired outcomes. Therefore, it is essential to employ the modeling methods with optimization algorithms for effective and precise modeling outputs.

Based on comparative analysis among the current and previous studies that applied the ANFIS modeling and optimization technique, Al-Mukhtar et al. (2019) [20] reported that ANFIS performed better than ANN and regression model in predicting TDS and EC. Moreover, better results of the ANFIS model with the particle swarm optimization algorithm (PSO) were reported by Azad et al. (2019) [5] in predicting various water quality parameters. Khadr et al. (2017) [51] and Tiwari et al. (2018) [52] concluded that the ANFIS model is efficient to forecast phosphorus and nitrogen and other water quality parameters. Furthermore, Azad et al. (2018) [21] reported that the proficiency of the ANFIS model in modeling water quality parameters could be improved with optimization algorithms. Sun et al. (2019) [53] used the variable mode decomposition (VDM) and least square support vector machine (LSSVR) methods for outlier detection and correction in water quality data. The authors reported the accurate performance of the aforementioned methods and improved water quality data. Alameddine et al. (2010) [54] compared the performance of three outlier detection approaches, namely minimum covariance determinant (MCD), minimum volume ellipsoid (MVE), and M-estimators, in detecting and removing the outliers from lake water quality data. The results of the study revealed the M-estimators as a robust and flexible method in dealing with inconsistent water quality data.

The available literature shows that a limited number of studies utilized the data pre-processing and the ANFIS modeling, coupled with an exhaustive search for inputs, which was successfully integrated into this study. Consistent water quality datasets, optimized modeling inputs, and a computational efficient ANFIS structure could be achieved by adopting the methods used in this study. Moreover, the integrated optimization algorithms are more effective in providing robustness models with enhanced outputs than standalone ANN, SVM, GEP, RF, and other regression models.

5. Conclusions

In developing countries, the financial constraint and lack of facilities and infrastructure encourage further research to develop accurate and computationally efficient models that require a minimum number of parameters for surface water quality prediction. The current study reported the development and applications of the ANFIS modeling technique for surface water quality prediction, that is, EC and TDS, in one of the major rivers in Asia, the upper Indus River Basin. The data inputs were Ca^{2+}, Mg^{2+}, Na^{+}, Cl^{-}, $SO_4{}^{2-}$, $HCO_3{}^{-}$, pH, EC, and TDS, collected monthly over a period of 30 years (1975–2005). The specific outputs of this study are as follows;

- The two-sided outlier detection approach was found to be efficient in data pre-processing and outlier removal to get homogenous and consistent data records for modeling.
- The input optimization process reduced the modeling complexity by evaluating the optimum number of inputs, which is helpful in reducing data processing and collection efforts, and therefore highlighting the strong ability of the exhaustive search method to reduce noise in the data.
- The developed ANFIS model showed strong agreement with the actual data for training as well as testing data. The ANFIS model was able to model the quality of surface water efficiently using the selected inputs. This could be attributed to the structure of the ANFIS model, which incorporates the advantages of fuzzy reasoning and the self-learning capability of neural networks.
- Conclusively, the ANFIS model could be efficiently utilized in water quality assessments and mitigation studies.

Author Contributions: Conceptualization, data collection and analysis, writing original draft, M.I.S.; conceptualization, software modeling and visualization, writing original draft, T.A.; formal analysis and modeling, M.F.J.; supervision, review and editing, F.B. and A.M.; investigation and review, A.A.; methodology, review and editing, M.A.U.R.T.; Validation, proofreading, review, A.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the National Research Foundation of South Africa (Grant number: 84166).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data sharing not applicable.

Acknowledgments: Open Access Funding by the Publication Fund of the TU Dresden. Amir Mosavi would like to thank Alexander von Humboldt Foundation.

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

References

1. Liou, S.-M.; Lo, S.-L.; Wang, S.-H. A generalized water quality index for Taiwan. *Environ. Monit. Assess.* **2004**, *96*, 35–52. [CrossRef]
2. Najah, A.; El-Shafie, A.; Karim, O.A.; El-Shafie, A.H. Application of artificial neural networks for water quality prediction. *Neural Comput. Appl.* **2013**, *22*, 187–201. [CrossRef]
3. Iqbal, M.M.; Shoaib, M.; Agwanda, P.; Lee, J.L. Modeling approach for water-quality management to control pollution concentration: A case study of Ravi River, Punjab, Pakistan. *Water* **2018**, *10*, 1068. [CrossRef]
4. Maqbool, F.; Malik, A.H.; Bhatti, Z.A.; Pervez, A.; Suleman, M. Application of regression model on stream water quality parameters. *Pak. J. Agri. Sci.* **2012**, *49*, 95–100.
5. Azad, A.; Karami, H.; Farzin, S.; Mousavi, S.F.; Kisi, O. Modeling river water quality parameters using modified adaptive neuro fuzzy inference system. *Water Sci. Eng.* **2019**, *12*, 45–54. [CrossRef]
6. Nazari-Sharabian, M.; Taheriyoun, M.; Ahmad, S.; Karakouzian, M.; Ahmadi, A. Water quality modeling of Mahabad Dam watershed–reservoir system under climate change conditions, using SWAT and system dynamics. *Water* **2019**, *11*, 394. [CrossRef]
7. Yan, J.; Xu, Z.; Yu, Y.; Xu, H.; Gao, K. Application of a hybrid optimized BP network model to estimate water quality parameters of Beihai Lake in Beijing. *Appl. Sci.* **2019**, *9*, 1863. [CrossRef]
8. Mohammadpour, R.; Shaharuddin, S.; Zakaria, N.A.; Ghani, A.A.; Vakili, M.; Chan, N.W. Prediction of water quality index in free surface constructed wetlands. *Environ. Earth Sci.* **2016**, *75*, 139. [CrossRef]
9. He, K.; Yang, Y.; Yang, Y.; Chen, S.; Hu, Q.; Liu, X.; Gao, F. HYDRUS simulation of sustainable brackish water irrigation in a winter wheat-summer maize rotation system in the North China Plain. *Water* **2017**, *9*, 536. [CrossRef]
10. Kim, H.; Jeong, H.; Jeon, J.; Bae, S. Effects of irrigation with saline water on crop growth and yield in greenhouse cultivation. *Water* **2016**, *8*, 127. [CrossRef]
11. Velmurugan, A.; Swarnam, P.; Subramani, T.; Meena, B.; Kaledhonkar, M.J. *Water Demand and Salinity*; IntechOpen: London, UK, 2020.
12. Jamei, M.; Ahmadianfar, I.; Chu, X.; Yaseen, Z.M. Prediction of surface water total dissolved solids using hybridized wavelet-multigene genetic programming: New approach. *J. Hydrol.* **2020**, *589*, 125335. [CrossRef]
13. Sattari, M.T.; Joudi, A.R.; Kusiak, A. Estimation of Water Quality Parameters with Data–Driven Model. *J. Am. Water Work. Assoc.* **2016**, *108*, E232–E239. [CrossRef]
14. Ayers, R.S.; Westcot, D.W. *Water Quality for Agriculture*; Food and Agriculture Organization of the United Nations: Rome, Italy, 1985.
15. Tung, T.M.; Yaseen, Z.M. A survey on river water quality modelling using artificial intelligence models: 2000–2020. *J. Hydrol.* **2020**, *585*, 124670.
16. Bozorg-Haddad, O.; Soleimani, S.; Loáiciga, H.A. Modeling water-quality parameters using genetic algorithm–least squares support vector regression and genetic programming. *J. Environ. Eng.* **2017**, *143*, 04017021. [CrossRef]
17. Shah, M.I.; Javed, M.F.; Abunama, T. Proposed formulation of surface water quality and modelling using gene expression, machine learning, and regression techniques. *Environ. Sci. Pollut. Res.* **2020**, *28*, 13202–13220. [CrossRef] [PubMed]
18. Deng, W.; Wang, G.; Zhang, X. A novel hybrid water quality time series prediction method based on cloud model and fuzzy forecasting. *Chemom. Intell. Lab. Syst.* **2015**, *149*, 39–49. [CrossRef]
19. Abunama, T.; Othman, F.; Ansari, M.; El-Shafie, A. Leachate generation rate modeling using artificial intelligence algorithms aided by input optimization method for an MSW landfill. *Environ. Sci. Pollut. Res.* **2019**, *26*, 3368–3381. [CrossRef]
20. Al-Mukhtar, M.; Al-Yaseen, F. Modeling water quality parameters using data-driven models, a case study Abu-Ziriq marsh in south of Iraq. *Hydrology* **2019**, *6*, 24. [CrossRef]
21. Azad, A.; Karami, H.; Farzin, S.; Saeedian, A.; Kashi, H.; Sayyahi, F. Prediction of water quality parameters using ANFIS optimized by intelligence algorithms (case study: Gorganrood River). *KSCE J. Civil Eng.* **2018**, *22*, 2206–2213. [CrossRef]
22. Chen, L.; Jamal, M.; Tan, C.; Alabbadi, B. A Study of Applying Genetic Algorithm to Predict Reservoir Water Quality. *Int. J. Model Opt.* **2017**, *7*, 98. [CrossRef]
23. Qasem, S.N.; Samadianfard, S.; Sadri Nahand, H.; Mosavi, A.; Shamshirband, S.; Chau, K.W. Estimating daily dew point temperature using machine learning algorithms. *Water* **2019**, *11*, 582. [CrossRef]
24. Sarkar, A.; Pandey, P. River water quality modelling using artificial neural network technique. *Aquatic Procedia* **2015**, *4*, 1070–1077. [CrossRef]
25. Ghavidel, S.Z.Z.; Montaseri, M. Application of different data-driven methods for the prediction of total dissolved solids in the Zarinehroud basin. *Stoch. Environ. Res. Risk Assess* **2014**, *28*, 2101–2118. [CrossRef]

26. Asadollahfardi, G.; Zangooi, H.; Asadi, M.; Tayebi Jebeli, M.; Meshkat-Dini, M.; Roohani, N. Comparison of Box-Jenkins time series and ANN in predicting total dissolved solid at the Zāyandé-Rūd River, Iran. *J. Water Supply Res. Technol. AQUA* **2018**, *67*, 673–684. [CrossRef]
27. Zounemat-Kermani, M.; Seo, Y.; Kim, S.; Ghorbani, M.A.; Samadianfard, S.; Naghshara, S.; Singh, V.P. Can decomposition approaches always enhance soft computing models? Predicting the dissolved oxygen concentration in the St. Johns River, Florida. *Appl. Sci.* **2019**, *9*, 2534. [CrossRef]
28. Maroufpoor, S.; Fakheri-Fard, A.; Shiri, J. Study of the spatial distribution of groundwater quality using soft computing and geostatistical models. *Ish. J. Hydraul. Eng.* **2019**, *25*, 232–238. [CrossRef]
29. Aryafar, A.; Khosravi, V.; Zarepourfard, H.; Rooki, R. Evolving genetic programming and other AI-based models for estimating groundwater quality parameters of the Khezri plain, Eastern Iran. *Environ. Earth Sci.* **2019**, *78*, 69. [CrossRef]
30. Banadkooki, F.B.; Ehteram, M.; Panahi, F.; Sammen, S.S.; Othman, F.B.; Ahmed, E.S. Estimation of total dissolved solids (TDS) using new hybrid machine learning models. *J. Hydrol.* **2020**, *587*, 124989. [CrossRef]
31. Ali, K.F.; De Boer, D.H. Spatial patterns and variation of suspended sediment yield in the upper Indus River basin, northern Pakistan. *J. Hydrol.* **2007**, *334*, 368–387. [CrossRef]
32. Khan, A.; Richards, K.S.; Parker, G.T.; McRobie, A.; Mukhopadhyay, B. How large is the Upper Indus Basin? The pitfalls of auto-delineation using DEMs. *J. Hydrol.* **2014**, *509*, 442–453. [CrossRef]
33. Khan, A.J.; Koch, M. Correction and informed regionalization of precipitation data in a high mountainous region (Upper Indus Basin) and its effect on SWAT-modelled discharge. *Water* **2018**, *10*, 1557. [CrossRef]
34. Tahir, A.A.; Chevallier, P.; Arnaud, Y.; Neppel, L.; Ahmad, B. Modeling snowmelt-runoff under climate scenarios in the Hunza River basin, Karakoram Range, Northern Pakistan. *J. Hydrol.* **2011**, *409*, 104–117. [CrossRef]
35. Ul Hasson, S. Future water availability from Hindukush-Karakoram-Himalaya Upper Indus Basin under conflicting climate change scenarios. *Climate* **2016**, *4*, 40. [CrossRef]
36. Ali, S.; Li, D.; Congbin, F.; Khan, F. Twenty first century climatic and hydrological changes over Upper Indus Basin of Himalayan region of Pakistan. *Environ. Res. Lett.* **2015**, *10*, 014007. [CrossRef]
37. Hewitt, K. Glacier change, concentration, and elevation effects in the Karakoram Himalaya, Upper Indus Basin. *Mt. Res. Dev.* **2011**, *31*, 188–201. [CrossRef]
38. Ramzan, S.; Zahid, F.M.; Ramzan, S. Evaluating multivariate normality: A graphical approach. *Middle East J. Sci. Res.* **2013**, *13*, 254–263.
39. Ahmed, A.N.; Othman, F.B.; Afan, H.A.; Ibrahim, R.K.; Fai, C.M.; Hossain, M.S.; Ehteram, M.; Elshafie, A. Machine learning methods for better water quality prediction. *J. Hydrol.* **2019**, *578*, 124084. [CrossRef]
40. Heddam, S.; Bermad, A.; Dechemi, N. ANFIS-based modelling for coagulant dosage in drinking water treatment plant: A case study. *Environ. Monit. Assess* **2012**, *184*, 1953–1971. [CrossRef]
41. Ying, H. General SISO Takagi-Sugeno fuzzy systems with linear rule consequent are universal approximators. *IEEE Trans. Fuzzy Syst.* **1998**, *6*, 582–587. [CrossRef]
42. Tang, J. ANFIS: Adaptive network based fuzzy inference systems. *IEEE Trans. Syst. Cybern* **1993**, *23*, 515–520.
43. Shah, M.I.; Memon, S.A.; Khan Niazi, M.S.; Amin, M.N.; Aslam, F.; Javed, M.F. Machine Learning-Based Modeling with Optimization Algorithm for Predicting Mechanical Properties of Sustainable Concrete. *Adv. Civ. Eng.* **2021**, *2021*, 6682283.
44. Kohavi, R. *A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection*; IJCAI: Montreal, QC, Canada, 1995.
45. Shah, M.I.; Amin, M.N.; Khan, K.; Niazi, M.S.K.; Aslam, F.; Alyousef, R.; Javed, M.F.; Mosavi, A. Performance Evaluation of Soft Computing for Modeling the Strength Properties of Waste Substitute Green Concrete. *Sustainability* **2021**, *13*, 2867. [CrossRef]
46. Prasad, K.; Gorai, A.K.; Goyal, P. Development of ANFIS models for air quality forecasting and input optimization for reducing the computational cost and time. *Atmos. Environ.* **2016**, *128*, 246–262. [CrossRef]
47. Chen, S.; Hong, X.; Harris, C.J.; Sharkey, P.M. Sparse modeling using orthogonal forward regression with PRESS statistic and regularization. *IEEE Trans. Syst. Man Cybern. Part B* **2004**, *34*, 898–911. [CrossRef]
48. Khan, J.A.; Aelst, S.V.; Zamar, R.H. Building a robust linear model with forward selection and stepwise procedures. *Comput. Stat. Data Anal.* **2007**, *52*, 239–248. [CrossRef]
49. Wang, X. Sparse support vector regression based on orthogonal forward selection for the generalised kernel model. *Neurocomputing* **2006**, *70*, 462–474. [CrossRef]
50. Montaseri, M.; Ghavidel, S.Z.Z.; Sanikhani, H. Water quality variations in different climates of Iran: Toward modeling total dissolved solid using soft computing techniques. *Stoch. Environ. Res. Risk Assess.* **2018**, *32*, 2253–2273. [CrossRef]
51. Khadr, M.; Elshemy, M. Data-driven modeling for water quality prediction case study: The drains system associated with Manzala Lake, Egypt. *Ain Shams Eng. J.* **2017**, *8*, 549–557. [CrossRef]
52. Tiwari, S.; Babbar, R.; Kaur, G. Performance evaluation of two ANFIS models for predicting water quality Index of River Satluj (India). *Adv. Civ. Eng.* **2018**, *2018*, 8971079. [CrossRef]
53. Sun, G.; Jiang, P.; Xu, H.; Yu, S.; Guo, D.; Lin, G.; Wu, H. Outlier detection and correction for monitoring data of water quality based on improved VMD and LSSVM. *Complexity* **2019**, *2019*, 9643921. [CrossRef]
54. Alameddine, I.; Kenney, M.A.; Gosnell, R.J.; Reckhow, K.H. Robust multivariate outlier detection methods for environmental data. *J. Environ. Eng.* **2010**, *136*, 1299–1304. [CrossRef]