water

MDPI

Article

Multi-Expression Programming (MEP): Water Quality Assessment Using Water Quality Indices

Ali Aldrees [1], Mohsin Ali Khan [2,3], Muhammad Atiq Ur Rehman Tariq [4,5], Abdeliazim Mustafa Mohamed [1], Ane Wai Man Ng [6,7,*] and Abubakr Taha Bakheit Taha [1]

1 Department of Civil Engineering, College of Engineering, Prince Sattam Bin Abdulaziz University, Al-Kharj 16273, Saudi Arabia; a.aldrees@psau.edu.sa (A.A.); a.bilal@psau.edu.sa (A.M.M.); a.taha@psau.edu.sa (A.T.B.T.)
2 Department of Civil Engineering, CECOS University of IT and Emerging Sciences, Peshawar 25000, Pakistan; mohsin.ali@cecos.edu.pk
3 Department of Structural Engineering, Military College of Engineering (MCE), National University of Science and Technology (NUST), Islamabad 44000, Pakistan
4 College of Engineering and Science, Victoria University, Melbourne, VIC 8001, Australia; atiq.tariq@yahoo.com
5 Institute for Sustainable Industries & Liveable Cities, Victoria University, P.O. Box 14428, Melbourne, VIC 8001, Australia
6 College of Engineering, IT & Environment, Charles Darwin University, Darwin, NT 0810, Australia
7 Energy and Resources Institute, Charles Darwin University, Darwin, Ellengowan Dr, Brinkin, NT 0810, Australia
* Correspondence: Anne.Ng@cdu.edu.au

Abstract: Water contamination is indeed a worldwide problem that threatens public health, environmental protection, and agricultural productivity. The distinctive attributes of machine learning (ML)-based modelling can provide in-depth understanding into increasing water quality challenges. This study presents the development of a multi-expression programming (MEP) based predictive model for water quality parameters, i.e., electrical conductivity (EC) and total dissolved solids (TDS) in the upper Indus River at two different outlet locations using 360 readings collected on a monthly basis. The optimized MEP models were assessed using different statistical measurements i.e., coefficient-of-determination (R^2), root-mean-square error (RMSE), mean-absolute error (MAE), root-mean-square-logarithmic error (RMSLE) and mean-absolute-percent error (MAPE). The results show that the R^2 in the testing phase (subjected to unseen data) for EC-MEP and TDS-MEP models is above 0.90, i.e., 0.9674 and 0.9725, respectively, reflecting the higher accuracy and generalized performance. Also, the error measures are quite lower. In accordance with MAPE statistics, both the MEP models shows an "excellent" performance in all three stages. In comparison with traditional non-linear regression models (NLRMs), the developed machine learning models have good generalization capabilities. The sensitivity analysis of the developed MEP models with regard to the significance of each input on the forecasted water quality parameters suggests that Cl and HCO_3 have substantial impacts on the predictions of MEP models (EC and TDS), with a sensitiveness index above 0.90, although the influence of the Na is the less prominent. The results of this research suggest that the development of intelligence models for EC and TDS are cost effective and viable for the evaluation and monitoring of the quality of river water.

Keywords: water quality monitoring; electric conductivity (EC); total dissolved solids (TDS); machine learning (ML); non-linear regression modelling (NLRM)

Citation: Aldrees, A.; Khan, M.A.; Tariq, M.A.U.R.; Mustafa Mohamed, A.; Ng, A.W.M.; Bakheit Taha, A.T. Multi-Expression Programming (MEP): Water Quality Assessment Using Water Quality Indices. *Water* **2022**, *14*, 947. https://doi.org/10.3390/w14060947

Academic Editor: Dimitrios E. Alexakis

Received: 11 February 2022
Accepted: 15 March 2022
Published: 17 March 2022

1. Introduction

Surface water connected with rivers and streams is a key component in controlling the hydrological cycle, ecology, social wellbeing, and economic growth [1]. Environmental events, including annual precipitation and erosion, and also social elements like agricultural, urban, and industrial production procedures, all have an impact on river water

quality [2]. Surface water contributes as a dominant and major source of fresh water throughout the world. Its diminishment can have serious implications on the accessibility of drinking water, and quite significant consequences in terms of economic and financial growth and technological strategies [3]. The interplay of river systems and their surroundings, as well as the interchange of industrial, agricultural, and urban wastes across their trajectory results in water contamination [4]. The poor and contaminated water quality is a critical issue that poses a vulnerable situation to human welfare, agricultural, and the environment [2].

One of the major factors contributing towards the poor and contaminated water quality is the salinity. The saline environment of waterbodies has continuously grown over the last decade, compromising the quality of drinking water, and water utilized for irrigation, and in industrial processing units [5]. The salt accumulation caused by saltiness creates an adverse hydrological condition within water, inhibiting its household, agricultural and commercial consumption. Water quality assessment and saltwater monitoring and regulation is becoming crucial, and thus the stability between water supply and demand has met the required threshold [6]. One of the key parameters that can be used as a substantial indicator for the assessment of water quality is total dissolved solids (TDS), which can be used to determine its appropriateness for irrigation and drinking [7]. TDS is primarily made up of dissolved inorganic salts like sodium (Na^+), calcium (Ca^{2+}), magnesium (Mg^{2+}), nitrates ($NO_3{}^-$), chloride (Cl^-), and sulfate ($SO_4{}^{2-}$), as well as other dissolved organic particles. Higher level of salts and organic matter indicates substandard water quality [8].

Electric conductivity (EC) and TDS measurements have been the subject of scientific laboratory analysis and experimental procedures [7]. However, manual laboratory experiments, have several drawbacks, like the time required, their unreliable nature, and their inaccurate and ambiguous results due to the systematic errors and thus the lack of ability to be generalized [9]. Moreover, larger projects with passing time limit are inappropriate for labor-intensive tests. As a result, computer simulated models can be utilized to evaluate and forecast the water quality [10,11]. Numerous scientific publications have attempted to analyze a diverse collection of different water quality attributes employing numerical, stochastic, and mechanistic methods [12]. Such conventional means can still provide predictions for linear and homogeneous data sources [13]. The researchers also used other different techniques for the modelling of water quality indices, such as the driver-pressure-state-impact-response (DPSIR) method [14] and the advanced classification techniques i.e., the ground directive (GWD) and the water framework directive (WFD) [15]. Recently, machine learning (ML) based supervised models have been claimed to handle broad, non-linear, asymmetric, and complicated mechanisms of ecological and hydrologic systems, thereby avoiding the limitations of previous traditional models. Consequently, novel methods that are computationally fast, reliable, and effective for the determination and evaluation of EC and TDS are necessary. These advancements in measuring, analyzing, and monitoring water quality can benefit the environmental engineering sector.

During the last couple of decades, the sub-field of artificial intelligence (AI) i.e., machine learning (ML) has been widely used to tackle varied ecological technical challenges, especially water quality index simulation [2,16–19]. ML techniques are indeed a technological breakthrough in the development on management and surveillance of many engineering activities [20–22]. These algorithmic procedures can help to make appropriate forecasts without needing sophisticated programming officials. ML models are eventually supported by the collected data and the interpretation of configurations among data. These models are accomplished via the application of algorithms with the help of data subsets, including training and validation sets, and testing (unseen dataset) for evaluating the performance of predictive models [23–25].

The literature review shows a great intention towards the use of AI techniques for water quality prediction [26]. Tripathi and Singal [27] focused on the development of a novel water quality index (WQI) calculation approach for the Indian Ganges River. The authors used a principal components analysis (PCA) approach to lessen the total twenty-eight

variables to just the nine best combinations of explanatory parameters, which includes Hydrogen Power (pH), EC, TDS, Sulfate (SO_4^{2-}), Dissolved Oxygen (DO), Chlorine (Cl^-), Total Coliform (TC), Magnesium (Mg), and Biochemical Oxygen Demand (BOD). The use of only nine variables results in quicker calculations, consequently reducing the computational duration. Similarly, Zali et al. [28] used the ML technique i.e., artificial neural networks (ANNs) to investigate the impacts of six key explanatory variables (i.e., Chemical Oxygen Demand (COD), Suspended Solids (SS), Nitrate (NO_3^-), BOD, DO, and pH) for the computation of WQI. Determining the comparative relevance of every variable in WQI prediction using a sensitivity analysis showed that DO, SS, and NO_3^- are indeed the essential input variables. Nigam and SM [29] compared the prediction performance of fuzzy based models and conventional computation techniques for the calculation of WQI of ground water, reporting comparatively the outburst predictive power of fuzzy (an intelligent model). Thus, this categorizes the water quality and surpasses the predictions of conventional calculation techniques. Srinivas and Singh [30] extended their study to an Interactive Fuzzy model (IFM) for establishment of a unique fuzzy decision-making technique for predicting WQI in rivers. The results of their research show a considerable enhancement in WQI predictive accuracy in comparison with a conventional fuzzy approach. Yaseen et al. [31] investigated the estimation efficiency of adaptive-neuro-fuzzy-inference-system (ANFIS)-based hybrid models combined with subtractive clustering (SC), Fuzzy C-mean data clustering (FCM), and grid partitioning (GP). They found that ANFIS-SC is the best and most consistent model. Radial-basis-function-neural-networks (RBFN) and back-propagation-neural-network (BPNN) algorithms were used to propose a model for the establishment of the relation between WQI and many biological variables (like COD, SS, BOD, DO, Nitrate, and pH) in tropical and subtropical environments [32]. The RBFNN model produced comparatively good predictive outcomes. Bozorg-Haddad et al. [10] tested the performance of genetic-programming (GP) and least-square-support-vector-regression (LSSVR) for the estimation of K, Na, Mg, EC, SO_4, EC, TDS, and pH, in the Sefidrood River located in Iran. For all the computed models, the R^2 values is greater than 0.9, indicating good correlation. Al-Mukhtar and Al-Yaseen [33] used ANN, ANFIS and multiple-linear-regression (MLR) techniques to assess the water quality of the Abu-Ziriq River, located in Iraq. They forecasted the EC and TDS with most significant input variables (nitrate, chloride, calcium, magnesium, hardness and sulfate) and found that the ANFIS technique yielded the best outcomes. Sarkar and Pandey [34] used the ANN approach to analyze the amount of dissolved oxygen (DO) in river water over three distinct sites using four different variables i.e., pH, temperature, DO, and biochemical oxygen demand (BOD) and reported a correlation coefficient (R) value above 0.90 between the forecasted and actual DO data. Zhang et al. [35] used the combined hybridized model of ANN and GP algorithm to forecast the production of drinking water from chemical treatment plants. The findings showed that these created models performed well in forecasting the output capacity of the water treatment processing unit. Incorporation of additional data to the algorithm during training enhanced the model performance significantly. Chen et al. [36] utilized a comprehensive database to examine the water quality predictive ability of ten distinct ML models (three ensemble and seven conventional). The study indicated that utilizing a larger number of datapoints for water quality assessment can improve the predictive accuracy of the model. Some other prominent methodologies have been used effectively for different meteorological, environmental, and hydrological challenges (like rainfall forecasting), and these include tree-based algorithmic procedures, like random forest (RF), decision tree (DT) and support vector machine (SVM) models. These models are also recognized as a remarkable ML approach for both linear and complex non-linear engineering problems. Different researchers have utilized these algorithms with excellent predictive performance in a variety of scientific challenges [37]. Granata et al. [38] generated the SVM and RF model to forecast the content of TDS, TSS, BOD and COD, finding that the SVM model provided superior predictions. However, the efficacy is reduced when subjected to unseen data. In

brief, the various mathematical models are developed that contributed to the betterment of human life [39–46].

Although conventional AI algorithms rely on ANN and ANFIS, are extensively used for WQP modeling. Environmentalists have been investigating new resilient and robust intelligent algorithms [18,19]. It is worthy to mention that the neural networks work like a black box and do not consider any physical phenomenon of the issue being resolved. Most of the neural networks deliver a complex expression for the prediction of outcome on the basis of inputs [47]. In fact, ANN based models are considered as a correlation amongst the inputs and outputs, and the relation is either linear or relies on the pre-defined base functions [48,49]. Also, the tree-based algorithms are only good at capturing the linear relationship [50]. To overcome these issues, researchers used EA algorithms like GEP for simulation of water WQPs [51,52]. The EAs are advantageous in the condition where realistic and practical expression with higher generalization and prediction capabilities is required. However, the GEP is unable to incorporate the diverging data for the establishment of an ultimate model and must be excluded from the training and validation phase for the enhancement of the model's performance. Also, the GEP encode a single chromosome (expression) and present it as a program. Thus, they are only appropriate when there exists a simple relationship between inputs and outputs [53]. In contrast, the MEP is a comparatively new variant of genetic programming (GP) and has an ability to code a multiple number of chromosomes (expressions) in just one computer program [53]. It has the potential to predict the outcome accurately given the unknown complexity of the targeted parameter [54]. Unlike other ML algorithms, MEP does not need the identification of the final expression. In addition, the evolution process can effectively read and eliminate the complex mathematical errors from the resulting expression. The decoding procedure of MEP is fairly intuitive in comparison with other ML methods. Given the immense benefits of MEP algorithmic process over other evolutionary algorithms, it is scarcely adopted by environmentalists.

In the present research, water quality parameters like EC and TDS of the Upper Indus Basin (UIB) at the Bisham Qilla monitoring stations were modeled employing MEP algorithmic procedures and the traditional non-linear regression (NLR) approach based on the most affecting variables. A comprehensive dataset of 360 monthly readings taken from the Water and Power Development Authority (WAPDA) is being partitioned into three sets (training, validation, and testing) to verify the efficiency of the training process. To guarantee model efficacy, reliability and applicability, an in-depth statistical error test and sensitivity study is performed on the developed MEP models. The models that effectively predict TDS and EC concentrations by employing a small set of variables considerably minimize the trouble and expense involved in environmental surveillance.

2. Materials and Methods

This section is specifically explaining the methodology of multi-expression programming (machine learning approach) and a traditional non-linear regression (NLR) approach for the assessment of water quality indicators (i.e., EC and TDS) alongside the study area chosen in the current research and the performance evaluation indicators.

2.1. Multi-Expression Programming

Genetic algorithms (GA) are the metaheuristic and stochastic approach for searching for and optimizing complex solutions depending upon the evolutionary genetics and biological selection principles [55]. GA generates a sequence of strings (binary) to describe the outcome by employing the conventional optimization methods. Genetic programming (GP) was established as an improved version of GA to generate string expressions into computer algorithms like tree construction or workable computer code [56]. GP is a symbolized optimization approach that employs Darwin's theory of evolution to resolve an issue using complex algorithms. The fundamental purpose of GP is just to find a code that combines the given inputs and known outcome using the fitness metric. In general, there are three

main forms of GP: linear-based GP, graph-based GP, and tree-based GP [57,58]. In comparison to the other two forms of GP, linear-GP is much more effective since it does not require sophisticated or fast analyzers. This results in a more relevant significance for such linear-GP, which improves the accurateness in actual timescales.

To forecast the water quality indices (i.e., EC and TDS), a relevant linear-GP technique known as multi-expression-programming (MEP) was used in this research, depending upon it's overall precision and reliability. The MEP records responses using linear chromosomes. A chromosome can contain many computer algorithms (productive solutions). The optimal recorded answer to reflect the chromosome is identified from a comparison of the fitness scores of the scripts (programs). The MEP algorithmic procedure begins with the creation of a randomized population of computer code. To design the best solution to the considered problem, the subsequent stages are performed until the exit condition is reached [53]:

1. Two individuals (parents) are chosen employing the simple binary tournament process, subsequently reconstituted using a specified crossover probability.
2. The two parents are then recombined to generate two offspring.
3. After the mutation of the offspring, the weakest individual is substituted with the finest among them within the present population.

MEP is represented in the same manner as C and Pascal compilers translate mathematical equations to coding [57]. A sequence (String or series) of formulas represents the MEP genes. The set of genes define the code length (chromosomal length), which remains unchanged across the computation. The gene is constituted with a function sign (component in a function set i.e., F) and one or two terminals (component terminal set i.e., T). To produce the contextually valid code, the initial gene of a chromosome should correspond to a terminal selected at random within a terminal set. The function gene must hold a reference to the function parameters. The generated terminal scores of a particular gene have lesser values than the gene's chromosomal location. The following example illustrates a simple MEP chromosome:

$$Y_0 : A_1$$

$$Y_1 : A_2$$

$$Y_2 : Pow\ A_0, A_1$$

$$Y_3 : A_3$$

$$Y_4 : /A_2, A_3$$

$$Y_5 : A_4$$

$$Y_6 : +A_4, A_5$$

In the given example, the function set $F = \{Pow, +, /\}$ and the terminal set $T = \{A_1, A_2, A_3, A_4\}$ are utilized to encode and MEP program. The MEP genes can be converted into a computational program with the decoding of the chromosomes right from the top to the bottom. The respective program is presented in the form of gene trees in Figure 1, where the 0, 1, 3 and 5 number genes are used to code one unique terminal i.e., $Y_0 = A_1$, $Y_1 = A_2$, $Y_3 = A_3$, and $Y_5 = A_4$, respectively. At chromosomal positions 0 and 1, gene 2 denotes the 'Pow' function upon the said operands, thus encoding the expression $Y_2 = A_1^{A_2}$. Likewise, at positions 2 and 3, gene 4 represents the operator $'/'$ upon the operands and encodes the expression as $Y_4 = A_1^{A_2}/A_3$. Eventually, $Y_4 = A_1^{A_2}/A_3 + A_4$ can be stated as a result of gene 6. Consequently, the chromosomal combination can be visualized as a forest of gene trees (see Figure 1), having several expressions. The optimal expression tree is determined following regulation of the efficiency of various expressions on an MEP chromosome [53].

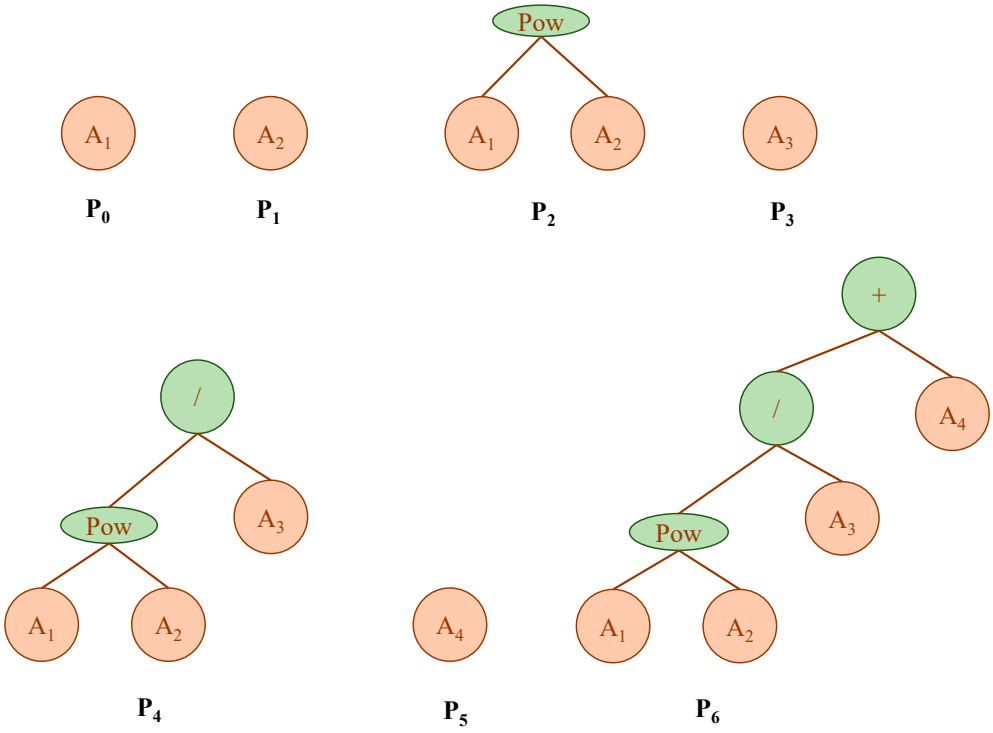


Figure 1. Trees (Genes) in MEP representing the expression using chromosome.

2.2. Non-Linear Regression Approach

Linear regressions employ a linear function for fitting the recorded data and to identify the relation among two or more variables and a single output parameter. Each record of the predictor variables corresponds to a recorded value of the outcome parameter. The following is a generalized version of multiple-linear regression:

$$Y = B_o + B_1X_1 + B_2X_2 + , \ldots , + B_nX_n \quad (1)$$

However, in cases when several dependent variables are non-linear, the logarithm based exponential adjustment can be employed to perform regression (see Equation (2)), which is then reverted for the prediction of the outcome using an antilogarithmic function (see Equation (3)).

$$\log(Y) = \log(B_o) + B_1 \log(X_1) + B_2 \log(X_2) + , \ldots , + B_n \log(X_n) \quad (2)$$

$$Y = B_o + X_1^{B_1} + X_2^{B_2} + , \ldots , + X_2^{B_2} \quad (3)$$

where $B_0, B_1, B_2, \ldots, B_n$ denotes the coefficient (constants) and $X_0, X_1, X_2, \ldots, X_n$ are the independent (input) variables. The above equation represents a multi-variable power expression which better captures the functional interdependency between numerous variables, and thus provides more realistic results [59].

The present research discusses the findings of the MEP (ML technique) and NLRM (traditional regression approach) established models to assess water quality indicators (i.e., EC and TDS). A comparison of outcomes from both methodologies was also performed to determine and confirm the superiority of the ML-based MEP approach. It is a widely adopted approach having significant implications in a variety of technological sectors [60–62]. The NLRM implemented in this research was created with the help of statistical software (i.e., SPSS: statistical package for social sciences), while the MEPX V.1.0. software is used for MEP modeling.

2.3. Study Area and Data Collection

The Indus River runs for 2880 km, with a mountainous, snow fed and glacierized catchment of the Upper Indus River basin (UIB) [63]. The UIB is located at the top of the Tarbela reservoir, which is basically associated with the Indus basin system, having a length and drain area of 1150 km and 165,400 km^2, respectively. The ice deposition of 2174 km^3 also exists at the same location. The height of UIB ranges between 455 to 8611 m, with the amount of annual precipitation between 100 and 200 mm, which changes the climatic condition within the basin proportional to the elevation [51,63].

The Pakistani Water and Power Development Authority (WAPDA) provided the dataset related to water quality which is utilized in this investigation. The finalized data set included 360 monthly instances recorded between 1975 and 2005 at the outlet location of Bisham Qilla and Doyian. The eight most prominent explanatory variables selected are the water temperature (°C), magnesium (Mg), calcium (Ca), sodium (Na), sulphate (SO_4), chloride (Cl), pH, and bicarbonates (HCO_3),, along with two well-known outputs i.e., EC and TDS. Table 1 presents the key descriptive statistical attributes of the collected data, which includes mean, maximum and minimum values and the dispersion statistics like kurtosis, skewness and standard deviation. TDS levels in the following research vary considerably between 60 and 524 ppm, whereas EC readings range from 88 to 770 µS/cm. According to WHO recommendations, the acceptable limit of TDS in drinkable water is 300 to 600 mg/L, whereas the authorized level in agricultural water is 450 to 2000 mg/L [7]. Table 1 presents that the concentrations of EC and TDS are both inside the allowable range; nonetheless, it is essential that significant water quality indices be measured precisely and without substantial work. Also, the kurtosis and skewness lie in the permissible range of [–10, 10] and [−3, +3], respectively, showing an acceptable dispersion and peakedness in the model parameters [64,65].

Table 1. Descriptive statical measures of collected data.

Independent and Dependent Variables	Descriptive Statistics						
	Range	**Minimum**	**Maximum**	**Mean**	**Std. Deviation**	**Skewness**	**Kurtosis**
Ca	103.39	0.61	104.00	1.82343	5.420	1.26	3.22
Mg	2.61	0.03	2.64	0.6149	0.343	1.92	4.42
Na	8.95	0.05	9.00	0.5427	0.672	1.86	4.44
HCO_3	7.29	0.11	7.40	1.7404	0.697	1.24	2.66
Cl	4.20	0.00	4.20	0.2968	0.312	1.25	5.80
SO_4	3.10	0.10	3.20	0.5758	0.383	2.18	1.87
PH	8.40	0.00	8.40	7.8413	0.621	−1.86	4.56
WT (°C)	20.11	1.00	21.11	12.1813	3.812	−0.32	−0.75
TDS (ppm)	464	60	524	148.35	62.231	4.584	37.225
EC (µS/cm)	690	88	770	250.42	98.691	3.801	31.808

Tables 2 and 3 show the correlation coefficient among the model variables and the model outcomes (i.e., EC and TDS). As per the current research, incorporating so many input variables having a poor relationship with the model outcome degrades model efficiency and enhances complication and computation complexity [66]. The inputs considered in this research have a strong relation with the concerned output. Also, no multicollinearity problem is observed, as the relation amongst the inputs are less than 0.8 [67,68]. As is customary, the data were divided randomly into three distinct sets i.e., 70% (252 recordings) for training, 15% (54 readings) for validation and 15% (54 readings) for testing the developed models.

Table 2. Correlation matrix of input variables corresponding to TDS.

Pearson's Correlation Matrix	Ca	Mg	Na	HCO_3	Cl	SO_4	pH	WT (°C)	TDS (mg/L)
Ca	1.000								
Mg	0.315	1.000							
Na	0.398	0.332	1.000						
HCO_3	0.626	0.482	0.594	1.000					
Cl	0.446	0.380	0.490	0.449	1.000				
SO_4	0.381	0.444	0.454	0.178	0.198	1.000			
pH	−0.133	0.116	−0.005	0.023	−0.051	−0.018	1.000		
WT (°C)	−0.464	−0.390	−0.249	−0.296	−0.239	−0.316	0.021	1.000	
TDS (mg/L)	0.704	0.619	0.759	0.769	0.587	0.546	−0.630	−0.664	1.000

Table 3. Correlation matrix of input variables corresponding to EC.

Pearson's Correlation Matrix	Ca	Mg	Na	HCO_3	Cl	SO_4	pH	WT (°C)	EC (µS/cm)
Ca	1.000								
Mg	0.315	1.000							
Na	0.398	0.332	1.000						
HCO_3	0.626	0.482	0.594	1.000					
Cl	0.446	0.380	0.490	0.449	1.000				
SO_4	0.381	0.444	0.454	0.178	0.198	1.000			
pH	−0.133	0.116	−0.005	0.023	−0.051	−0.018	1.000		
WT (°C)	−0.464	−0.390	−0.249	−0.296	−0.239	−0.316	0.021	1.000	
EC (µS/cm)	0.715	0.601	0.728	0.767	0.555	0.508	−0.616	−0.649	1.000

2.4. Statistical Indicators for Response Evaluation of Models

The functionality of the simulated models in the training or testing phase is monitored by estimating statistical measures like mean absolute percent error (MAPE), mean root mean squared logarithmic error (RMSLE), root mean square error (RMSE), coefficient of determination (R^2) (also known as root square value) and mean absolute error (MAE). R^2 is recognized to be the superior amongst these for examining the effectiveness of the models. The R^2 score from 0.65 to 0.75 implies outstanding performance, whereas below 0.50 indicates poor performance [69]. The formula used to get the R^2 value is given in Equation (4), where, the 'P' and 'E' denotes the model predicted and experimental findings, respectively. And 'm' is the total number of readings.

$$R^2 = 1 - \frac{\sum_{i=1}^{m}(P_i - E_i)^2}{\sum_{i=1}^{m}\left(P_i - \overline{P}_i\right)^2} \tag{4}$$

MAE represents the imbalances between estimated and observed results in terms of the mean of the absolute magnitude of the error (removing the negative sign) in the same unit as the output. It captures the magnitude of lower error values effectively [70] and is computed using the provided Equation (5).

$$MAE = \frac{\sum_{i=1}^{m}|P_i - E_i|}{m} \tag{5}$$

RMSLE helps to determine the proportional difference among the forecasted and observed result by incorporating the enlarged logarithmic error. It is useful for right-skewed outcomes because the log conversion depicts the targeted distribution quite intuitively. The Equation (6) shows the expression used for the calculation of *RMSLE*.

$$RMSLE = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(\log(E_i + 1) - \log(P_i + 1))^2} \tag{6}$$

An *RMSE* is used for the effective assessment of the magnitude of larger error values and significant deviations, like outliers, which were weighted more heavily [70]. It measures the error by considering the root of squared error. The the value of *RMSE*, the better the effectiveness and performance of the simulated model will be. It is computed using Equation (7).

$$RMSE = \sqrt{\frac{\sum_{i=1}^{m}(E_i - P_i)^2}{m}} \quad (7)$$

Furthermore, the mean absolute percent error (*MAPE*) depicts the percentage of absolute error (positive or negative) values. It is the best indicator to categorize the statistical model based on performance. Using *MAPE*, the predictive performance of a model can be considered excellent ($0\% \leq MAPE \leq 10\%$), good ($10\% < MAPE \leq 20\%$), acceptable ($20\% \leq MAPE < 50\%$), or inaccurate ($50\% \leq MAPE$) [65].

$$MAPE = \frac{1}{m}\sum_{i=1}^{m}\left|\frac{E_i - P_i}{E_i}\right| \quad (8)$$

2.5. Tunning the Modeling Hyper-Parameters

To produce robust, reliable and adaptable models, certain MEP hyper-parameters are set up before starting the modeling. Using the existing suggestions in the literature and a hit-and-trial process, the correct parameters are set [71]. The population size is used to determine the total programs that will emerge in the population. A larger population size can take a long time until convergence is achieved, and it may be complex but will also be realistic and accurate. However, the rise in size above a particular threshold led to the overfitting problem in the developed model. Table 4 shows the optimized hyper-parameters that were chosen for the two developed models (EC and TDS) produced in the current study. The modelling began with the assumption of a total of 10 sub-population. Initially, a simple arithmetic operator (i.e., $+$, $-$, $\times$, and $\div$) was selected to obtain a simple final formulation that was later on extended to a logarithmic and square root function to achieve the convergence at a higher degree of accuracy and robustness. The iterations are also used to specify the accurateness that the algorithm must attain before being terminated. A simulation having several iterations might result in a model with the least inaccurate results. Consequently, the crossover and mutation speed are the measures to control the likelihood of the offspring currently undergoing these genetic activities. Considering these probabilities, the crossover rate varies between 50% and 95%. After testing multiple combinations of these parameters, the best possible options were chosen depending on the model performance, as presented in Table 4. Performance overfitting of the developed model is the frequently occurring challenge with machine learning modeling. A model appears to work successfully on the sample data provided, however its efficacy suffers dramatically when extended to new unseen data. To solve the discussed problem, the literature suggested that the developed (trained) model must be tested on the testing dataset (fresh unseen) [72]. Thus, the entire data collected in the database was arbitrarily partitioned among three distinct sets i.e., training (70%–252% reading), validation (15%–54% readings), and testing (15%–54% readings). It was verified that the data distribution was consistent across all sets. Both the training and validation data were utilized in the computation of the optimized model. The performance of the validated model was further evaluated on the unseen data i.e., the testing data (third set). The resulting models outperformed in all three stages. A publicly accessible computer tool (i.e., MEPX v1.0), was used to execute the MEP algorithmic process.

Table 4. The optimized hyper-parameters of MEP model developed for EC and TDS.

Hyper Parameters	Optimized Setting
Operators	Addition, subtraction, multiplication, division, Sqrt, Ln
Num subpopulations	50
Subpopulation size	250
Code length	50
Crossover probability	0.9
Crossover type	uniform
Mutation probability	0.01
Tournament size	2
Operators probability	0.5
Variables probability	0.5
Num generations	1000

The method initiates with the generation of a population containing acceptable solutions. The operations are repetitive, with every iteration bringing us closer to an optimal solution. Within the whole solution population, the efficiency of every iteration is assessed. The process will keep evolving unless the stagnation of pre-defined fitness function (i.e., RMSE or R^2) is reached. In case the model findings for each set of data are inaccurate, the procedure is continued again, steadily boosting the size and number of subpopulations. The optimal result is then chosen depending on the lowest RMSE and highest R^2. It was noted that the accuracy of certain models in the training phase was better than in the testing phase, indicating the overfitting, which must be prevented. It is worthy of mention that the evolution time and generations have an influence on the accuracy of the produced models. A model would keep developing endlessly with these types of algorithms owing to the inclusion of additional variables into the process. However, in the current work, the modeling was terminated after a thousand generations or when the improvement in fitness value falls below 0.1%. Furthermore, an optimum solution must fulfill several performance metrics, as discussed in Section 2.4.

3. Results and Discussions

3.1. Formulation of EC and TDS Using MEP Model

The generated optimized MEP code for the future prediction of EC and TDS using nine variables is presented in Appendix A as A-1 and A-2, respectively. The experimental and modeling results of EC of all three sets is plotted in Figure 2 along with the slope. The ideal position of the regression line is 45°, i.e., slope equaling 1, while for strong correlation it must be nearer to 1 [70]. As depicted in Figure 2, the slope for all three sets is 0.9885 (training), 0.9921 (validation), and 0.9918 (testing), indicating a substantial relationship among the experimental and resulted modeling values of EC. Furthermore, the results seem fairly comparable to one another and approach a perfect fit for the sets, implying the proper training of the model [65]. Thus, it possesses a strong generalization potential, i.e., it functions pretty similarly on new data (testing set).

A corresponding assessment was also undertaken for the TDS findings, which are presented in Figure 3. It is quite clear and obvious that the generated model for TDS has been trained well using the input data to accurately follow and capture the experimental TDS. The regression line slope for all three sets is 0.9924 (training), 1.0119 (validation), 0.9989 (testing). Like an EC model, the TDS model also functions exceptionally well in the testing phase [73]. Thus, this demonstrates that the problem of over-fitting was addressed effectively [70].

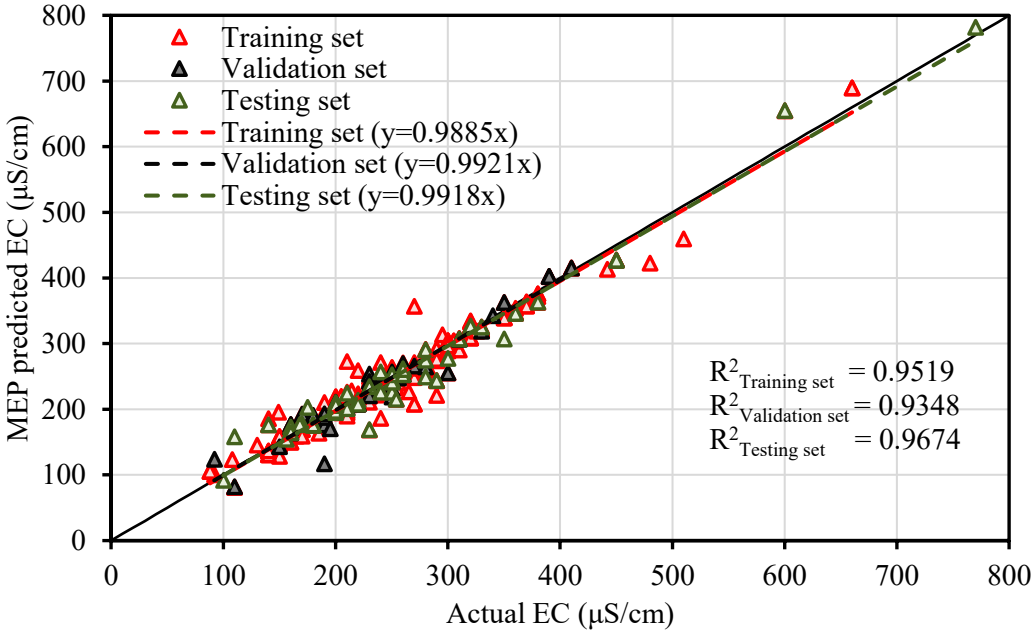


Figure 2. Regression plot of MEP models established for EC.

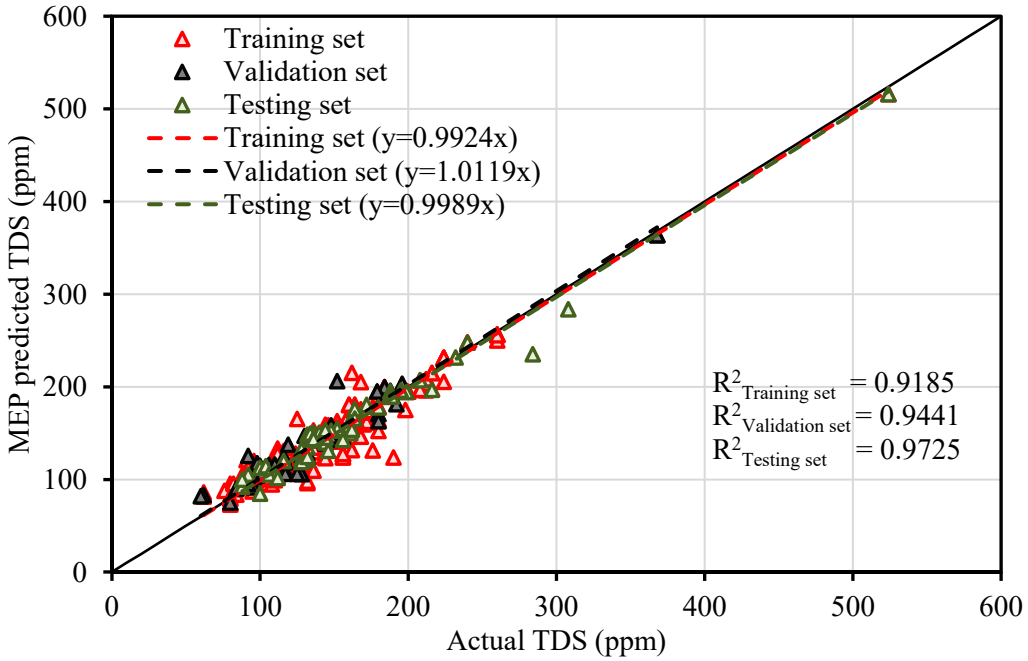


Figure 3. Regression plot of MEP models established for TDS.

Furthermore, the R^2 of both EC and TDS MEP models is above 0.9 in each stage i.e., for training, validation and testing. It can be seen clearly from Figures 2 and 3, the scatter of datapoints is near the 45° lines (1:1), resulting in the higher prediction accuracy of the suggested EC-MEP and TDS-MEP models, respectively. The R^2-values for the EC model are 0.9519 (training), 0.9348 (validation) and 0.9674 (testing), while for the TDS model they are 0.9185 (training), 0.9441 (validation) and 0.9725 (testing). Both of the developed MEP models shows a generalized performance, as the measures in each stage are closer with a little difference [74].

Additionally, the number of data points employed for modeling has a substantial significance upon the efficiency as well as adaptability of these simulation equations [75].

Inside the compiled database, the maximum points i.e., 360 were picked for both EC and TDS, thus, resulting in a higher accuracy with minimal errors [75].

3.2. Overall Performance of Developed Models

The reliability of a model is directly related to the number of instances/readings being used in the establishment of the model. It is mentioned in the literature that the instances to independent variables ratio must exceed five [76]. For both the EC and TDS, this ratio is 31.5 and 6.75 for the training and validation/testing set, respectively. An R^2-score consistently above 0.8 indicates a good association amongst measured and model predicted results [77]. Both EC and TDS are significantly associated with all the selected input variables. Conversely, the investigations revealed that R^2 identifies the linear relationship of outcome and independent factors. Therefore, evaluating the presented models (EC and TDS) just on the slope or inclination of the trendline and the regression coefficient is inadequate [76]. Thus, multiple statistical measures are used to examine the robustness and reliability of the generated MEP models.

3.2.1. EC-MEP MODEL

To evaluate the robustness of the developed MEP-EC model, the statistical error measures (i.e., MAE, RMSE, MAPE, and RMSLE) of all three sets is presented in Table 5. The MAE, RMSE and RMSLE are graphically interpreted via the mean absolute error plot (Figure 4), while the absolute percent error histogram (Figure 5) is presented for the examination of MAPE. The MAE and RMSE have their own significance in evaluation of model performance. The RMSE gives higher weight to larger error values as the error is squared before taking the average. Conversely, MAE is significant in capturing the smaller error values effectively and is thus always smaller than RMSE. The MAE in training and validation are 12.36, and 12.14, respectively, while the RMSE values are 18.54, and 17.19, respectively, with slightly better performance in the testing stage having MAE and RMSE equaling 11.12 and 16.43, respectively, thus fulfilling the stated condition (MAE < RMSE) [37]. Considering the whole data set of 360 readings, the minimum and maximum values of the absolute error are 0.016 and 98.16, respectively. Additionally, the RMSLE also nearly equals to zero at each stage, replicating the outburst functioning as it penalized the larger measured/prediction values efficiently. Furthermore, the percent error histograms show that 282 readings (78.33% of data) have an absolute error below 20%, with no prediction having an error above 40%. The MAPE values fall below 10%, i.e., 5.317%, 6.145% and 6.119% for training, validation and the testing phase, respectively. In accordance with the criteria explained in Section 2.4, the developed EC-MEP model can be categorized as "excellent" [7,8]. The developed model can be effectively used for future prediction with higher accuracy and minimal error measures, thus assisting the practitioners in avoiding human and machine errors [26,35].

Table 5. Statistical error measurement of non-linear regression models developed for EC and TDS.

Statistical Measure	MEP-EC			MEP-TDS		
	Training	**Validation**	**Testing**	**Training**	**Validation**	**Testing**
R^2	0.9519	0.9348	0.9674	0.9186	0.9442	0.9725
MAPE	5.317	6.145	6.119	7.40	8.25	5.54
MAE	12.36	12.14	11.22	9.75	9.80	8.27
RMSE	18.54	17.19	16.43	13.36	13.33	11.36
RMSLE	0.000582	0.000469	0.000402	0.000868	0.00033	0.00031

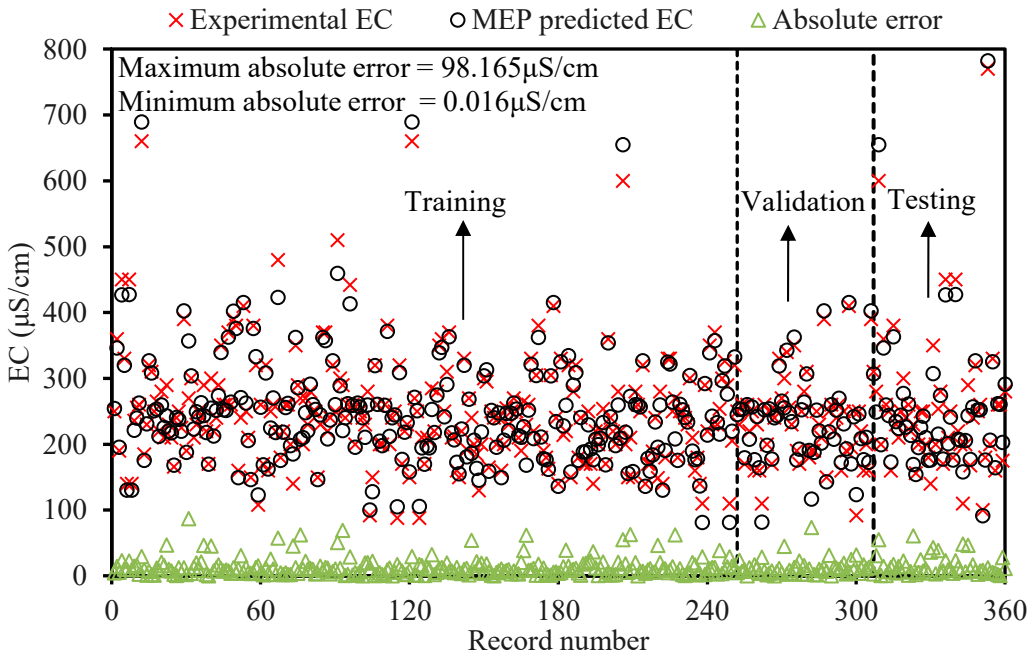


Figure 4. Absolute error plot of MEP model established for EC.

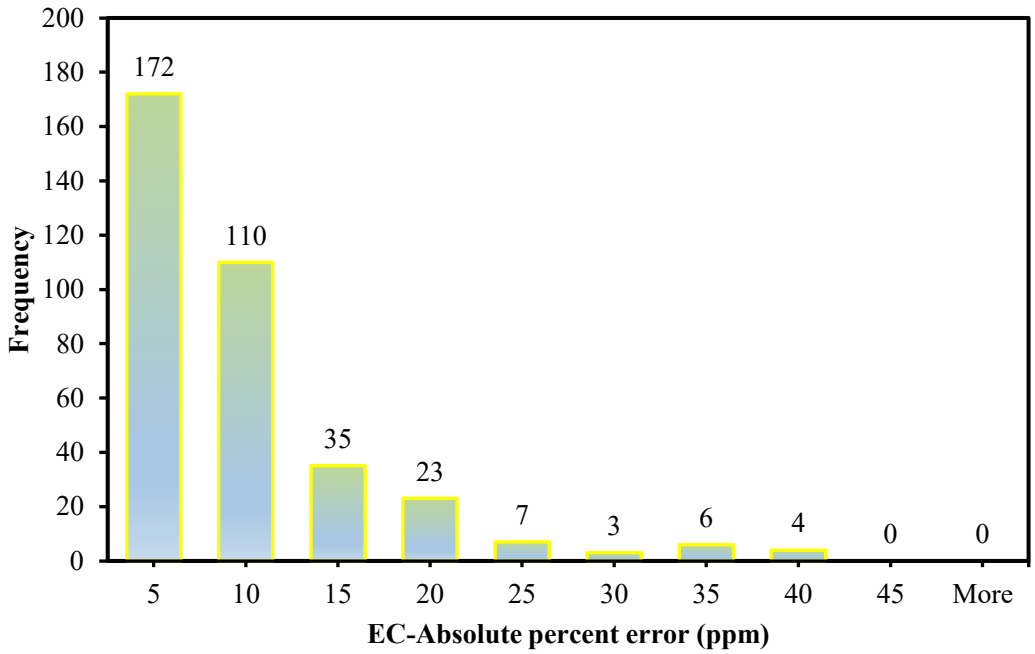


Figure 5. Histogram representing the percent error ranges of MEP-EC model.

3.2.2. TDS-MEP Model

Similar to the EC-MEP model, the TDS-MEP model also shows higher accuracy based on the slope of the regression line, R^2, and the reliable predictive performance considering the error metrics (i.e., MAE, RMSE, RMSLE, and MAPE). As shown in Figure 6, the distribution of absolute error near the *x*-axis witnesses the outburst performance of the developed TDS model. As can be seen in Table 5, the MAE is lesser than RMSE in all three stages, with the least values in the testing stage i.e., 8.27 ppm and 11.36 ppm, respectively. The maximum absolute error is 66.25 ppm, while the minimum was found to be 0.046 ppm, signifying the effectiveness of the developed TDS model [7]. In addition, the error histogram (see Figure 7) reflects zero error predictions above 45% of the absolute

percent error with 309 readings (85.83% data points) having absolute percent error below 10%. Like the EC mode, the MAPE for TDS model in training, validation and for unseen data (testing) also falls below 10% i.e., 7.40%, 8.25%, and 5.54%, respectively, and can be similarly classified as "excellent" [7].

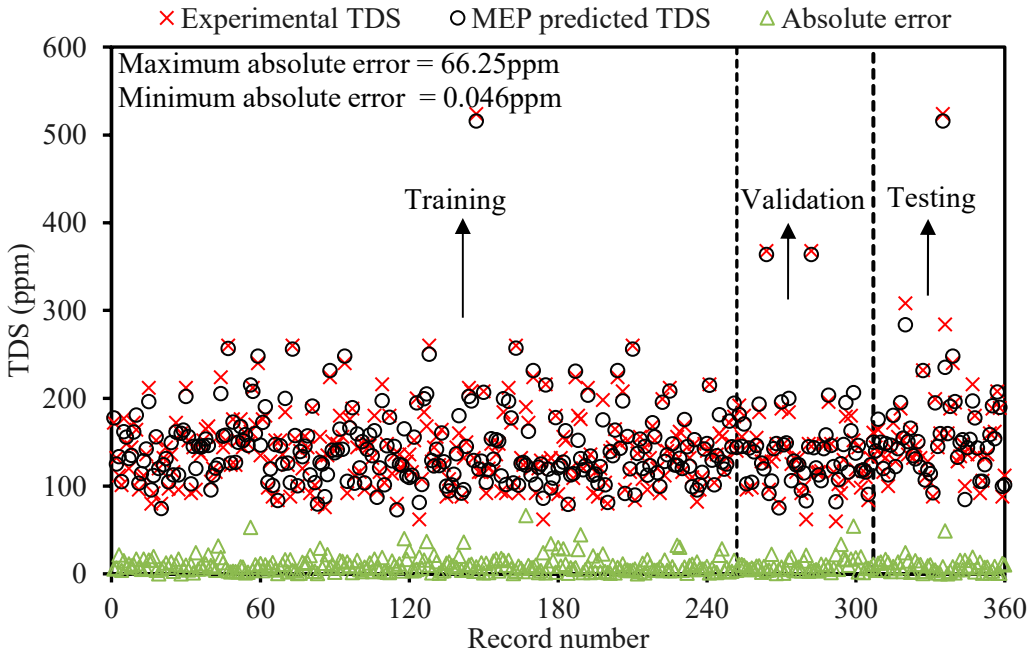


Figure 6. Absolute error plot of MEP model established for TDS.

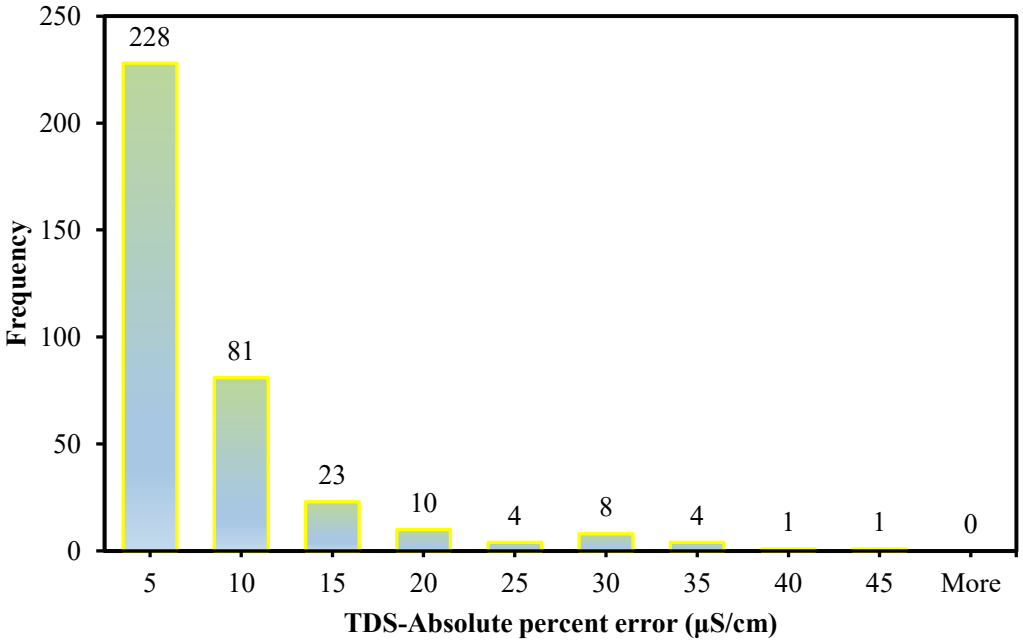


Figure 7. Histogram representing the percent error ranges of MEP-TDS model.

3.3. Comparison between the MEP Models and NLRMs

The non-linear regression method was applied to develop a mathematical model for predicting the EC and TDS based on the inputs, using the same data sets [37,61]. Equations (9) and (10) denotes the developed regression equations for EC and TDS. Figure 8a,b illustrates the deviation of the experimental and regression model predicted results of EC and TDS, respectively. The predicted results of both models (EC and TDS) largely deviate from their targeted

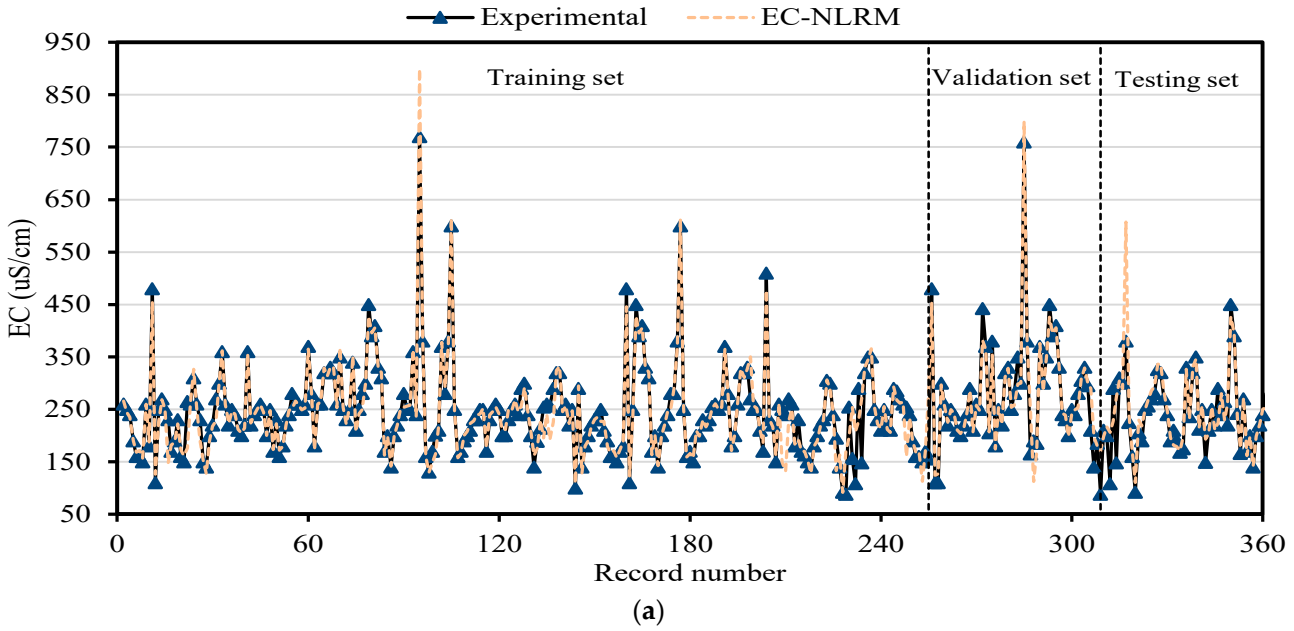
values, which make the performance and reliability of developed traditional regression models doubtful. The statistical measure presented in Table 6 shows that on average the performance of EC-NLRM is 28.36% (R^2), 29.67% (MAE), 54.67% (RMSE), 85.62% (RMSLE), and 27.19% (MAPE) lower than EC-MEP model. Consequently, TDS-NLRM gives 29.26% (R^2), 27.34% (MAE), 45.67% (RMSE), 78.58% (RMSLE), and 34.59% (MAPE), which is an inaccurate prediction as compared to TDS-MEP model. In the testing phase, the R^2 and MAPE of EC-NLRM are 0.7156 and 25.88% respectively, and 0.7295 and 28.813% for TDS-NLRM models. Thus, the developed NLRMs can be categorized as "acceptable" models for prediction but are less accurate than MEP-based models. It can be observed from Table 6 that the performance measurements of NLRM models get worse when subjected to unseen (testing data), replicating the inconsistency and irregularities in the performance [35,36,78]. In essence, the traditional regression models (i.e., NLRMs) are not useful for the prediction of complex problems because of their inefficiency and lesser generalization capability [35,78].

$$\mathrm{EC}(\mu\mathrm{S/cm}) = 3.21 + 0.19\mathrm{Ca}^{0.29} + 44.75\mathrm{Mg}^{0.15} + 34.65\mathrm{Na}^{0.28} + 58.33\mathrm{HCO_3}^{1.23} + 92.4\mathrm{Cl}^{0.95} + 79.31\mathrm{SO_4}^{0.97} - 1.05\mathrm{pH}^{0.17} - 2.72\mathrm{WT}^{0.57} \quad (9)$$

$$\mathrm{TDS}(\mathrm{ppm}) = 43.05 + 0.21\mathrm{Ca}^{0.88} + 9.33\mathrm{Mg}^{0.82} + 11.07\mathrm{Na}^{0.66} + 27.5\mathrm{HCO_3}^{1.33} + 56.55\mathrm{Cl}^{0.95} + 48.78\mathrm{SO_4}^{1.15} - 1.21\mathrm{pH}^{0.91} - 0.76\mathrm{WT}^{0.81} \quad (10)$$

Table 6. Statistical error measurement of non-linear regression models developed for EC and TDS.

Statistical Measure	NLRM-EC			NLRM-TDS		
	Training	Validation	Testing	Training	Validation	Testing
R^2	0.9133	0.9343	0.7156	0.9204	0.9552	0.7295
MAPE	16.148	17.863	25.88	15.896	16.017	28.813
MAE	13.320	15.738	22.912	8.135	8.258	12.95
RMSE	25.599	27.742	43.004	14.544	14.234	33.15
RMSLE	0.0013	0.00253	0.00351	0.00124	0.00753	0.00853



(**a**)

Figure 8. *Cont.*

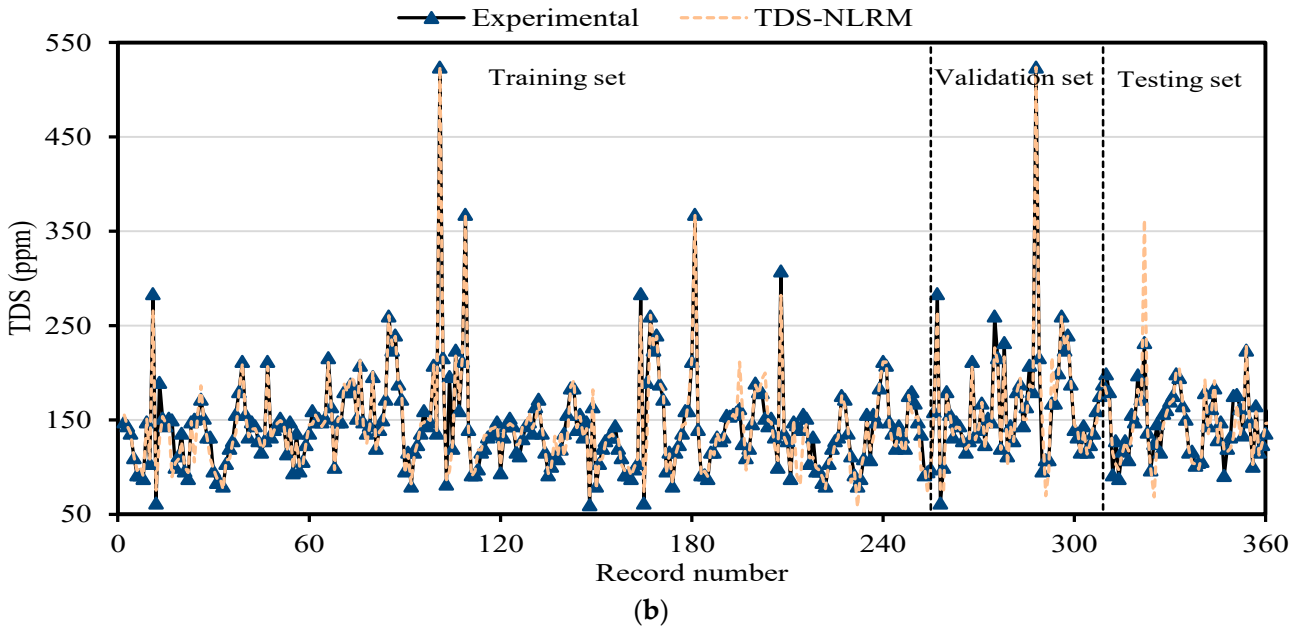


Figure 8. Deviation of observed and predicted results of NLRMs (**a**) EC; (**b**) TDS.

3.4. Sensitivity Analysis of MEP Models

In machine learning modeling, it is imperative to do many assessments to verify that the developed models (EC and TDS) are reliable and function successfully on a variety of data configurations. The higher efficacy on an established set of data, which includes the training, validation, and testing phases, does not ensure a model's superiority. A sensitivity analysis was presented in numerous studies and is used in the current work to assess that the model works well and is not just a relationship of inputs and outputs. The total data is subjected to a sensitivity analysis to get a deeper interpretation of the relevance of the independent variables on the projected water quality indicators. The sensitivity "R_{sen}" for a given input variables "I_j" are computed as Equation (11) [79,80].

$$R_{sen} = \frac{\sum_{j=1}^{m}\left(I_j \times P_j\right)}{\sqrt{\sum_{j=1}^{m} I_j^{\,2} \times \sum_{j=1}^{m} P_j^{\,2}}} \tag{11}$$

In the above equation, "P_j" are the forecasted outcome of the established predictive model (EC or TDS), and "m" represents the number of the instances/readings (m = 360). The R_{sen} score typically runs from zero to one, indicating the intensity of the relationship between a single input and the projected outcome (EC or TDS). When the R_{sen} score approaches one, the particular input has a greater effect upon a certain forecasted output.

Figure 9 represents the sensitivity (R_{sen}) of the inputs for the prediction of EC and TDS. All the eight variables considered in the current research influence the prediction of the water quality parameters (EC and TDS), utilizing the suggested models with distinct effects (all above 0.5). The sensitivity of the chlorides, sulphates, and carbonates concentration remains high for both water quality parameters, suggesting that all three of these inputs are the most relevant attributes influencing the forecasting outcomes [35]. However, the least contributing input is the sodium with a sensitiveness equal to 0.53 and 0.51 for EC and TDS, respectively.

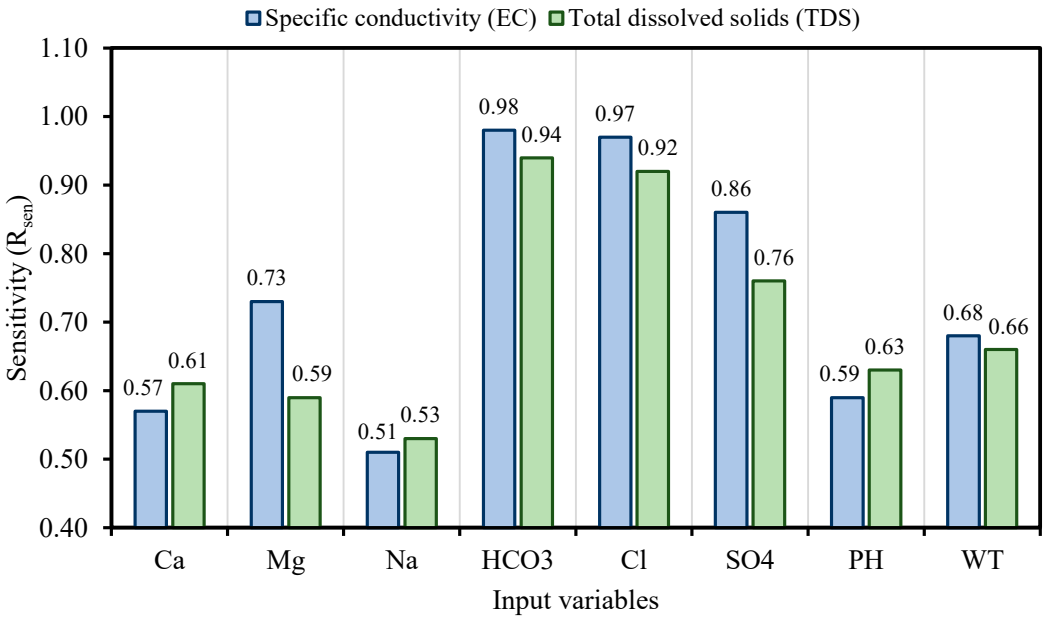


Figure 9. Contribution of input variables for the prediction of EC and TDS.

4. Recommendations and Suggestions

The current study concentrated on a case-based analysis of the upper Indus River basin. The results of this analysis provide a helpful and comprehensive knowledge of water-related challenges which can be applied for the assessment of the groundwater quality and performance evaluation of machine learning techniques. The results revealed that predictions based on evolutionary algorithms (i.e., MEP) are robust and can be utilized confidently for water quality concerns, which will be useful for government, practitioners, designers, and policy makers in order to save the available limited amount of water resources, thus conserving the environment. Furthermore, it is highly suggested that additional studies be undertaken employing other AI methodologies including ensemble simulations to better develop their performance with least hyper-parameters tunning involved in the estimation of water quality metrics. Additional water attributes should also be included when evaluating the predictive power of the aforementioned methodologies. The spatio-temporal assessment will provide a complete perspective and pinpoint the mechanisms causing water quality degradation.

5. Conclusions

The contamination of water is a big issue that directly threatens both human health and the environment, and also poses a serious issue for agricultural productivity. This study proposed evolutionary algorithm- (EA) based multi-expression programming (MEP) models for the prediction of specific conductivity (EC) and total dissolved solids (TDS). A bigger water quality dataset of 360 readings collected on a monthly basis was used in the modeling process. The eight most influential water quality input variables were selected, i.e., water temperature (°C), magnesium (Mg), calcium (Ca), sodium (Na), sulphate (SO_4), chloride (Cl), pH, and bicarbonates (HCO_3). The accuracy, reliability and generalization of the established models were evaluated using various well-known of statistical measures, i.e., slope and coefficient of determination (R^2), mean-absolute-present error (MAPE), mean-absolute-error (MAE), root-mean-square-logarithmic error (RMSLE), and root-mean-square error (RMSE). The performance of the models was compared with traditional multiple non-linear regression (NLRM) models. The regression results of EC-MEP and TDS-MEP showed excellent accuracy with coefficient of regression (R^2) and slope above 0.95 in the testing phase on unseen data. Also, the error statistics are minimum, showing the generalized and reliable performance. The projected (RMSE and MAE) in EC prediction were (18.54 µS/cm

and 12.36 μS/cm), (17.19 μS/cm and 12.14 μS/cm) and (16.43 uS/cm and 11.22 μS/cm) for training, validation and testing sets, respectively, and for the TDS modeling they were (13.36 ppm and 9.75 ppm), (13.33 ppm and 9.80 ppm) and (11.36 ppm and 8.27 ppm), respectively. The RMSLE approaches 0, indicating an outburst performance. According to MAPE, the performance of the established models was categorized as "excellent" and thus can be confidently used for future predictions. The predictions of NLRMs show a significant deviation from the targeted results, reflecting the reduced performance statistics that made the reliability of the NLRMs doubtful. However, the MAPE of the NLRMs also falls within acceptable limits i.e., below 50%. In essence, the traditional regression models (i.e., NLRMs) are not useful for the prediction of complex problems because of their inefficiency and least generalization capability. Furthermore, the sensitivity analysis of the developed MEP models revealed that all of the eight variables considered in current research influences the prediction of the water quality parameters (EC and TDS), with distinct effects having a sensitiveness index above 0.5. Thus, the developed EA-based MEP models are not merely the correlations but can be helpful for practitioners and decision-makers that will eventually save the time and money required for monitoring water quality parameters.

Author Contributions: Conceptualization, A.A. and A.T.B.T.; methodology, A.W.M.N.; software, M.A.K.; validation, M.A.U.R.T., A.W.M.N. and M.A.K.; formal analysis, A.W.M.N.; investigation, A.W.M.N.; resources, A.A.; data curation, A.T.B.T.; writing—original draft preparation, A.A.; writing—review and editing, A.W.M.N.; visualization, A.M.M.; supervision, A.W.M.N.; project administration, M.A.U.R.T.; funding acquisition, A.A. and A.M.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Acknowledgments: The authors extended their appreciation to the Deputyship for Research & Innovation, Ministry of Education in Saudi Arabia for funding this research work through the project number (IF-PSAU-2021/01/18623).

Conflicts of Interest: The authors declare that they have no conflict of interest.

Appendix A

Table A1. A-1: MEP generated code for prediction of Electric Conductivity (EC).

```
#include <math.h>
#include <stdio.h>

void mepx(double *x/*inputs*/, double *outputs)
{
   double prg[38];
   prg[0] = x[6];
   prg[1] = x[5];
   prg[2] = x[3];
   prg[3] = prg[1] * prg[1];
   prg[4] = log(prg[3]);
   prg[5] = prg[2] + prg[1];
   prg[6] = prg[5] + prg[5];
   prg[7] = prg[2] * prg[3];
   prg[8] = prg[5] + prg[6];
   prg[9] = prg[0] + prg[8];
```

Table A1. *Cont.*

```
    prg[10] = x[4];
    prg[11] = prg[10] − prg[0];
    prg[12] = prg[9]/prg[2];
    prg[13] = prg[4] ∗ prg[12];
    prg[14] = prg[9] ∗ prg[9];
    prg[15] = x[4];
    prg[16] = prg[0] − prg[13];
    prg[17] = prg[13] − prg[0];
    prg[18] = x[0];
    prg[19] = prg[14] + prg[5];
    prg[20] = x[0];
    prg[21] = log(prg[20]);
    prg[22] = x[7];
    prg[23] = prg[16] ∗ prg[21];
    prg[24] = prg[13] − prg[18];
    prg[25] = prg[12] + prg[23];
    prg[26] = prg[19] + prg[25];
    prg[27] = prg[7]/prg[24];
    prg[28] = prg[16] − prg[17];
    prg[29] = prg[26] + prg[24];
    prg[30] = prg[28] ∗ prg[15];
    prg[31] = prg[27] + prg[29];
    prg[32] = prg[30] + prg[31];
    prg[33] = prg[32] + prg[21];
    prg[34] = prg[33] − prg[7];
    prg[35] = prg[22]/prg[11];
    prg[36] = prg[30] + prg[34];
    prg[37] = prg[35] + prg[36];

    outputs[0] = prg[37];
}

int main(void)
{

//example of utilization...

    double x[8];
    x[0] = 1.680000;
    x[1] = 0.730000;
    x[2] = 0.320000;
    x[3] = 1.550000;
    x[4] = 0.690000;
    x[5] = 0.480000;
    x[6] = 7.900000;
    x[7] = 5.555556;

    double outputs[1];

    mepx(x, outputs);
    printf("%lf", outputs[0]);
    getchar();
}
```

Table A2. A-2: MEP generated code for prediction of Total Dissolved Solid (TDS).

```
#include <math.h>
#include <stdio.h>

void mepx(double *x/*inputs*/, double *outputs)
{
   double prg[44];
   prg[0] = x[6];
   prg[1] = prg[0] ∗ prg[0];
   prg[2] = x[5];
   prg[3] = x[5];
   prg[4] = x[3];
   prg[5] = sqrt(prg[4]);
   prg[6] = x[3];
   prg[7] = prg[5]/prg[6];
   prg[8] = prg[5] − prg[2];
   prg[9] = prg[7] ∗ prg[0];
   prg[10] = prg[6] * prg[8];
   prg[11] = prg[7] − prg[2];
   prg[12] = prg[7] + prg[7];
   prg[13] = prg[9] + prg[9];
   prg[14] = prg[13] + prg[9];
   prg[15] = prg[6] ∗ prg[14];
   prg[16] = prg[11] + prg[12];
   prg[17] = prg[14] + prg[5];
   prg[18] = prg[17] + prg[9];
   prg[19] = prg[16] + prg[11];
   prg[20] = x[7];
   prg[21] = x[1];
   prg[22] = prg[3] + prg[4];
   prg[23] = prg[1] ∗ prg[22];
   prg[24] = prg[19] ∗ prg[19];
   prg[25] = prg[22] + prg[12];
   prg[26] = x[4];
   prg[27] = sqrt(prg[25]);
   prg[28] = prg[9] + prg[23];
   prg[29] = prg[26] ∗ prg[15];
   prg[30] = prg[17] − prg[1];
   prg[31] = prg[20]/prg[6];
   prg[32] = prg[13]/prg[20];
   prg[33] = prg[28] + prg[29];
   prg[34] = prg[10] − prg[29];
   prg[35] = prg[31] ∗ prg[27];
   prg[36] = prg[30] + prg[24];
   prg[37] = prg[35]/prg[34];
   prg[38] = prg[37] + prg[36];
   prg[39] = prg[32] − prg[26];
   prg[40] = prg[21] + prg[33];
   prg[41] = prg[38] + prg[40];
   prg[42] = prg[39] + prg[18];
   prg[43] = prg[42] + prg[41];

   outputs[0] = prg[43];
}

int main(void)
{

//example of utilization...
```

Table A2. *Cont.*

```
    double x[8];
    x[0] = 1.680000;
    x[1] = 0.730000;
    x[2] = 0.320000;
    x[3] = 1.550000;
    x[4] = 0.690000;
    x[5] = 0.480000;
    x[6] = 7.900000;
    x[7] = 5.600000;

    double outputs[1];

    mepx(x, outputs);
    printf("%lf", outputs[0]);
    getchar();
}
```

References

1. Pandhiani, S.M.; Sihag, P.; Shabri, A.B.; Singh, B.; Pham, Q.B. Time-series prediction of streamflows of Malaysian rivers using data-driven techniques. *J. Irrig. Drain. Eng.* **2020**, *146*, 04020013. [CrossRef]
2. Singh, A.P.; Dhadse, K.; Ahalawat, J. Managing water quality of a river using an integrated geographically weighted regression technique with fuzzy decision-making model. *Environ. Monit. Assess.* **2019**, *191*, 378. [PubMed]
3. Shahzad, G.; Rehan, R.; Fahim, M. Rapid performance evaluation of water supply services for strategic planning. *Civ. Eng. J.* **2019**, *5*, 1197–1204. [CrossRef]
4. Solangi, G.S.; Siyal, A.A.; Siyal, P. Analysis of Indus Delta groundwater and surface water suitability for domestic and irrigation purposes. *Civ. Eng. J.* **2019**, *5*, 1599–1608. [CrossRef]
5. 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]
6. Velmurugan, A.; Swarnam, P.; Subramani, T.; Meena, B.; Kaledhonkar, M. Water demand and salinity. In *Desalination-Challenges and Opportunities*; IntechOpen: London, UK, 2020.
7. 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]
8. Jagaba, A.; Kutty, S.; Hayder, G.; Baloo, L.; Abubakar, S.; Ghaleb, A.; Lawal, I.; Noor, A.; Umaru, I.; Almahbashi, N. Water quality hazard assessment for hand dug wells in Rafin Zurfi, Bauchi State, Nigeria. *Ain Shams Eng. J.* **2020**, *11*, 983–999. [CrossRef]
9. 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]
10. 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]
11. Salami, E.; Salari, M.; Ehteshami, M.; Bidokhti, N.; Ghadimi, H. Application of artificial neural networks and mathematical modeling for the prediction of water quality variables (case study: Southwest of Iran). *Desalination Water Treat.* **2016**, *57*, 27073–27084. [CrossRef]
12. El Osta, M.; Masoud, M.; Alqarawy, A.; Elsayed, S.; Gad, M. Groundwater Suitability for Drinking and Irrigation Using Water Quality Indices and Multivariate Modeling in Makkah Al-Mukarramah Province, Saudi Arabia. *Water* **2022**, *14*, 483. [CrossRef]
13. 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]
14. Alexakis, D.E. Linking DPSIR Model and Water Quality Indices to Achieve Sustainable Development Goals in Groundwater Resources. *Hydrology* **2021**, *8*, 90. [CrossRef]
15. Alexakis, D.E. Meta-evaluation of water quality indices. application into groundwater resources. *Water* **2020**, *12*, 1890. [CrossRef]
16. Dehghani, M.; Saghafian, B.; Nasiri Saleh, F.; Farokhnia, A.; Noori, R. Uncertainty analysis of streamflow drought forecast using artificial neural networks and Monte-Carlo simulation. *Int. J. Climatol.* **2014**, *34*, 1169–1180.
17. Mandal, S.; Mahapatra, S.; Adhikari, S.; Patel, R. Modeling of arsenic (III) removal by evolutionary genetic programming and least square support vector machine models. *Environ. Process* **2015**, *2*, 145–172. [CrossRef]
18. Alizadeh, M.J.; Kavianpour, M.R.; Danesh, M.; Adolf, J.; Shamshirband, S.; Chau, K.-W. Effect of river flow on the quality of estuarine and coastal waters using machine learning models. *Eng. Appl. Comput. Fluid Mech.* **2018**, *12*, 810–823. [CrossRef]
19. Kargar, K.; Samadianfard, S.; Parsa, J.; Nabipour, N.; Shamshirband, S.; Mosavi, A.; Chau, K.-W. Estimating longitudinal dispersion coefficient in natural streams using empirical models and machine learning algorithms. *Eng. Appl. Comput. Fluid Mech.* **2020**, *14*, 311–322. [CrossRef]

20. Sihag, P.; Tiwari, N.; Ranjan, S. Prediction of unsaturated hydraulic conductivity using adaptive neuro-fuzzy inference system (ANFIS). *ISH J. Hydraul. Eng.* **2019**, *25*, 132–142. [CrossRef]
21. Sihag, P.; Tiwari, N.; Ranjan, S. Modelling of infiltration of sandy soil using gaussian process regression. *Modeling Earth Syst. Environ.* **2017**, *3*, 1091–1100. [CrossRef]
22. Yaseen, Z.M.; Sulaiman, S.O.; Deo, R.C.; Chau, K.-W. An enhanced extreme learning machine model for river flow forecasting: State-of-the-art, practical applications in water resource engineering area and future research direction. *J. Hydrol.* **2019**, *569*, 387–408. [CrossRef]
23. Najafzadeh, M.; Tafarojnoruz, A. Evaluation of neuro-fuzzy GMDH-based particle swarm optimization to predict longitudinal dispersion coefficient in rivers. *Environ. Earth Sci.* **2016**, *75*, 157. [CrossRef]
24. Najafzadeh, M.; Tafarojnoruz, A.; Lim, S.Y. Prediction of local scour depth downstream of sluice gates using data-driven models. *ISH J. Hydraul. Eng.* **2017**, *23*, 195–202. [CrossRef]
25. Najafzadeh, M.; Rezaie-Balf, M.; Tafarojnoruz, A. Prediction of riprap stone size under overtopping flow using data-driven models. *Int. J. River Basin Manag.* **2018**, *16*, 505–512. [CrossRef]
26. Tung, T.M.; Yaseen, Z.M. A survey on river water quality modelling using artificial intelligence models: 2000–2020. *J. Hydrol.* **2020**, *585*, 124670.
27. Tripathi, M.; Singal, S.K. Use of principal component analysis for parameter selection for development of a novel water quality index: A case study of river Ganga India. *Ecol. Indic.* **2019**, *96*, 430–436. [CrossRef]
28. Zali, M.A.; Retnam, A.; Juahir, H.; Zain, S.M.; Kasim, M.F.; Abdullah, B.; Saadudin, S.B. Sensitivity analysis for water quality index (WQI) prediction for Kinta River, Malaysia. *World Appl. Sci. J.* **2011**, *14*, 60–65.
29. Nigam, U.; SM, Y. Development of computational assessment model of fuzzy rule based evaluation of groundwater quality index: Comparison and analysis with conventional index. In Proceedings of the International Conference on Sustainable Computing in Science, Technology and Management (SUSCOM), Amity University Rajasthan, Jaipur, India, 26–28 February 2019.
30. Srinivas, R.; Singh, A.P. Application of fuzzy multi-criteria approach to assess the water quality of river Ganges. In *Soft Computing: Theories and Applications*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 513–522.
31. Yaseen, Z.M.; Ramal, M.M.; Diop, L.; Jaafar, O.; Demir, V.; Kisi, O. Hybrid adaptive neuro-fuzzy models for water quality index estimation. *Water Resour. Manag.* **2018**, *32*, 2227–2245. [CrossRef]
32. Hameed, M.; Sharqi, S.S.; Yaseen, Z.M.; Afan, H.A.; Hussain, A.; Elshafie, A. Application of artificial intelligence (AI) techniques in water quality index prediction: A case study in tropical region, Malaysia. *Neural Comput. Appl.* **2017**, *28*, 893–905. [CrossRef]
33. 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]
34. Sarkar, A.; Pandey, P. River water quality modelling using artificial neural network technique. *Aquat. Procedia* **2015**, *4*, 1070–1077. [CrossRef]
35. Zhang, Y.; Gao, X.; Smith, K.; Inial, G.; Liu, S.; Conil, L.B.; Pan, B. Integrating water quality and operation into prediction of water production in drinking water treatment plants by genetic algorithm enhanced artificial neural network. *Water Res.* **2019**, *164*, 114888. [CrossRef] [PubMed]
36. Chen, K.; Chen, H.; Zhou, C.; Huang, Y.; Qi, X.; Shen, R.; Liu, F.; Zuo, M.; Zou, X.; Wang, J. Comparative analysis of surface water quality prediction performance and identification of key water parameters using different machine learning models based on big data. *Water Res.* **2020**, *171*, 115454. [CrossRef] [PubMed]
37. Hazarika, B.B.; Gupta, D.; Ashu; Berlin, M. A Comparative Analysis of Artificial Neural Network and Support Vector Regression for River Suspended Sediment Load Prediction. In *First International Conference on Sustainable Technologies for Computational Intelligence*; Springer: Singapore, 2020; pp. 339–349.
38. Granata, F.; Papirio, S.; Esposito, G.; Gargano, R.; De Marinis, G. Machine Learning Algorithms for the Forecasting of Wastewater Quality Indicators. *Water* **2017**, *9*, 105. [CrossRef]
39. Zha, T.-H.; Castillo, O.; Jahanshahi, H.; Yusuf, A.; Alassafi, M.O.; Alsaadi, F.E.; Chu, Y.-M. A fuzzy-based strategy to suppress the novel coronavirus (2019-NCOV) massive outbreak. *Appl. Comput. Math.* **2021**, *20*, 160–176.
40. Nazeer, M.; Hussain, F.; Khan, M.I.; El-Zahar, E.R.; Chu, Y.-M.; Malik, M. Theoretical study of MHD electro-osmotically flow of third-grade fluid in micro channel. *Appl. Math. Comput.* **2022**, *420*, 126868. [CrossRef]
41. Zhao, T.H.; Khan, M.I.; Chu, Y.M. Artificial neural networking (ANN) analysis for heat and entropy generation in flow of non-Newtonian fluid between two rotating disks. *Math. Methods Appl. Sci.* **2021**. [CrossRef]
42. Chu, H.-H.; Zhao, T.-H.; Chu, Y.-M. Sharp bounds for the Toader mean of order 3 in terms of arithmetic, quadratic and contraharmonic means. *Math. Slovaca* **2020**, *70*, 1097–1112. [CrossRef]
43. Zhao, T.-H.; He, Z.-Y.; Chu, Y.-M. On some refinements for inequalities involving zero-balanced hypergeometric function. *AIMS Math.* **2020**, *5*, 6479–6495. [CrossRef]
44. Zhao, T.-H.; Wang, M.-K.; Chu, Y.-M. A sharp double inequality involving generalized complete elliptic integral of the first kind. *AIMS Math.* **2020**, *5*, 4512–4528.
45. Zhao, T.-H.; Zhou, B.-C.; Wang, M.-K.; Chu, Y.-M. On approximating the quasi-arithmetic mean. *J. Inequalities Appl.* **2019**, *2019*, 42. [CrossRef]
46. Zhao, T.-H.; Wang, M.-K.; Zhang, W.; Chu, Y.-M. Quadratic transformation inequalities for Gaussian hypergeometric function. *J. Inequal. Appl.* **2018**, *2018*, 1–15. [CrossRef] [PubMed]

47. Azimi, S.; Azhdary Moghaddam, M.; Hashemi Monfared, S.A. Prediction of annual drinking water quality reduction based on Groundwater Resource Index using the artificial neural network and fuzzy clustering. *J. Contam. Hydrol.* **2019**, *220*, 6–17. [CrossRef] [PubMed]
48. Ismael, M.; Mokhtar, A.; Farooq, M.; Lü, X. Assessing drinking water quality based on physical, chemical and microbial parameters in the Red Sea State, Sudan using a combination of water quality index and artificial neural network model. *Groundw. Sustain. Dev.* **2021**, *14*, 100612. [CrossRef]
49. Kim, J.; Seo, D.; Jang, M.; Kim, J. Augmentation of limited input data using an artificial neural network method to improve the accuracy of water quality modeling in a large lake. *J. Hydrol.* **2021**, *602*, 126817. [CrossRef]
50. Zhang, Q.; Li, Z.; Zhu, L.; Zhang, F.; Sekerinski, E.; Han, J.-C.; Zhou, Y. Real-time prediction of river chloride concentration using ensemble learning. *Environ. Pollut.* **2021**, *291*, 118116. [CrossRef]
51. 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. Int.* **2021**, *28*, 13202–13220. [CrossRef]
52. Jiang, L.; Chui, T.F.M. A review of the application of constructed wetlands (CWs) and their hydraulic, water quality and biological responses to changing hydrological conditions. *Ecol. Eng.* **2022**, *174*, 106459. [CrossRef]
53. Oltean, M.; Grosan, C. A comparison of several linear genetic programming techniques. *Complex Syst.* **2003**, *14*, 285–314.
54. Arabshahi, A.; Gharaei-Moghaddam, N.; Tavakkolizadeh, M. Development of applicable design models for concrete columns confined with aramid fiber reinforced polymer using Multi-Expression Programming. *Structures* **2020**, *23*, 225–244. [CrossRef]
55. Goldberg, D.E. *Genetic Algorithms*; Pearson Education India: London, UK, 2006.
56. Koza, J.R.; Koza, J.R. *Genetic Programming: On the Programming of Computers by Means of Natural Selection*; MIT Press: Cambridge, MA, USA, 1992; Volume 1.
57. Alavi, A.H.; Gandomi, A.H.; Nejad, H.C.; Mollahasani, A.; Rashed, A. Design equations for prediction of pressuremeter soil deformation moduli utilizing expression programming systems. *Neural Comput. Appl.* **2013**, *23*, 1771–1786. [CrossRef]
58. Li, P.; Khan, M.A.; El-Zahar, E.R.; Awan, H.H.; Zafar, A.; Javed, M.F.; Khan, M.I.; Qayyum, S.; Malik, M.; Wang, F. Sustainable Use of Chemically modified Tyre Rubber in Concrete: Machine Learning based Novel Predictive Model. *Chem. Phys. Lett.* **2022**, *793*, 139478. [CrossRef]
59. FM Zain, M.; M Abd, S. Multiple regression model for compressive strength prediction of high performance concrete. *J. Appl. Sci.* **2009**, *9*, 155–160. [CrossRef]
60. Faradonbeh, R.S.; Hasanipanah, M.; Amnieh, H.B.; Armaghani, D.J.; Monjezi, M. Development of GP and GEP models to estimate an environmental issue induced by blasting operation. *Environ. Monit. Assess.* **2018**, *190*, 351. [CrossRef] [PubMed]
61. Ilyas, I.; Zafar, A.; Javed, M.F.; Farooq, F.; Aslam, F.; Musarat, M.A.; Vatin, N.I. Forecasting Strength of CFRP Confined Concrete Using Multi Expression Programming. *Materials* **2021**, *14*, 7134. [CrossRef]
62. Chu, H.-H.; Khan, M.A.; Javed, M.; Zafar, A.; Khan, M.I.; Alabduljabbar, H.; Qayyum, S. Sustainable use of fly-ash: Use of gene-expression programming (GEP) and multi-expression programming (MEP) for forecasting the compressive strength geopolymer concrete. *Ain Shams Eng. J.* **2021**, *12*, 3603–3617. [CrossRef]
63. 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]
64. Khan, M.A.; Farooq, F.; Javed, M.F.; Zafar, A.; Ostrowski, K.A.; Aslam, F.; Malazdrewicz, S.; Maślak, M. Simulation of Depth of Wear of Eco-Friendly Concrete Using Machine Learning Based Computational Approaches. *Materials* **2022**, *15*, 58. [CrossRef]
65. Khan, S.; Ali Khan, M.; Zafar, A.; Javed, M.F.; Aslam, F.; Musarat, M.A.; Vatin, N.I. Predicting the Ultimate Axial Capacity of Uniaxially Loaded CFST Columns Using Multiphysics Artificial Intelligence. *Materials* **2022**, *15*, 39. [CrossRef]
66. Khan, M.A.; Shah, M.I.; Javed, M.F.; Khan, M.I.; Rasheed, S.; El-Shorbagy, M.; El-Zahar, E.R.; Malik, M. Application of random forest for modelling of surface water salinity. *Ain Shams Eng. J.* **2021**, *13*, 101635. [CrossRef]
67. Jalal, F.E.; Xu, Y.; Iqbal, M.; Javed, M.F.; Jamhiri, B. Predictive modeling of swell-strength of expansive soils using artificial intelligence approaches: ANN, ANFIS and GEP. *J. Environ. Manag.* **2021**, *289*, 112420. [CrossRef] [PubMed]
68. Azim, I.; Yang, J.; Iqbal, M.F.; Mahmood, Z.; Javed, M.F.; Wang, F.; Liu, Q.-F. Prediction of catenary action capacity of RC beam-column substructures under a missing column scenario using evolutionary algorithm. *KSCE J. Civ. Eng.* **2021**, *25*, 891–905. [CrossRef]
69. Nafees, A.; Amin, M.N.; Khan, K.; Nazir, K.; Ali, M.; Javed, M.F.; Aslam, F.; Musarat, M.A.; Vatin, N.I. Modeling of Mechanical Properties of Silica Fume-Based Green Concrete Using Machine Learning Techniques. *Polymers* **2022**, *14*, 30. [CrossRef] [PubMed]
70. Iqbal, M.F.; Javed, M.F.; Rauf, M.; Azim, I.; Ashraf, M.; Yang, J.; Liu, Q.-F. Sustainable utilization of foundry waste: Forecasting mechanical properties of foundry sand based concrete using multi-expression programming. *Sci. Total Environ.* **2021**, *780*, 146524. [CrossRef]
71. Mousavi, S.; Alavi, A.; Gandomi, A.; Arab Esmaeili, M.; Gandomi, M. A data mining approach to compressive strength of CFRP-confined concrete cylinders. *Struct. Eng. Mech.* **2010**, *36*, 759. [CrossRef]
72. Qiu, R.; Wang, Y.; Wang, D.; Qiu, W.; Wu, J.; Tao, Y. Water temperature forecasting based on modified artificial neural network methods: Two cases of the Yangtze River. *Sci. Total Environ.* **2020**, *737*, 139729. [CrossRef]
73. Jalal, F.E.; Xu, Y.; Iqbal, M.; Jamhiri, B.; Javed, M.F. Predicting the compaction characteristics of expansive soils using two genetic programming-based algorithms. *Transp. Geotech.* **2021**, *30*, 100608. [CrossRef]

74. Khan, M.A.; Zafar, A.; Farooq, F.; Javed, M.F.; Alyousef, R.; Alabduljabbar, H.; Khan, M.I. Geopolymer Concrete Compressive Strength via Artificial Neural Network, Adaptive Neuro Fuzzy Interface System, and Gene Expression Programming with K-Fold Cross Validation. *Front. Mater.* **2021**, *8*, 621163. [CrossRef]
75. Gholampour, A.; Gandomi, A.H.; Ozbakkaloglu, T. New formulations for mechanical properties of recycled aggregate concrete using gene expression programming. *Constr. Build. Mater.* **2017**, *130*, 122–145. [CrossRef]
76. Liu, Q.-F.; Iqbal, M.F.; Yang, J.; Lu, X.-Y.; Zhang, P.; Rauf, M. Prediction of chloride diffusivity in concrete using artificial neural network: Modelling and performance evaluation. *Constr. Build. Mater.* **2021**, *268*, 121082. [CrossRef]
77. Farooq, F.; Ahmed, W.; Akbar, A.; Aslam, F.; Alyousef, R. Predictive modeling for sustainable high-performance concrete from industrial wastes: A comparison and optimization of models using ensemble learners. *J. Clean. Prod.* **2021**, *292*, 126032. [CrossRef]
78. Iqbal, M.F.; Liu, Q.-F.; Azim, I.; Zhu, X.; Yang, J.; Javed, M.F.; Rauf, M. Prediction of mechanical properties of green concrete incorporating waste foundry sand based on gene expression programming. *J. Hazard. Mater.* **2020**, *384*, 121322. [CrossRef] [PubMed]
79. Ardakani, A.; Kordnaeij, A. Soil compaction parameters prediction using GMDH-type neural network and genetic algorithm. *Eur. J. Environ. Civ. Eng.* **2019**, *23*, 449–462. [CrossRef]
80. Wang, H.-L.; Yin, Z.-Y. High performance prediction of soil compaction parameters using multi expression programming. *Eng. Geol.* **2020**, *276*, 105758. [CrossRef]