Water Quality Prediction Model Based Support Vector Machine Model for Ungauged River Catchment under Dual Scenarios

water

MDPI

Article

Water Quality Prediction Model Based Support Vector Machine Model for Ungauged River Catchment under Dual Scenarios

Abobakr Saeed Abobakr Yahya [1], Ali Najah Ahmed [1,*], Faridah Binti Othman [2], Rusul Khaleel Ibrahim [2], Haitham Abdulmohsin Afan [2,*], Amr El-Shafie [3], Chow Ming Fai [1], Md Shabbir Hossain [4], Mohammad Ehteram [5] and Ahmed Elshafie [2]

1 Institute of Energy Infrastructure (IEI), Universiti Tenaga Nasional, Kajang 43000, Selangor, Malaysia; ale.a.najah@gmail.com (A.S.A.Y.); Chowmf@uniten.edu.my (C.M.F.)
2 Department of Civil Engineering, Faculty of Engineering, University Malaya, Kuala Lumpur 50603, Malaysia; faridahothman@um.edu.my (F.B.O.); rusul.alqaisy@yahoo.com (R.K.I.); elshafie@um.edu.my (A.E.)
3 Civil Engineering Department El-Gazeera High Institute for Engineering Al Moqattam, Cairo 11311, Egypt; amrhuss63@gmail.com
4 Department of Civil Engineering, School of Energy, Geoscience, Infrastructure and Society, Heriot-Watt University, Putrajaya 62200, Malaysia; m.hossain@hw.ac.uk or realism007@gmail.com
5 Department of Water Engineering and Hydraulic Structures, Faculty of Civil Engineering, Semnan University, Semnan 35131-19111, Iran; mohammdehteram@semnan.ac.ir
* Correspondence: mahfoodh@uniten.edu.my (A.N.A.); haitham.afan@gmail.com (H.A.A.)

Received: 29 April 2019; Accepted: 15 May 2019; Published: 13 June 2019

Abstract: Water quality analysis is a crucial step in water resources management and needs to be addressed urgently to control any pollution that may adversely affect the ecosystem and to ensure the environmental standards are being met. Thus, this work is an attempt to develop an efficient model using support vector machine (SVM) to predict the water quality of Langat River Basin through the analysis of the data of six parameters of dual reservoirs that are located in the catchment. The proposed model could be considered as an effective tool for identifying the water quality status for the river catchment area. In addition, the major advantage of the proposed model is that it could be useful for ungauged catchments or those lacking enough numbers of monitoring stations for water quality parameters. These parameters, namely pH, Suspended Solids (SS), Dissolved Oxygen (DO), Ammonia Nitrogen (AN), Chemical Oxygen Demand (COD), and Biochemical Oxygen Demand (BOD) were provided by the Malaysian Department of Environment (DOE). The differences between dual scenarios 1 and 2 depend on the information from prior stations to forecast DO levels for succeeding sites (Scenario 2). This scheme has the capacity to simulate water-quality accurately, with small prediction errors. The resulting correlation coefficient has maximum values of 0.998 and 0.979 after the application of Scenario 1. The approach with Type 1 SVM regression along with 10-fold cross-validation methods worked to generate precise results. The MSE value was found to be between 0.004 and 0.681, with Scenario 1 showing a better outcome.

Keywords: support vector machine; water quality; dissolved oxygen

1. Introduction

Water plays a crucial role in environmental and social life. It occupies the largest area of our planet, and therefore, among all other natural resources, water resources gain the most special place. However, the ongoing increase in urbanization and industrialization processes generate wide ranges of hazardous contaminants that deteriorate the quality of river waters. The produced wastes are

discharged into water bodies in different forms, for instance, organic pollutants (pesticides, insecticides, phenols, hydrocarbons, etc.), heavy metals (lead, arsenic, copper, cadmium mercury, etc.), and microbial pathogens. All these water pollutants cause adverse effects on public health and the surrounding wildlife. Therefore, there is an exigent demand to monitor the levels of water quality. Rapid population growth and climate change in the past few decades resulted in the scarcity of fresh water, as well as degrading the water-quality levels across the globe [1]. The Water Quality Index (WQI) summarizes water quality information in a readily-understood scale format. Index numbers range between 1 and 100, wherein higher numbers indicate better water-quality levels. Generally, river stations that score 80 or more indicate that water quality meets expectations for "clean rivers", stations that score from 40 to 80 indicate "slightly polluted rivers", whereas river stations that score below 40 indicate that the water quality expectations are not being met, and are regarded as "polluted rivers".

In past research, various water-quality parameters are integrated into a general index that forms the water-quality index (WQI). The development of WQI and its use in assessing water-quality levels have risen exponentially [2]. One of the challenges faced in recent researches on water quality is the need to reduce costs while developing smarter computer-aided means of assessing water-quality levels. WQI represents the best means to communicate and categorize water-quality levels in assessments of water suitability for various applications. Nevertheless, a variety of WQI types have been planned for quick and efficient assessments of the general level of water quality throughout particular regions. The WQI scale relied on by the Canadian Council of Ministers of Environment is utilized in Canada and many other countries [3].

1.1. Background

Polluted water has been widely spread all over the globe in the past few decades, and thus water-quality level assessment and forecasting have become more essential to the ecological management organisations of many countries. A set of six parameters is typically used in DOE analyses for evaluating surface water characteristics in terms of pH, Suspended Solids (SS), Dissolved Oxygen (DO), Ammonia Nitrogen (AN), Chemical Oxygen Demand (COD), and Biochemical Oxygen Demand (BOD). The prediction of water-quality parameters based on accurate values could enable better early cautionary of contaminated water and timely decision-making. To this day, dual scenarios are extensively employed to forecast water-quality trends. Scenario 1 is designed to validate the DO prediction scheme at each station according to five input parameters, while Scenario 2 is designed to validate the DO prediction scheme according to five input parameters as well as DO predictions from a prior (upstream) station [4–6].

In this study, time-series predictive techniques for water quality were discussed, which predict the value of water quality parameters at interval t through use of preceding time series, based on similar and other parameters. For many decades, numerous statistical analyses and AI-based modelling strategies have been used to influence time-series predictive techniques for water quality and in water resources management [6–15]. To this day, the term water quality is utilized for describing water conditions, such as physical, chemical, and biological properties. Among the most important factors for water quality is the dissolved oxygen (DO), which refers to the presence of oxygen gas molecules (O_2) in terms of mg/L concentrations. Plant and animal species cannot directly utilise the oxygen that forms part of water molecules (H_2O). The dissolved oxygen content can be decreased through respiration processes, wherein oxygen enters surface water from the air as a photosynthetic product of river plants. Generally, a constant high level of DO is best for ecosystems.

The level of DO varies according to many factors such as depth, altitude, and season, rates of flow, water temperatures, and time of the day. Water coursing at increased temperatures and at higher altitudes normally contains lesser amounts of DO. Moreover, DO levels peak during the daytime, whereas during night time, the DO levels lower along with the decreased photosynthetic action due to the oxygen-consuming activities including oxidation and respiration, until just before dawn. DO levels directly affect fish cultures: (1) At low DO levels (0–2 mg/L), oxygen is insufficient to support

fish cultures; (2) at middle DO levels (2–4 mg/L), few river fish and insect species can survive; (3) at higher DO levels (4–7 mg/L), oxygen is sufficient to support diverse river life, particularly cold-water fish species; (4) For the majority of stream fish, the optimal DO range is 7–11 mg/L. Human activities that influence DO levels in surface water include the entry of additional oxygen-consuming organic wastes, including sewage, nutrient additions, changing water flows, rising water temperatures, and added chemicals [16].

Selangor, Malaysia, has long experience with river water pollution issues linked to land-use trends. The Langat River forms a principal basin that drains into densely-populated and highly-developed areas in Selangor. For four decades, the river has served almost half the populace of Selangor as a water supply and remains a source of hydroelectric power and flood discharge controls. More than two-thirds of the region's population resides in the floodplain, which features highly fertile soil for agricultural as well as for industrial, residential, and recreational purposes. This situation has brought about conflict between economic development and river ecosystems, which has led to increasing amounts of wastes in riverine channels [17]. A total of 42 tributaries surveyed in Peninsular Malaysia are categorised as highly polluted, including Langat River Basin. In 1999, a total of 13 tributaries across Malaysia along with 36 rivers were polluted as a result of industrial, construction, and agricultural activities along the tributaries [18]. Until 1990, a total of 48 rivers could still be classified as clean, in comparison to only 32 that remained clean as of 1999 [15]. Some 60% of all main rivers are subject to regulation for residential, agricultural, and commercial purposes. The primary pollution sources in Malaysia that affect its river systems are sewage disposal, discharge from small- and medium-sized industries that improperly treat effluents, and various land-clearing and earthwork processes. By 1999, some 42% of all river basins were reportedly contaminated with suspended solids (SS) as a result of badly planned and controlled land-clearing processes, some 30% with heavy biological oxygen demand (BOD) due to industrial effluent, and some 28% with ammoniacal nitrogen (AN) due to animal husbandry activity and domestic disposal of sewage [19].

Surface water pollution has been identified as the primary problem that affects the Langat River. The expansion of human development in the basin has led to increased wastes entering its channels. In efforts to prevent further pollution, the Department of Environment (DOE), under Malaysia's Ministry of Natural Resources and Environment, has set up telemetry stations all along the river in a programme that monitors the water-quality levels on a continuing basis. Based on water-quality data, the water-quality index (WQI) was applied in order to assess water-quality status and river category. The measure offers a useful basis for forecasting changes in water-quality levels, through the consideration of several factors. WQI comprises six water-quality level variables, including SS, pH, AN, dissolved oxygen (DO), chemical oxygen demand (COD), and biochemical oxygen demand (BOD) [14].

1.2. Problem Statement

The level of pollution has been increasing in Selangor with regards to the increased amounts of wastes entering the Langat River leading to water shortage within the basin. Numerous sources of wastes contaminate the Langat River, including emissions of industrial effluent that are 58% of all pollutants, with 28% discharged from domestic sewage treatment plants, while construction projects account for some 12% and pig farms only 2% of the total. In 2011, the WQI for Langat River was between 72 and 75, which led to its classification as a slightly polluted river. WQI remains at a class I level at the upstream station in the Hulu Langat District prior to the Lui River, which indicates a low-pollution area. However, WQI changes to a class III level at the downstream station past the Batang Benar and Balak tributaries [20]. This paper examines the efficiency of Support Vector Machine (SVM) models in predicting water-quality levels for Langat River within Malaysia. A network was devised and trained using six major parameters to estimate water-pollutant levels as well as water-quality levels and classes. Its dataset includes records that have been gathered over the past decade since 2006. Artificial Neural Network (ANN) and Support Vector Machine (SVM) systems have been used to

forecast water-quality levels. SVM offers more advantages than its ANN equivalents, for it can resolve small samples in term of nonlinear, high-dimensional, localised minima, and other partial elements. SVM also features a modular scheme that enables independent implementation of component designs. This research demonstrates the use of SVM in forecasting water-quality characteristics based on these six parameters, with dynamic processes masked in the very measurement data.

1.3. Objectives

The objectives and goals of the project are as follows:

(1) To provide effective methods of water-quality prediction to decision makers, towards better water resource planning and management.
(2) To present a system-independent method of water-quality forecasting that utilises SVM in place of statistical modelling methods.
(3) To provide recommendations to water-quality management agencies that accord with the research findings on the Langat River Basin, regarding how such findings may be integrated into ecological strategies for catchment management.

2. Materials and Methods

2.1. Case Study

The Langat basin (Figure 1) includes a key catchment that provides for raw water supplies and further resources to some 1.2 million people residing there. Key stretches serve various towns, namely Kajang, Cheras, Bangi, Putrajaya Government Centre, and so on. Dual reservoirs are located in the catchment, namely Semenyih and Hulu Langat Dam, along with eight water-treatment plants that provide for safe and clean and water supplies. Langat River features a total catchment area of some 2350 km^2 and is located at latitudes between 2°40′M 152″ N and 3°16′M 15″ N and along longitudes between 101°19′M 20″ E and 102°1′M 10″ E. The principal river has a length of 141 km that stretches some 40 km east of the capital of Kuala Lumpur and features several tributaries, the largest of which are the Lui, Semenyih, and Beranang Rivers. There are two reservoirs that the Langat and Semenyih dams supply via particular catchment. The newer Langat dam was constructed in 1981 with a catchment area totalling 54 km^2, whereas Semenyih was constructed in 1942 and features a catchment area totalling 41 km^2. The Langat River represents more than just a supply of water, for it serves further purposes that include recreational activities, fishing, irrigation, effluent discharge, as well as sand mining. As a result of the growth of numerous activities along its streams, it has become necessary to evaluate related water demand and resource supply.

This research investigates water-quality forecast methods for the Langat catchment, which drains into three major tributaries, the Langat, Semenyih, and Labu Rivers. Nevertheless, this research only covers the upper portion of the Langat catchment. The primary Langat tributary flows some 182 km from the major range (Banjaran Titiwangsa) in the northeast of Hulu, through Langat District in the south-southwest direction, and drains via the Straits of Malacca. The Langat and Semenyih Rivers both originate in hilly and forested regions along the western slopes of Banjaran Titiwangsa, which is northeast of the Hulu Langat District. Semenyih River runs mainly from the Semenyih Dam, in a stretch that flows south-southwest into and through the towns of Semenyih and Bangi Lama, which then joins with the Langat River some 4 km east of Bangi Lama. Semenyih River is joined by the Beranang and Pajam Rivers as well, and both originate in the northern area of the Seremban District, namely Negeri Sembilan. The data has been gathered from seven water stations along the Langat River.

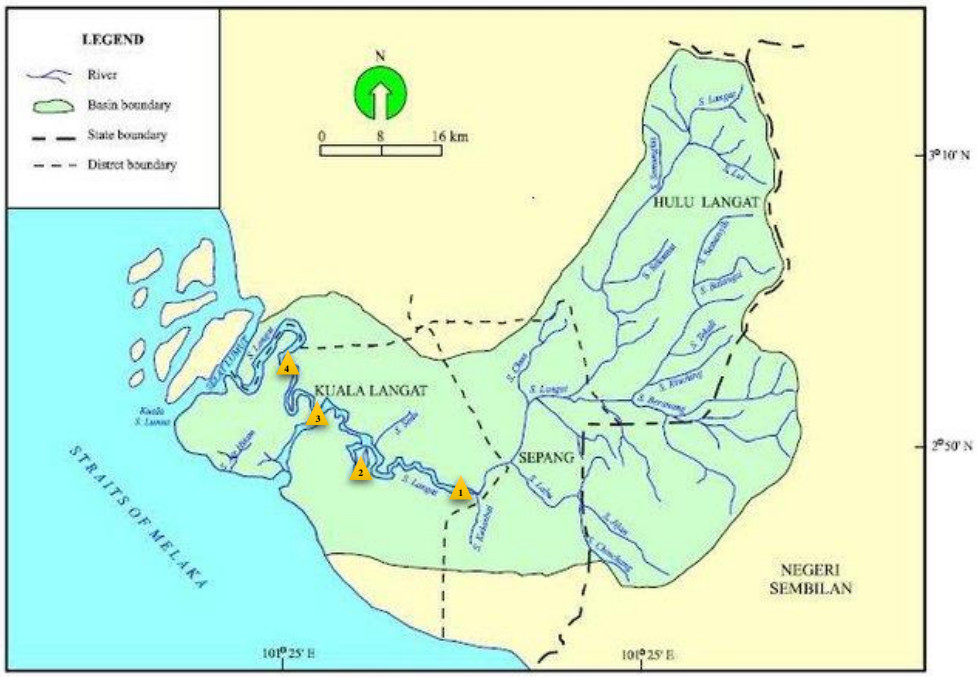

Figure 1. Water station on Langat River basin map.

2.2. Select Appropriate Inputs

The water-quality data utilized in this research was obtained through the Malaysian Department of Environment (DOE). Time-series data for selected parameter sets through four stations from 2006 until 2016 were utilized in this research. Six water-quality parameters were chosen for SVM modelling in this research, specifically pH, Suspended Solids (SS), Dissolved Oxygen (DO), Ammonia Nitrogen (AN), Chemical Oxygen Demand (COD), and Biochemical Oxygen Demand (BOD). Simple statistics for measured water-quality parameters in the Langat River from 2006 to 2016 are displayed in Table 1.

Table 1. Basic statistics of the measured water quality parameter in Langat River.

Sampling Site	Basic Statistic	DO (mg/L)	BOD (mg/L)	COD (mg/L)	pH	SS (mg/L)	AN (mg/L)
Station 1	Mean	4.34	6.29	29.77	6.46	160.45	0.59
	Min	0.83	1	4	2.41	0.1	0.01
	Max	7.53	23	99	7.36	1050	1.69
	SD	1.354	4.691	17.07	0.69	199	0.443
	CV	31.2	74.5	57.3	10.6	124	74.9
Station 2	Mean	4.34	5.77	25.54	6.58	121.3	0.96
	Min	0.87	1	7	3.8	1	0.01
	Max	7.5	24	70	7.92	821	3.73
	SD	1.11	3.59	12.74	0.70	116.58	0.662
	CV	25.67	62.25	49.90	10.70	96.11	69.23
Station 3	Mean	5.81	5.29	23.72	7.01	182.9	1.11
	Min	3.55	1	2	6.10	5	0.01
	Max	7.62	17	66	8.28	1400	5.75
	SD	0.77	2.50	11.87	0.323	202.94	0.995
	CV	13.2	47.34	50.06	4.61	110.96	90.01
Station 4	Mean	5.50	8.03	30.50	7.07	272.44	1.56
	Min	2.39	2	5	5.84	3	0.01
	Max	8.2	27	84.10	7.91	1910	6.65
	SD	1.19	4.84	15.24	0.29	350.52	1.27
	CV	21.65	60.29	49.9	4.20	128.66	81.44

2.3. Structure of SVM

Support vectors are the training points that are the nearest to the separating hyperplane and the basic concept of SVM is illustrated in Figure 2. There are decision functions that are accountable for example, hyperplanes that are able to delineate the positive and negative data that has marked the maximum margins. This shows the range from the nearest positive sample to a hyperplane and the range between the nearest negative sample and the hyperplane shall be maximized [21].

$$y(x) = w^T \phi(x) + b \tag{1}$$

where, $\phi(x)$ represents the high dimensional feature spaces, which is nonlinearly mapped from the input space x. The coefficient are w and b are estimated by minimizing the regularized function $R(C)$:

$$R(C) = C\frac{1}{N}\sum_{i=1}^{N} L_\varepsilon(d_i, y_i) + \frac{1}{2}\|w\|^2 \tag{2}$$

where

$$L_\varepsilon(d_i, y_i) = \begin{cases} \left|d_i - y_i\right| - \varepsilon & if \left|d_i - y_i\right| \geq \varepsilon \\ 0 & otherwise \end{cases} \tag{3}$$

To obtain the estimation of w and b, Equation (2) is transformed to the primal function given by Equation (4) by introducing the positive slack variables ξ_i and ξ_i^* as follows:

$$Minimize: \; R(C) = C\left(\sum_{i=1}^{N}\left(\xi_i + \xi_i^*\right)\right) + \frac{1}{2}\|w\|^2 \tag{4}$$

$$Subject\ to \begin{cases} d_i - w\varnothing(x_i) - b_i \leq \varepsilon + \xi_i, \; i = 1, 2, \ldots, N \\ w\varnothing(x_i) + b_i - y_i \leq \varepsilon + \xi_i^*, \; i = 1, 2, \ldots, N \\ \xi_i, \xi_i^* \geq 0 \; i = 1, 2, \ldots, N \end{cases}$$

The first term (l/2||w||2) is the weights vector norm, d_i the desired value, and C is referred as regularized constant determining the tradeoff between the empirical error and the regularized term. ε is called the tube size of SVM as shown in Figure 2.

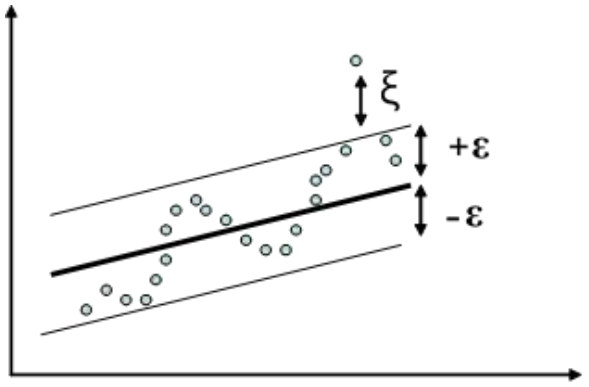

Figure 2. Hyperplane and the basic concept of support vector machine (SVM) [21].

It is similar to the equation accuracy that is related in the training data. The multipliers that are non zero are known as support vectors. This is where the variable slacks and brought into the study. ξ_i

and ξ_i^* are introduced and by introducing this and by exploiting constraint optimality the function decision by Equation (1) gives the following:

$$y(x) = \sum_{i=1}^{N} \left(\alpha_i - \alpha_i^*\right) K\left(x, x_i\right) + b \tag{5}$$

In Equation (5), α_i and α_i^*, are the so-called Lagrange multipliers. They satisfy the equalities $\alpha_i, \times \alpha_i^* = 0$, $\alpha_i, \geq 0$ and $\alpha_i^* \geq 0$ where, $i = 1, 2, \ldots, n$ and are obtained by maximizing the dual function of Equation (4) which has the following form:

$$R\left(\alpha_i, \alpha_i^*\right) = \sum_{i=1}^{N} d_i\left(\alpha_i - \alpha_i^*\right) - \varepsilon \sum_{i=1}^{N}\left(\alpha_i + \alpha_i^*\right) - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} d_i\left(\alpha_i - \alpha_i^*\right)\left(\alpha_j - \alpha_j^*\right) K\left(x_i, x_j\right) \tag{6}$$

with the constraints

$$\sum_{i=1}^{N}(\alpha_i) = \sum_{i=1}^{N}\left(\alpha_i^*\right) \text{ and } \alpha_i, \alpha_i^* \in [0, C],\ i = 1, 2, \ldots, N \tag{7}$$

$K\left(x_i, x_j\right)$ is defined as the kernel function. The value of the kernel is equal to the inner product of two vectors x_i and x_j in the feature space $\phi(x_i)$ and $\phi(x_j)$, that is, $K\left(x_i, x_j\right) = \phi(x_i) \times \phi(x_j)$.

Four common kernel function types of SVM are given as follows:

$$\text{Linear kernel}: \ K\left(x_i, x_j\right) = x_i^T \times x_j \tag{8}$$

$$\text{Polynomial kernel}: \ K\left(x_i, x_j\right) = \left(\gamma x_i^T \times x_j + r\right)^d,\ \gamma > 0 \tag{9}$$

$$\text{Radial basis kernel}: \ K\left(x_i, x_j\right) = \exp(-\|x_i - x_j\|^2),\ \gamma > 0 \tag{10}$$

$$\text{Sigmoid kernel}: \ k\left(x_i, x_j\right) = tanh\left(x_i^T \times x_j + r\right) \tag{11}$$

Here, γ, r and d are kernel parameters. The kernel parameters should be carefully chosen as it implicitly defines the structure of the high dimensional feature space $\phi(x)$ and thus controls the complexity of the final solution.

There are two types of SVM regression; both have the general formula given in Equation (1). The first type of SVM regression is known as Type 1 or Epsilon. This type of error function is given by the formula shown in Equation (4). The second type of regression is known as Nu.

2.4. Statistical Indexes

Model performances were assessed based on three statistical indexes.

2.4.1. Coefficient of Efficiency (CE)

To assess model performance, Coefficient of Efficiency (CE) was utilized [5].

$$CE = 1 - \frac{\sum_{i=1}^{n}(m - p)^2}{\sum_{i=1}^{n}(m - \overline{m})^2} \tag{12}$$

wherein n is the number of observations, m and p correspond to the measured and predicted data, while $\overline{m}$ denotes the average of the measured data.

2.4.2. Mean Square Error (MSE)

Mean square error (MSE) is utilized to establish and define the fit of the network outputs to the desired scheme. Lower values for MSE enable higher performance (values nearer 0 indicate reliable and accurate results), which is expressed as follows:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(m-p)^2 \tag{13}$$

2.4.3. Coefficient of Correlation (CC)

Coefficient of correlation (CC) is utilized to assess the linear relationship between the observed and predicted data. It is defined as follows:

$$CC = \frac{\sum_{i=1}^{n}(m-\overline{m})(p-\overline{p})}{\sqrt{\sum_{i=1}^{n}(m-\overline{m})^2\sum_{i=1}^{n}(m-\overline{p})^2}} \tag{14}$$

2.5. Sensitivity Analysis

To assess the effect of input parameters on this model, dual evaluation processes are utilized. Firstly, performance evaluations for various possible parameter combinations are investigated using Mean Square Error (MSE) and Coefficient of Efficiency (CE) and methods, in order to establish the most effective parameters for the output. Generally, six networks were compared, each of which demonstrate the significance of eliminated parameters that influence network accuracy [22].

3. Results and Discussion

3.1. Kernel Functions of SVM

In this study, SVM was developed with the use of various kernel function types such as Linear, Polynomial, Radial Basis, and Sigmoid functions. Trial and error method were conducted in order to establish the best types of kernel functions through cross-validation-based processes. Table 2 displays a comparative analysis of predictive performances for the SVM model based on four different kernel function types. Regarding the coefficient of correlation value when utilising RBF kernel, the predictive precision of the test data was most reliable at around 0.801, followed by the Polynomial 0.496, Linear 0.438, and Sigmoid 0.336 kernels.

Table 2. Performance of SVM model with four kernel function.

Kernel Function	Mean Squared Error		Correlation Coefficient	
	Training	Testing	Training	Testing
Linear	0.229	0.189	0.564	0.438
Polynomial	0.177	0.217	0.732	0.496
RBF	0.177	0.322	0.998	0.801
Sigmoid	155.886	418.735	0.105	0.336

3.2. Epsilon-RBF Model and Nu-RBF Model

The search for model parameter sets does have a vital role in acquiring reliable predictive performances for SVM. Settings for hyper-parameters (C, γ) and kernel parameters (Nu and Epsilon) are considered to be vital in shaping SVM generalization performances (approximation accuracy).

In the search for the best SVM architecture for particular applications, dual test types exist for RBF kernel functions, namely Epsilon (ε) and Nu. Epsilon-RBF is utilized for predictions, so C is set to 8.5 and γ = 8, with ε set to values ranging from 0.001 to 0.5. Results for model performance in testing

and training are displayed in Table 3, which shows that MSE increases with increases in values of ε while every correlation coefficient value and the numbers of support vectors decreases. Lastly, an ε value of 0.001 yields minimal generalization errors, with acceptable numbers of the support vector (36) selected. Where the Nu model increases, the predictive precision of the test data will gradually rise to its highest value (at Nu = 0.4), with the remaining values set as displayed. Table 4 also displays the numbers that associate with support vectors increasing as Nu values increase. In accordance with these results, it is noted that the best-fitting values for correlation coefficient and MSE correspond to 0.998 and 0.001, respectively, where obtained by utilizing Nu-RBF Model with Gamma equal to 8 and Capacity equal to 8.5.

Table 3. The result of various Epsilon parameter ε, of SVM model where C = 8.5 and gamma = 0.2 of total behavior prediction model.

ε	Mean Error Squared		Correlation Coefficient			No. of Support Vectors
	(Train)	(Test)	(Train)	(Test)	(Overall)	
0.001	0.008	0.305	0.976	0.286	0.785	36
0.1	0.030	0.353	0.910	0.198	0.703	27
0.2	0.074	0.226	0.802	0.166	0.654	14
0.3	0.118	0.212	0.688	0.130	0.567	9
0.4	0.144	0.222	0.633	0.126	0.519	6
0.5	0.206	0.275	0.538	0.104	0.443	5

Table 4. Nu-RBF performance utilizing different values of Nu with fixed (gamma = 8, capacity = 8.5) of Total behavior prediction model.

Nu	Mean Error Squared		Correlation Coefficient			No. of Support Vectors
	(Train)	(Test)	(Train)	(Test)	(Overall)	
0.001	0.357	0.363	0.416	0.057	0.339	2
0.1	0.018	0.340	0.946	0.241	0.743	33
0.2	0.003	0.302	0.992	0.306	0.802	41
0.3	0.001	0.326	0.998	0.272	0.797	41
0.4	0.001	0.321	0.998	0.278	0.801	44
0.5	0.001	0.331	0.998	0.268	0.795	42

3.3. Optimising the Parameters of SVM

There are two essential parameters that are required to be selected in order to achieve the best structure of the SVM model. These parameters are gamma γ and capacity parameter C. The prediction accuracy shows a very high sensitivity to the selection of C, as a small value of C can possibly cause the model to undervalue the target at some stage in the training data. This is primarily because the use of moderately small weight in the training data will lead to larger values of the predictor as the model is examined through the testing dataset and vice-versa. In addition to this, when the value of C is large, its significance will diminish when the mapping between the input and the output is detected. On the contrary, a large value of C could indicate an extensive range of support vector's values. In accordance to it, the supplementary data records can possibly be selected for the optimisation of the support vectors. Taking this aspect into consideration, the optimal values for γ and C need to be selected through numerous trials and error methods. The selection of the optimal values of these parameters will substantially improve the probability to attain a high level of accuracy in estimating the desired data.

In the present study, the objective is to ascertain the appropriate values of the parameters γ and C. The replication processes that involved a number of parameters were also considered as the highlights of SVM and regarded as a key step in behaviour prediction model. At the initial stage of building the recommended model and throughout the training session of the input-output data, a constant value of

γ equal to 0.1 was considered, while the values of C were changeable and in the range of 0 and 10. As a result, the prediction error was calculated as MSE, while correlation coefficient was computed as the number of the support vectors. Figure 3 illustrates that an increase in the used value of C leads to a small decrease in the value of MSE and in the number of the support vectors but, in contrast, causes an increase in the value of correlation coefficient. Furthermore, by keeping the emphasis of the observation on the parameter C, it was found that the highest correlation coefficient value was 0.998 and the lowest value of MSE point was 0.001. According to the observation, with an increase in the value of parameter C, there is a marginal reduction in the value of MSE, and then it shows an upward trend after the value of C reaches the optimal point. Therefore, it is advantageous to choose the parameter C to be 8.5 for station 4, and the remaining figures for every other station. In reality, it is important to first ascertain the applicable values for the hyper parameters C and γ because it will be a significant step in the implementation of any SVM model. Hence, conducting a trial and error procedure is important and necessary. The generalised accuracy has been assessed using different values of kernel hyper in relation to this context. Due to the shortcoming of parameter γ, it was resolved that its optimal value will be searched so that it is within the range of [0.001, 10] at increments of 1 for γ, while the values are fixed for C (C = 8.5) and Nu (Nu = 0.3). As a consequence, the optimal selection value of γ is derived using 10-fold cross-validation that involves repetition of ten times with the purpose of improving the reliability of the model outcomes. The final architecture of the model in the competition during the training session is such that the parameters yield minimum errors during the course of validation. The relationship between correlation coefficient and γ is presented in Figure 4. The figure illustrates where the value of correlation coefficient begins to rise as the value of γ increases, and attains its peak value at 8 for station 4. Thereafter, the selection of values of parameters takes place that offers the minimum generalisation error. For the model in training and forecasting phase, the best outcome is taken into consideration when choosing γ = 8 with a satisfactory number of support vectors 41 for station 4, while the remaining figures are for other stations.

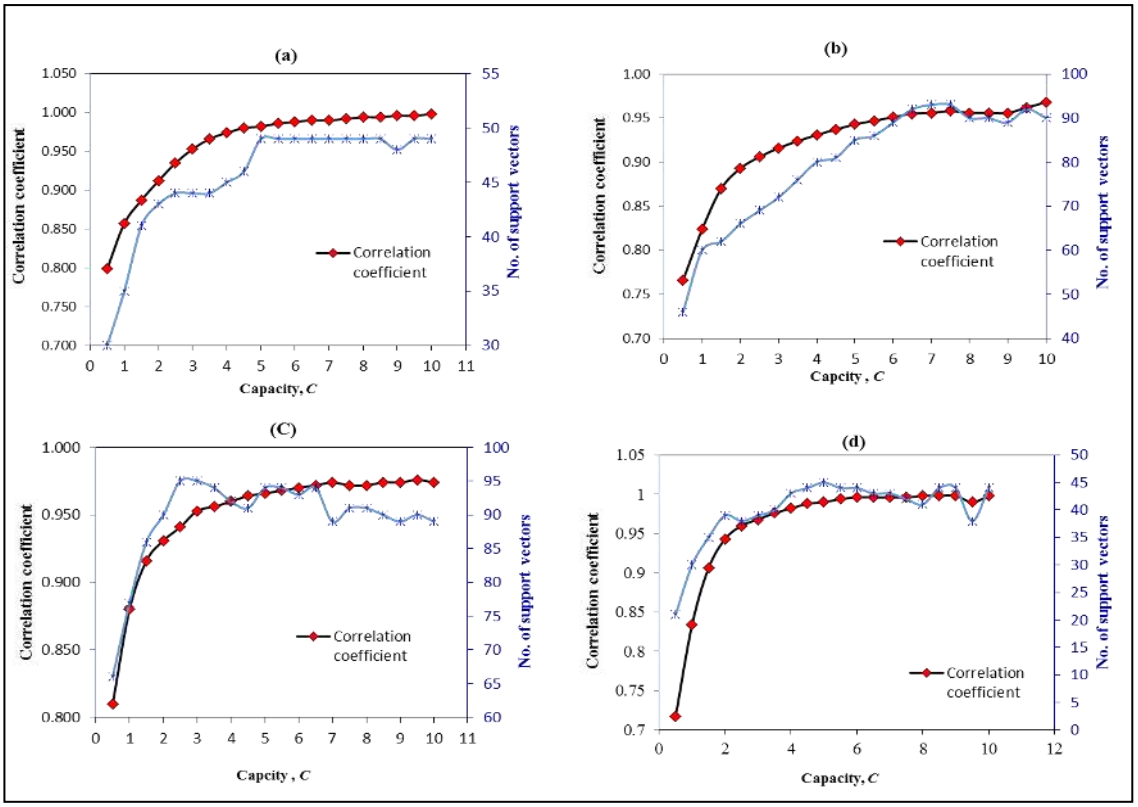

Figure 3. The result of various capacity parameter C, of SVM model, for (**a**) station 1; (**b**) station 2; (**c**) station 3; and (**d**) station 4.

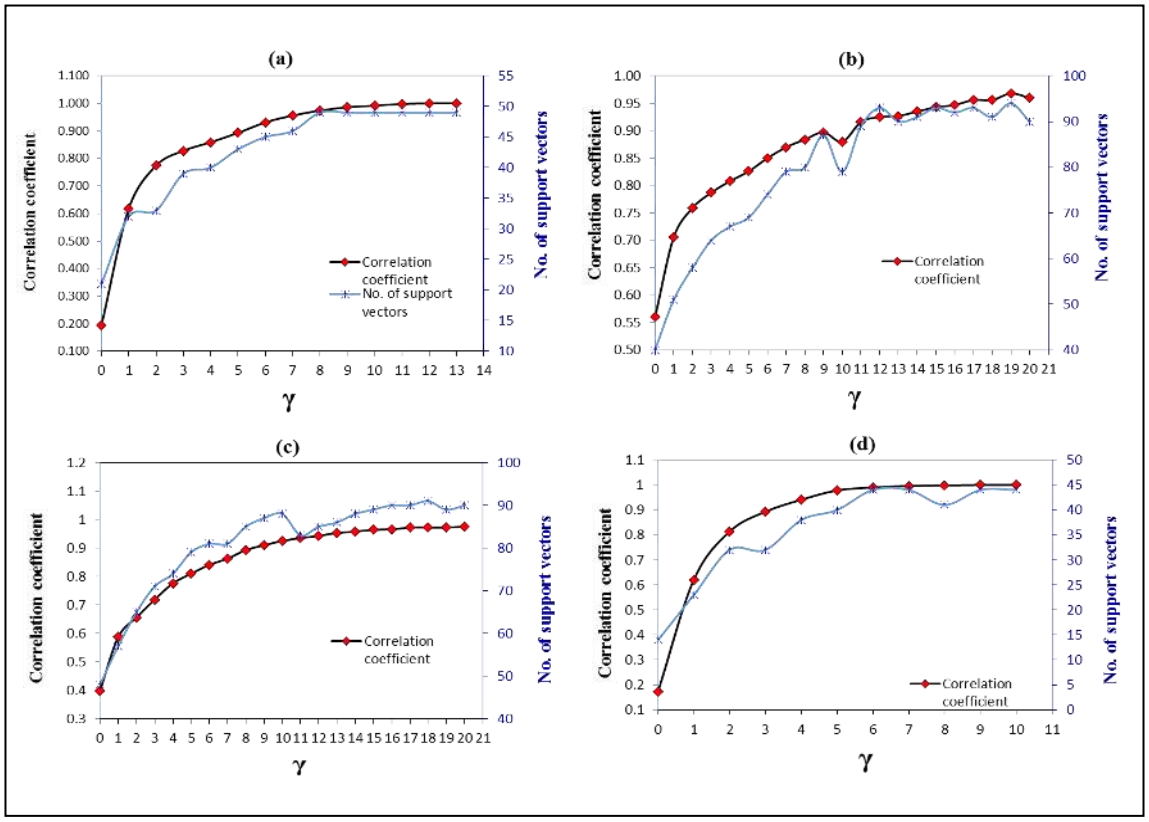

Figure 4. The result of various capacity parameter Gamma, of SVM for (**a**) station 1; (**b**) station 2; (**c**) station 3; and (**d**) station 4.

3.4. SVM Model Development

The cross validation procedure is regarded as one of the extensively used approaches for assessing the values of the model architecture parameters. The training dataset is unsystematically fragmented into a set number of V-folds (V_1, V_2, … V_n) so that the V-folds cross validation can be used. The chosen SVM model is then implemented in sequence as per the observations in relation to the V-1 folds. The expected results of the performing architecture can then be attained using the process with specific parameters on the sample V, so that the error defined by one of the statistical indices can be figured out. In other words, the sample or fold that was not visible during training of the SVM model was identified. The key limitation of this procedure is that the average accuracy for the v times causes a tendency for a consistent measure model error and impacts its stability, that is, the strength of the model for analysing the unseen session data. As shown in Table 5, the RMSE and MAE values attained by using the 10-fold cross-validation results in the best goodness of fit and performance. After the selection of the optimal kernel parameters, the Nu-RBF model is selected as the optimal model for evaluating the whole training data. Figure 5 presents a comparison of actual versus expected behaviour of DO concentration for testing data set at station 4: (a) Nu-SVM type model; (b) Epsilon-SVM type model. It can be seen that the both proposed model capable of predicting DO accurately but, in general, Nu-SVM outperformed Epsilon-SVM. While Figure 6 depicts the comparison between actual and predicted value of DO for all stations during testing.

Table 5. Statistical evaluation using 3, 5, 7, 10, and 15-fold cross-validation for Epsilon-RBF and Nu-RBF models.

Statistical Evaluation	V-Fold				
	3	**5**	**7**	**10**	**15**
ε-RBF Model:					
RMSE	0.458	0.164	0.167	0.164	0.164
MAE	0.393	0.538	0.530	0.538	0.538
CR	0.6050	0.933	0.933	0.933	0.933
Nu-RBF model:					
RMSE	0.089	0.070	0.089	0.089	0.077
MAE	0.648	0.646	0.646	0.648	0.632
CR	0.980	0.986	0.980	**0.986**	0.984

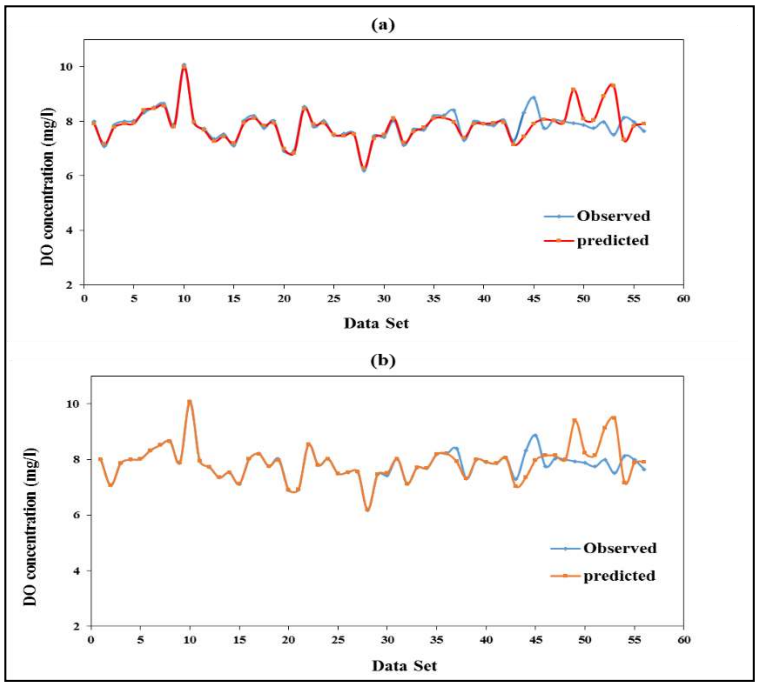

Figure 5. Comparison of actual versus predicted behavior Dissolved Oxygen (DO) concentration: (**a**) Nu-SVM type model and (**b**) Epsilon-SVM type model.

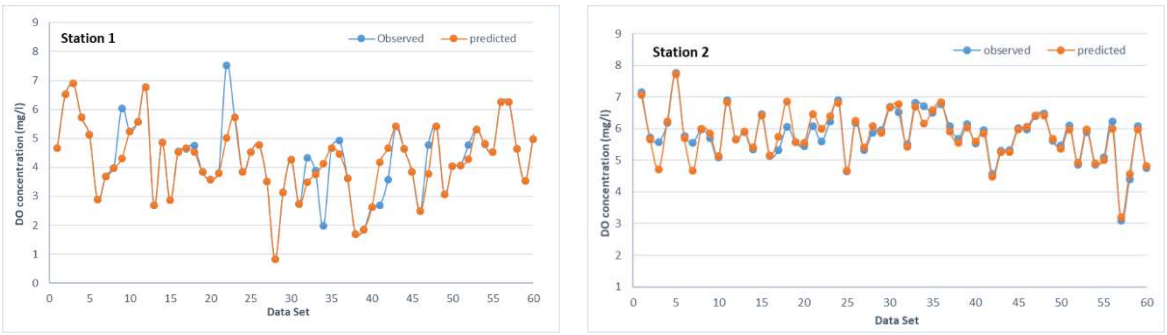

Figure 6. *Cont.*

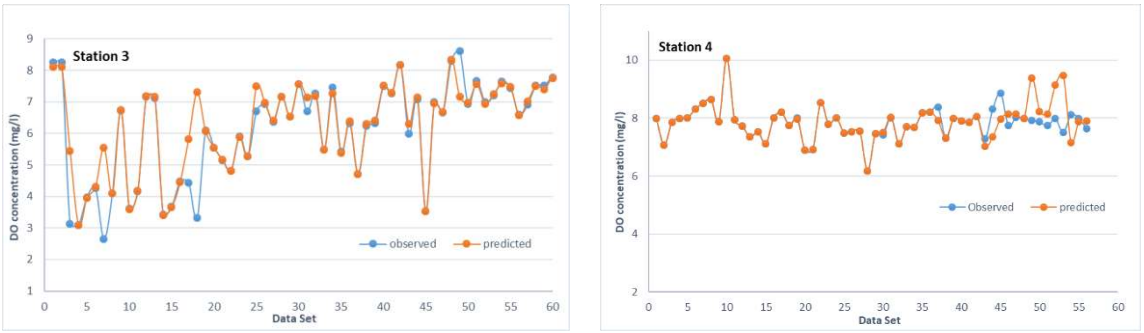

Figure 6. Comparison between actual and predicted value for DO.

3.5. Compression between Scenario One and Two

The compression between the two scenarios reveals that scenario 1 yields a more accurate result than scenario 2 in predicting DO in a two-week time horizon. Tables 6 and 7 present the results. As observed in Table 6, the sensitivity analysis produces more accurate results when all the input parameters are included. In scenario 1, the best model is the first model, which produces the highest value for correlation coefficient equal to 0.99 while the lowest mean square error value is equal to 0.001 at station 4 after including all the six parameters as input to the model. Moreover, this model performed well in all stations where the mean square error is in the range of 0.001 to 0.44 and the correlation coefficient is in the range of 0.93 to 0.99 for all stations. Figure 7 presents a Scatter plot of the observed and expected values for Scenario 1 for all four stations. On the other hand, for scenario 2, most model proposed model performed poorly in predicting DO except the model where the predicted DO from all previous three stations were used as input to predict DO as station 4 which is considered as the best model that produces the highest value for correlation coefficient equal to 0.979. Table 8 depicts the performance of the proposed model in prediction DO in different horizon at all stations in terms of maximum residual error.

Table 6. Correlation coefficient for Scenario 1.

Model	Input Parameters	Correlation Coefficient				Mean Square Error			
		DO 1	DO 2	DO 3	DO 4	DO 1	DO 2	DO 3	DO 4
1	DO, BOD, COD, SS, pH, AN	0.99	0.99	0.93	0.95	0.004	0.044	0.035	0.001
2	DO, BOD, COD, SS, pH	0.92	0.85	0.64	0.72	0.141	0.281	0.212	0.008
3	DO, BOD, COD, SS, AN	0.90	0.82	0.82	0.73	0.175	0.131	0.203	0.071
4	DO, BOD, COD, pH, AN	0.97	0.94	0.88	0.92	0.058	0.084	0.052	0.009
5	DO, BOD, SS, pH, AN	0.95	0.91	0.83	0.85	0.081	0.124	0.105	0.040
6	DO, COD, SS, pH, AN	0.96	0.92	0.79	0.82	0.068	0.160	0.135	0.038

Table 7. Results for scenario 2.

Model	Input Parameters	CC		
		DO 2	DO 3	DO 4
1	DO 1	0.279		
2	DO 1		0.236	
3	DO 2		0.258	
4	DO 1, DO2		0.571	
5	DO 1			0.503
6	DO 2			0.344
7	DO 3			0.361
8	DO 1, DO2			0.746
9	DO 1, DO3			0.656
10	DO 2, DO3			0.701
11	DO 1, DO2, DO 3			0.979

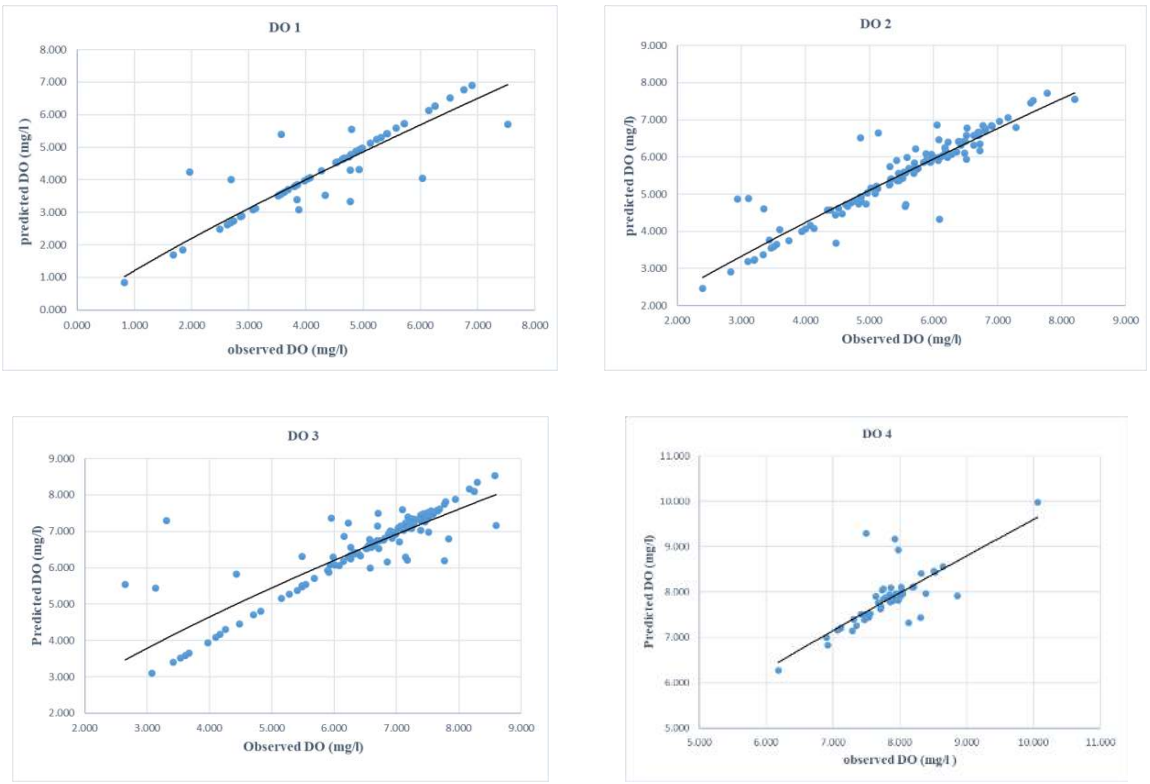

Figure 7. Scatter plots between the observed and predicted value for Scenario 1.

Table 8. Maximum residual error for DO.

Station	Maximum Residual Error		
	1 Week	2 Weeks	Month
1	2.264	2.068	1.409
2	1.924	0.634	0.525
3	3.992	3.04	2.559
4	0.994	1.129	0.450

4. Conclusions

The prediction of water quality is very vital in monitoring the pollution and in sustaining the availability of potable water resources. Undoubtedly, it can afford early warnings when the water quality changes as well as it can reduce the adverse consequences resulting from the poor water quality. Herein, the SVM approach was introduced to estimate the water quality of Langat River Basin using six parameters. The presented model accurately estimated the water quality factors with relatively minor prediction errors, proving a quite efficient and robust performance. The model also can help in the optimization of water quality monitoring plans by decreasing the frequency, quantity of sampling sites, and water quality factors. Prediction precision with the maximum error was equal to 1% and CC was equal to 0.9987. Even though the outcomes seem to be reasonable, the application of water quality parameters is quite sensitive to the error level. 1% as a maximum error is comparatively on the higher side in such an application which triggers the need to improve it. In this regard, it is suggested to deploy the optimal kernel parameters determined and choose the Nu-RBF model as the optimal model.

Author Contributions: Formal analysis, A.S.A.Y., A.N.A., H.A.A., C.M.F. and A.E.; Methodology, A.S.A.Y., A.N.A., H.A.A., M.S.H. and M.E.; Writing—original draft, F.B.O., R.K.I. and A.E.; Writing—review & editing, A.E.-S.

Funding: This research was funded by Innovation & Research Management Center (iRMC), Universiti Tenaga Nasional, Malaysia, Bold 2025 grant coded RJO 10436494 and J510050822 and funded by the University of Malaya grant coded UMRG RP025A-18SUS.

Acknowledgments: The authors appreciate so much the facilities support by the Civil Engineering Department, Faculty of Engineering, University of Malaya, Malaysia.

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

References

1. Tarmizi, A.; Ahmed, A.N.; El-Shafie, A. Dissolved Oxygen Prediction Using Support Vector Machine in Terengganu River. *Middle-East. J. Sci. Res.* **2014**, *21*, 2182–2188.
2. Behmel, S.; Damour, M.; Ludwig, R.; Rodriguez, M. Water quality monitoring strategies—A review and future perspectives. *Sci. Total Environ.* **2016**, *571*, 1312–1329. [CrossRef] [PubMed]
3. Lumb, A.; Sharma, T.C.; Bibeault, J.-F. A Review of Genesis and Evolution of Water Quality Index (WQI) and Some Future Directions. *Water Qual. Expo. Health* **2011**, *3*, 11–24. [CrossRef]
4. Najah, A.; El-Shafie, A.; Karim, O.A.; Jaafar, O. Integrated versus isolated scenario for prediction dissolved oxygen at progression of water quality monitoring stations. *Hydrol. Earth Syst. Sci. Discuss.* **2011**, *8*, 6069–6112. [CrossRef]
5. Najah, A.; El-Shafie, A.; Karim, O.A.; El-Shafie, A.H. Performance of ANFIS versus MLP-NN dissolved oxygen prediction models in water quality monitoring. *Environ. Sci. Pollut. Res.* **2014**, *21*, 1658–1670. [CrossRef] [PubMed]
6. Ahmed, A.N.; El-Shafie, A.; Karim, O.A.; El-Shafie, A.H. An augmented wavelet de-noising technique with neuro-fuzzy inference system for water quality prediction. *Int. J. Innov. Comput. Inf. Control* **2012**, *8*, 7055–7082.
7. Yoon, H.; Kim, Y.; Ha, K.; Lee, S.-H.; Kim, G.-P. Comparative Evaluation of ANN- and SVM-Time Series Models for Predicting Freshwater-Saltwater Interface Fluctuations. *Water* **2017**, *9*, 323. [CrossRef]
8. Wang, Y.; Wu, L.; Engel, B. Prediction of Sewage Treatment Cost in Rural Regions with Multivariate Adaptive Regression Splines. *Water* **2019**, *11*, 195. [CrossRef]
9. Liu, C.; Hu, Y.; Yu, T.; Xu, Q.; Liu, C.; Li, X.; Shen, C. Optimizing the Water Treatment Design and Management of the Artificial Lake with Water Quality Modeling and Surrogate-Based Approach. *Water* **2019**, *11*, 391. [CrossRef]
10. Elzwayie, A.; El-shafie, A.; Yaseen, Z.M.; Afan, H.A.; Allawi, M.F. RBFNN-based model for heavy metal prediction for different climatic and pollution conditions. *Neural Comput. Appl.* **2017**, *28*, 1–13. [CrossRef]
11. El-Shafie, A.; Najah, A.; Alsulami, H.M.; Jahanbani, H. Optimized Neural Network Prediction Model for Potential Evapotranspiration Utilizing Ensemble Procedure. *Water Resour. Manag.* **2014**, *28*, 947–967. [CrossRef]
12. Wang, K.; Wen, X.; Hou, D.; Tu, D.; Zhu, N.; Huang, P.; Zhang, G.; Zhang, H. Application of Least-Squares Support Vector Machines for Quantitative Evaluation of Known Contaminant in Water Distribution System Using Online Water Quality Parameters. *Sensors* **2018**, *18*, 938. [CrossRef] [PubMed]
13. Chang, M.-J.; Chang, H.-K.; Chen, Y.-C.; Lin, G.-F.; Chen, P.-A.; Lai, J.-S.; Tan, Y.-C. A Support Vector Machine Forecasting Model for Typhoon Flood Inundation Mapping and Early Flood Warning Systems. *Water* **2018**, *10*, 1734. [CrossRef]
14. Wang, Z.; Liu, J.; Li, J.; Zhang, D.D. Multi-Spectral Water Index (MuWI): A Native 10-m Multi-Spectral Water Index for Accurate Water Mapping on Sentinel-2. *Remote Sens.* **2018**, *10*, 1643. [CrossRef]
15. 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]
16. Chen, Y.F. Influence and control strategies of agricultural nonpoint source pollution on water quality. *China Resour. Compr. Utilization* **2016**, *34*, 54–56.
17. Akoteyon, I.S.; Omotayo, A.O.; Soladoye, O.; Olaoye, H.O. Determination of water quality index and suitability of urban river for municipal water supply in Lagos-Nigeria. *Eur. J. Sci. Res.* **2011**, *54*, 263–271.

18. Malaysia.Jabatan Alam Sekitar. *Malaysia Environmental Quality Report 2007*; Department of Enviroment: Petaling Jaya, Malaysia, 2008; 84p, ISBN 9770127643008.
19. River Water Quality (RWQ) | Malaysia Environmental Performance Index. Available online: http://www.epi.utm.my/v4/?page_id=102 (accessed on 5 April 2019).
20. Heng, L.Y.; Abdullah, M.P.; Yi, C.S.; Mokhtar, M.; Ahmad, R. Development of Possible Indicators for Sewage Pollution for the Assessment Of Langat River Ecosystem Health. *Malays. J. Anal. Sci.* **2006**, *10*, 15–26.
21. Aljanabi, Q.A.; Chik, Z.; Allawi, M.F.; El-Shafie, A.H.; Ahmed, A.N.; El-Shafie, A. Support vector regression-based model for prediction of behavior stone column parameters in soft clay under highway embankment. *Neural Comput. Appl.* **2018**, *30*, 1–11. [CrossRef]
22. Afiq, H.; Ahmed, E.; Ali, N.; Othman, A.; Aini, K.; Mukhlisi, H.M. Daily forecasting of dam water levels: Comparing a support vector machine (SVM) model with adaptive neuro fuzzy inference system (ANFIS). *Water Resour. Manag.* **2013**, *27*, 3803–3823. [CrossRef]