Mathematical Biosciences and Engineering

MBE, 19 (12): 12744–12773.
DOI: 10.3934/mbe.2022595
Received: 06 July 2022
Revised: 15 August 2022
Accepted: 18 August 2022
Published: 01 September 2022

http://www.aimspress.com/journal/MBE

Research article

Estimating the incubated river water quality indicator based on machine learning and deep learning paradigms: BOD_5 Prediction

Sungwon Kim[1,*], Meysam Alizamir[2], Youngmin Seo[3], Salim Heddam[4], Il-Moon Chung[5], Young-Oh Kim[6], Ozgur Kisi[7] and Vijay P. Singh[8]

[1] Department of Railroad Construction and Safety Engineering, Dongyang University, Yeongju, 36040, Republic of Korea
[2] Department of Civil Engineering, Hamedan Branch, Islamic Azad University, Hamedan, Iran
[3] Department of Constructional and Environmental Engineering, Kyungpook National University, Sangju, 37224, Republic of Korea
[4] Faculty of Science, Agronomy Department, Hydraulics Division, Laboratory of Research in Biodiversity Interaction Ecosystem and Biotechnology, University 20 Août 1955, Route El Hadaik, BP 26, Skikda, Algeria
[5] Department of Hydro Science and Engineering Research, Korea Institute of Civil Engineering and Building Technology, Goyang-si 10223, Republic of Korea
[6] Department of Civil Engineering, Seoul National University, Seoul, Republic of Korea
[7] Department of Civil Engineering, University of Applied Sciences, 23562, Lübeck, Germany
[8] Department of Biological and Agricultural Engineering & Zachry Department of Civil Engineering, Texas A&M University, College Station, Texas, 77843-2117, USA

* **Correspondence:** Email: swkim1968@dyu.ac.kr.

Abstract: As an indicator measured by incubating organic material from water samples in rivers, the most typical characteristic of water quality items is biochemical oxygen demand (BOD_5) concentration, which is a stream pollutant with an extreme circumstance of organic loading and controlling aquatic behavior in the eco-environment. Leading monitoring approaches including machine leaning and deep learning have been evolved for a correct, trustworthy, and low-cost prediction of BOD_5 concentration. The addressed research investigated the efficiency of three standalone models including machine learning (extreme learning machine (ELM) and support vector regression (SVR)) and deep learning (deep echo state network (Deep ESN)). In addition, the novel double-stage synthesis models (wavelet-extreme learning machine (Wavelet-ELM), wavelet-support vector regression (Wavelet-SVR), and

wavelet-deep echo state network (Wavelet-Deep ESN)) were developed by integrating wavelet transformation (WT) with the different standalone models. Five input associations were supplied for evaluating standalone and double-stage synthesis models by determining diverse water quantity and quality items. The proposed models were assessed using the coefficient of determination (R^2), Nash-Sutcliffe (NS) efficiency, and root mean square error (RMSE). The significance of addressed research can be found from the overall outcomes that the predictive accuracy of double-stage synthesis models were not always superior to that of standalone models. Overall results showed that the SVR with 3^{th} distribution (NS = 0.915) and the Wavelet-SVR with 4^{th} distribution (NS = 0.915) demonstrated more correct outcomes for predicting BOD_5 concentration compared to alternative models at Hwangji station, and the Wavelet-SVR with 4^{th} distribution (NS = 0.917) was judged to be the most superior model at Toilchun station. In most cases for predicting BOD_5 concentration, the novel double-stage synthesis models can be utilized for efficient and organized data administration and regulation of water pollutants on both stations, South Korea.

Keywords: biochemical oxygen demand; wavelet transformation; deep echo state network; extreme learning machine; support vector regression

1. Introduction

The role of water quality in streams, lakes, and seas can be stated as organic, synthetic, and environmental condition of waterbody [1,2]. The items for water quality include the diverse features such as dissolved oxygen (DO), chemical oxygen demand (COD), biochemical oxygen demand (BOD_5), total organic carbon (TOC), total phosphorus (T-P), total nitrogen (T-N), suspended solids (SS), turbidity (TU), potential of Hydrogen (pH), electrical conductivity (EC), water temperature (WT), and chlorophyll-a (Chl-*a*) and so on. The quantitative evaluation of water quality items is significant for the transaction of integrated water resources [3].

The water quality items can be identified in three ways: situ-measurement class (e.g., TU, pH, EC, DO, and WT), lab-measurement class (e.g., T-P, T-N, TOC, SS, COD, and Chl-*a*), and incubated-measurement class (e.g., BOD_5). Among water quality items, BOD_5, the representative incubated-measurement indicator, was considered as a reference to appraise the organic pollution of waterbody by the American Public Health Association Standard Methods Committee (APHASMC) [4]. Also, the concentration of BOD_5 can be recommended as the necessity of DO to cut down the organic matter of fluid at specific temperature [5]. It can, therefore, be estimated using the quantity of oxygen used up per liter of inspected data dependent on the 5-day period at 20 Celsius (°C) [6], and was assessed as one of essential river water quality items for the preservation and management of eco-environmental systems [7].

Although different machine learning and deep learning paradigms have been implemented for estimating the incubated-measurement indicator in rivers, this article proposes the unique technique for the accurate prediction of BOD_5 concentration. Hybrid neuroscience approaches involving the diverse data preprocessing coupled with the neuroscience techniques promote the evolution of more complex models based on the higher precision of estimated problems in natural behavior [2,8]. The double-stage synthesis models, one of hybrid neuroscience approaches, combining the wavelet transformation (WT) and different neuroscience models were developed and implemented to boost the

predictive accuracy of BOD_5 concentration in Hwangji and Toilchun stations, South Korea. The standalone models such as extreme learning machine (ELM), support vector regression (SVR), and deep echo state network (Deep ESN) were also employed for integrating and evaluating novel double-stage synthesis model's scheme clearly. The novel double-stage synthesis models (i.e., Wavelet-ELM, Wavelet-SVR, and Wavelet-Deep ESN), therefore, demonstrates the efficient and accurate estimation of highly complex and nonstationary problem in rivers. The distinguished attraction of double-stage synthesis models motivates to explore the accurate prediction of BOD_5 concentration.

To the best of our knowledge and recognition from the previous information such as published articles, documents and reports, the double-stage synthesis models in the addressed article have not been frequently implemented for predicting BOD_5 concentration among the various water quality items. This article discusses the performance of implemented models (ELM, SVR, Deep ESN, Wavelet-ELM, Wavelet-SVR, and Wavelet-Deep ESN) for predicting BOD_5 concentration. They are evaluated by utilizing three mathematical formulae (R^2, NS, and RMSE) and four graphical aids (Scatter diagram, boxplot, violin plot, and Taylor diagram), respectively.

The rest of addressed research is arranged as follows. A brief review of BOD_5 concentration estimation and prediction is presented in section 2. The detailed description of machine learning and deep learning paradigms are provided in section 3. Also, the wavelet transformation is discussed. In section 4, report for data available and the criteria of model assessment are provided in detail. In section 5, a case study is presented by using the standalone and double-stage synthesis models based on water quantity and quality items collected in Hwangji and Toilchun stations, South Korea. In section 6, the advantages of standalone and double-stage synthesis models using mathematical formulae and graphical aids are discussed. In the end, the conclusions are drawn up.

2. Literature review on BOD_5 concentration estimation and prediction

Various machine learning and deep learning paradigms for the estimation and prediction issues of water quality have been extensively reported in numerous articles and documents. [9] developed the hybrid model utilizing SVR and firefly algorithm (FFA) for predicting water quality indicator in the Euphrates River, Iraq. They found that the SVR-FFA model could predict the water quality indicator accurately. [10] implemented four standalone and twelve hybrid models to predict the Iran water quality indicator. The BA-RT, one of hybrid models, provided the best performance to predict the Iran water quality indicator. [11] employed seven standalone and three hybrid models to predict the diverse water quality indicators in China. Results showed that the decision tree (DT), random forest (RF), and deep cascade forest (DCF) models produced the outstanding achievements to predict water quality indicators in major rivers and lakes. [12] reviewed the recent advances in water quality remote sensing system using 200 datasets of water quality indicators. They demonstrated that the deep learning model outperformed the other proposed models to predict water quality indicators in Midwestern United States. [13] investigated the ELM, RF, group method of data handling (GMDH), classification and regression tree (CART), and Bat-ELM models to predict the chlorophyll-a concentration in river and lake systems, USA. They concluded that the Bat-ELM model predicted the chlorophyll-a concentration precisely compared to other models. [14] proposed the deep learning models including the recurrent neural network (RNN), long-short term memory (LSTM), and gated recurrent unit (GRU) to predict the drainage water quality indicator in Southern China. They showed that the deep learning models produced better prediction compared to the multiple linear regression (MLR) and multilayer perception (MLP) models.

However, limited techniques and methods have been implemented to estimate and predict BOD_5 concentration [15–18]. [19] employed the regression tree (RT) and SVR models to estimate total suspended solids (TSS), total dissolved solid (TDS), COD, and BOD_5 concentration using the datasets from National Stormwater Quality Database (NSQD), USA. Results showed that the applied models could estimate BOD_5 concentration accurately. [20] developed the adaptive neuro-fuzzy inference system (ANFIS) and wavelet SVR (WSVR) models to predict BOD_5 concentration in Karun River, Iran. They demonstrated that the WSVR model provided better prediction compared to the ANFIS model. [1] estimated BOD_5 concentration employing the RF, gradient boosting regression tree (GBRT), ELM, and Deep ESN in the Han River, South Korea. It can be found from [1]'s article that the Deep ESN5 model supplied the most accurate predictions of BOD_5 concentration among the developed models. Also, [2] developed two-stage and standalone neuroscience models to predict BOD_5 concentration in the Nakdong River, South Korea. Considering the developed models, the DWT-RF5 and DWT-GRNN4 models were the best model for predicting BOD_5 concentration. [21] utilized the SVR, RF, artificial neural networks (ANNs), long short-term memory (LSTM), convolutional neural networks (CNN)-LSTM, and Bi-LSTM models for forecasting COD and BOD_5 concentrations in the Yamuna River, India. This investigation provided that the Bi-LSTM model supplied the best performance for forecasting COD and BOD_5 concentrations. [22] implemented four standalone (ANN, RF, support vector machines (SVMs), and gradient boosting machines (GBM)) and six hybrid (RF-SVMs, ANN-SVMs, GBM-SVMs, RF-ANNs, GBM-ANNs, and RF-GBM) models to predict BOD_5 concentration in the Buriganga River, Bangladesh. They found that the RF-SVMs model provided the best predictive accuracy among the developed models. In contrast with the above-mentioned machine learning and deep learning paradigms, the novel double-stage synthesis models were introduced to find the optimal models between BOD_5 concentration and well-known water quality items based on five input associations. The addressed research can highlight how the novel double-stage synthesis models enhance the predictive results of BOD_5 concentration.

3. Implemented models and supplementary method

The implemented models in the addressed article were machine learning (ELM and SVR) and deep learning (Deep ESN) paradigms, and the supplementary method was classified as the wavelet transformation, which is one of data preprocessing techniques used in various research fields. It can be seen from Figure 1 that the comprehensive mechanism of research process is underlined. Successive sub-phases explain the implemented models and supplementary method.

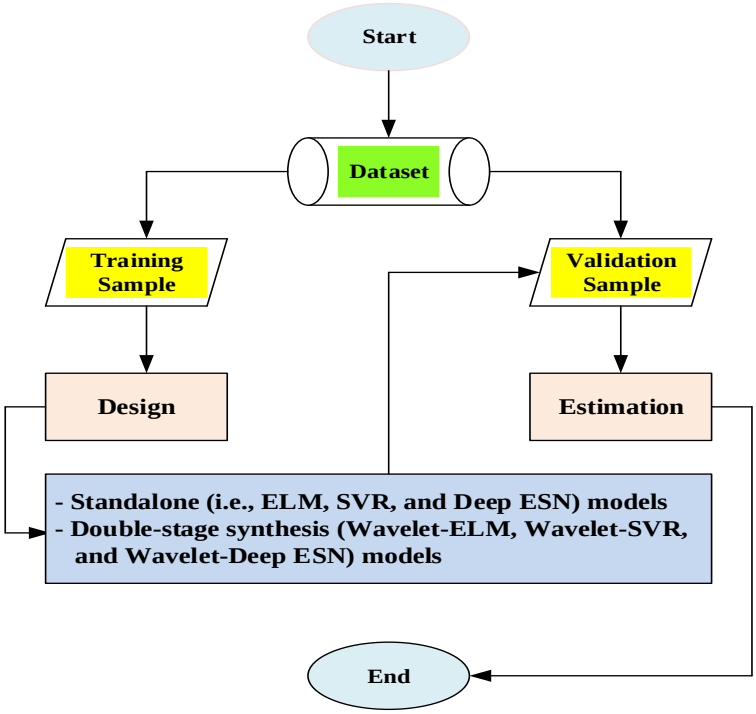


Figure 1. Comprehensive mechanism of research process.

3.1. Extreme Learning Machine (ELM)

[23] initially recommended the ELM model as a rapid and effective category of feedforward neural networks (FFNN) (refer to Figure 2). It involves a single-middle-layer, which receives a particular scheme for training the parameters of networks compared to the conventional multilayer perceptron (MLP) model. The ELM model can map using a single-middle-layer with M independent input indicators and be written as follows:

$$f(x) = \sum_{j=1}^{M} \sum_{i=1}^{L} \beta_i g_i (\mathrm{w}_i x_j + b) \tag{1}$$

where *g(.)* is the activation function, which supplies the output in the middle layer; β_i is the weight of output for connecting the middle neurons to the output neuron; and L is the neuron number of the middle layer. The output indicator can be given by the following formula (2):

$$y = \sum_{j=1}^{M} \sum_{i=1}^{L} \beta_i g_i (\mathrm{w}_i x_j + b) = t + \varepsilon \tag{2}$$

where ε is the error. The Gaussian and sigmoid functions are the most employed mapping ones in the ELM model's category. The underlying formula (3) expresses the Gaussian function:

$$g(x_i) = h(a, c, x_i) = \exp(-a \| x_i - c \|^2) \tag{3}$$

where *a* and *c* refer to the activation functions. During training phase, the connection weight is fixed in the ELM model's category. That is, random values are allowed directly to neurons' activation functions instead of requesting an iterative process to update them. The connection weights for output neuron can be achieved continuously utilizing the least squares method. In other words, the fitting error ought to be reduced by computing $\|\mathbf{H\beta} - \mathbf{T}\|^2$ for the connection weight ($\boldsymbol{\beta}$), where $\mathbf{T}$ is the matrix for target and $\mathbf{H}$ is the randomized matrix corresponding to the middle layer:

$$\mathbf{H} = \begin{bmatrix} g(x_1) \\ \cdot \\ \cdot \\ g(x_N) \end{bmatrix}, \quad \mathbf{T} = \begin{bmatrix} t_1^T \\ \cdot \\ \cdot \\ t_N^T \end{bmatrix} \tag{4}$$

The connection weight for output is resolved, based on the linear equation system such as $\boldsymbol{\beta}=\mathbf{H}^{+}\mathbf{T}$, where $\mathbf{H}^{+}$ is the generalized inverse function of Moore-Penrose [1,24].

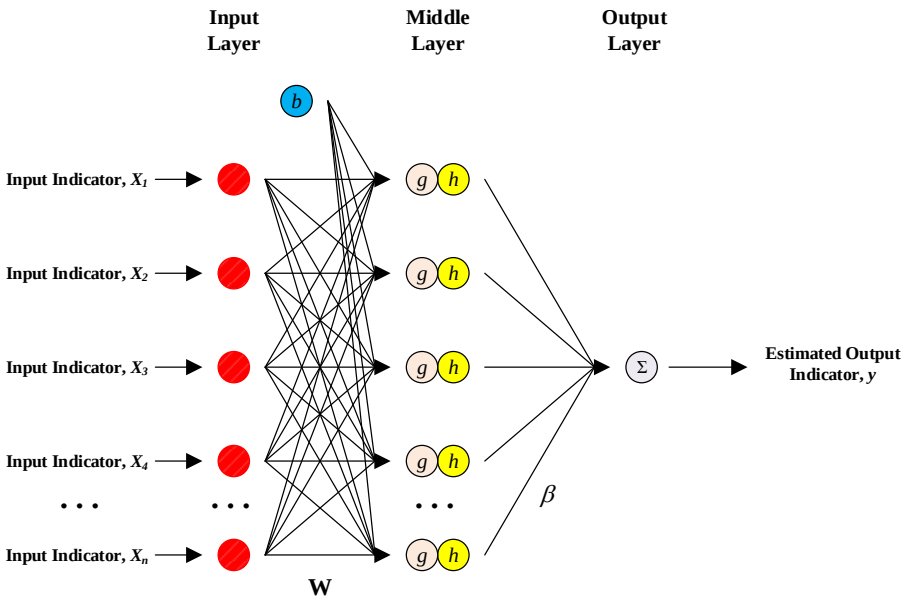


Figure 2. A schematic diagram of extreme learning machine (ELM) model.

3.2. Support Vector Regression (SVR)

The SVR model (refer to Figure 3), which is a special type of SVMs, has been applied in various fields, including stock index prediction, bioinformatics engineering, chemical synthesis, and production process control and so on [25,26]. The generalization of conventional ANNs models may reduce to a local optimized generalization, while a universal optimization is insured for the SVR model [27–29].

The fundamental concepts of SVR model are as follows. Recognizing the training sample (x_i, y_i),

where $x_i \in \Re^n$ is a specific value of input indicator x, and $y_i \in \Re^n$ is the matching value of surveyed model output. Also, a nonlinear transfer function ($\Phi(\cdot)$) and a linear function ($f(\cdot)$) can be defined between input and output indicators. The actual output, therefore, is expressed by formula (5):

$$\bar{y} = f(x) = w^T \Phi(x) + B \tag{5}$$

where $\bar{y}$ is the actual output; and w and B are the adjustable parameters of the model. In the SVR model, the empirical risk can be written as the following formula (6):

$$R_{emp} = \frac{1}{N} \sum_{i=1}^{N} \left| y_i - \bar{y}_i \right|_\varepsilon \tag{6}$$

where R_{emp} is the empirical risk; and $\left| y_i - \bar{y}_i \right|_\varepsilon$ is Vapnik's ε-insensitive loss function. The adjustable parameters (i.e., w and B in formula (5)) of the model can be calculated by obtaining the minimum cost function [27]. In the addressed article, the following cost function was used:

$$\psi_\varepsilon(w, \xi, \xi^*) = \frac{1}{2} w^T w + C \sum_{i=1}^{N} (\xi_i + \xi_i^*) \tag{7}$$

where $\psi_\varepsilon(w, \xi, \xi^*)$ is the cost function; ξ_i, ξ_i^* are the positive slack variables; and C is the cost constant. In addition, the constraints for formula (7) can be classified as: (1) $y_i - \bar{y}_i \le \varepsilon + \xi_i \quad i = 1, 2, \ldots, N$; (2) $-y_i + \bar{y}_i \le \varepsilon + \xi_i^* \quad i = 1, 2, \ldots, N$; and (3) $\xi_i, \xi_i^* \ge 0 \quad i = 1, 2, \ldots, N$.

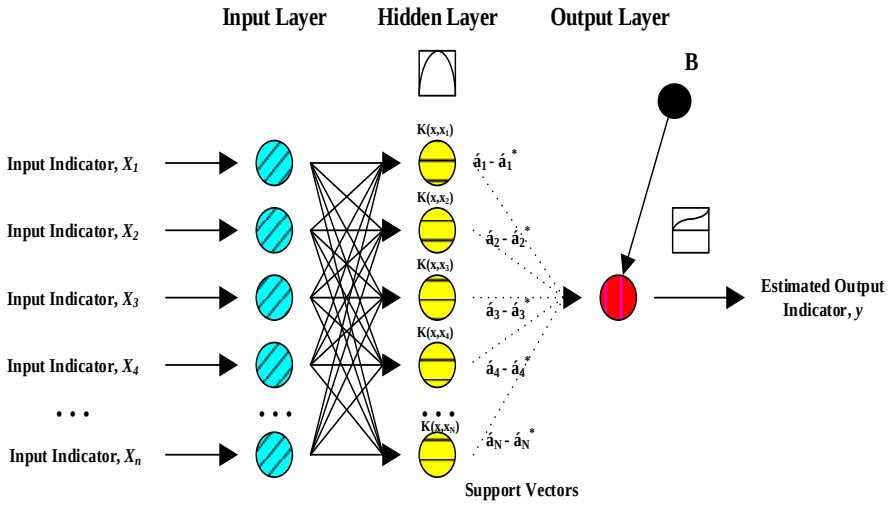


Figure 3. A schematic diagram of support vector regression (SVR) model.

3.3 Deep Echo State Network (Deep ESN)

The recurrent neural networks (RNN), including the echo state network (ESN), has the most broadly utilized reservoir computing (RC) method [1,31,32]. Since the RNN model is powerful and accurate for computing the complicated and nonlinear problems, the Deep ESN model is effective for historical data. The Deep ESN model contains an order of deformed recurrent layers named as reservoir, where every layer output performs as the following layer input. Figure 4 explains a conceptual diagram of Deep ESN model with the recurrent structure of reservoir from the viewpoint of a discrete-time dynamic system. The reservoir dynamic state can be renewed by recognizing the leaky integration ESN (i.e., LI-ESN) as below:

$$x(t) = (1-\alpha)\,x(t-1) + \alpha \tanh(W_{in}u(t) + W_R\,x(t-1)) \tag{8}$$

where α is the leaky coefficient; *u(t)* is the outside input based on time *t*; *x(t)* is the reservoir state in the corresponding layer based on time *t*; W_{in} is the matrix of input connection weight for the reservoir; and W_R is the matrix of recurrent connection weight. Because the input indicator to the following reservoir can be supplied by the output indicator of its prior reservoir, an ordinary equation for the Deep ESN model is organized for expanding the function of state transition as below.

$$\begin{aligned} &x^{\ell}(t) = (1-\alpha^{\ell})\,x^{\ell}(t-1) + \alpha^{\ell}\tanh(W^{\ell}i^{\ell}(t) + W^{\ell}{}_{R}\,x^{\ell}(t-1));\\ &i^{0}(t) = u(t)\ \&\ i^{\ell}(t) = x^{\ell-1}(t);\quad \ell = 1,2,...,L \end{aligned} \tag{9}$$

where ℓ is a layer (reservoir) in the structure of RC; $W^{\ell}{}_{R}$ are the connection weights between the layer ℓ and the prior one $\ell-1$; and *L* is the quantity for the layers of reservoir [1,32,33]. Here, five layers were employed in the reservoir for the Deep ESN model to estimate BOD_5 concentration.

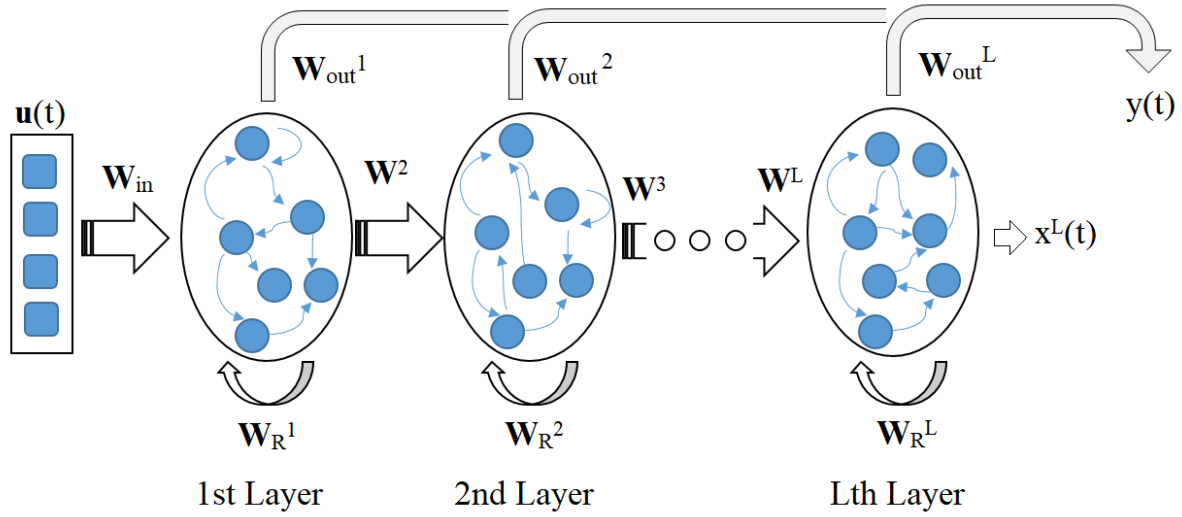


Figure 4. A conceptual diagram of deep echo state network (Deep ESN) model [1].

3.4 Wavelet Transformation (WT)

The WT method, which is one of multi-resolution signal procedure methods, is employed to build the double-stage synthesis models. The original data can be isolated into various frequency components involving an approximation and numerous details handling the WT method. In the addressed article, discrete-based wavelet transformation, which has been utilized for the data preprocessing in diverse fields, was selected. In fact, the discrete-based WT method can be accomplished by implementing the Mallat method [34]. The bottom line of Mallat method is two-route filters comprising two filters as low-pass and high-pass [2,35]. [36] defined that the coefficients for the wavelet (high-pass) and scaling (low-pass) in the j^{th} level of decomposition is outlined as

$$W_{j,t} \equiv \sum_{l=0}^{L-1} h_l V_{j-1,\, 2t+1-l \bmod N_{j-1}}, \quad V_{j,t} \equiv \sum_{l=0}^{L-1} g_l V_{j-1,\, 2t+1-l \bmod N_{j-1}}, \quad t = 0, 1, \cdots, N_j - 1 \tag{10}$$

where $W_{j,t}$ and $V_{j,t}$ are the elements for corresponding $\mathbf{W}_j$ and $\mathbf{V}_j$. The WT decomposes the complex and original input time series into the components (approximation and details) which show relatively simpler patterns than the original input time series. The different components obtained from WT were implemented as input association of corresponding double-stage synthesis model. Evolving double-stage synthesis models for the components separately and summing their predicted values can improve the predictive accuracy of double-stage synthesis models compared to performance of standalone models for the original input time series with high complexity. A flowchart for dual-step discrete-based WT is shown in figure 5. Here, two details (D_1 and D_2) and an approximation (A_2) are achieved from the original input time series. Also, Figure 6 illustrates the sequential diagram for evolving the double-stage synthesis models.

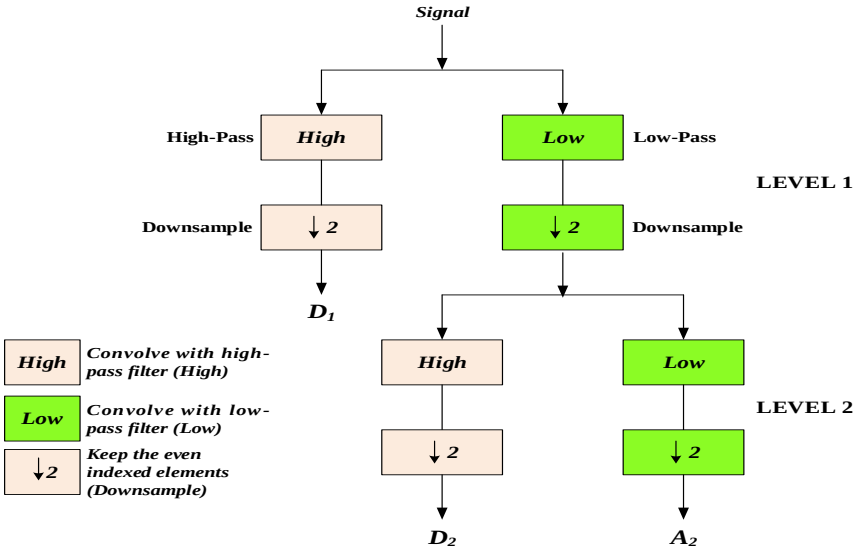


Figure 5. Dual-step discrete-based WT decomposition.

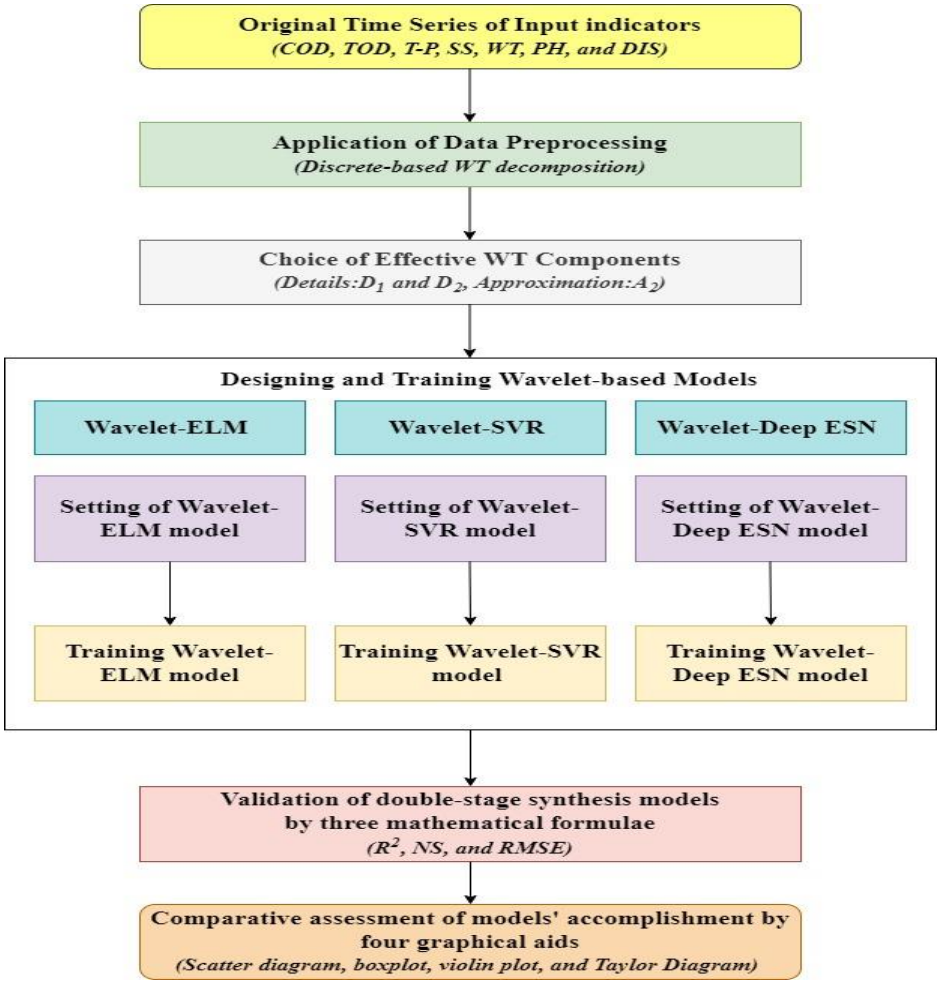


Figure 6. Sequential diagram for evolving double-stage synthesis models.

4. Report for data and assessment criteria

4.1.Preparation of utilized data

The original data can be isolated into various frequency In the addressed article, Hwangji (longitude 129°05′07″E; latitude 37°06′74″N) and Toilchun (longitude 128°44′46″E; latitude 36°47′09″N) stations were employed to predict BOD_5 concentration using diverse physical and chemical variables such as T-N, T-P, TOC, DO, WT, SS, COD, pH, EC, and station discharge (DIS) in South Korea. Figure 7 shows the illustrative map of Hwangji and Toilchun stations.

The surveyed data (2008/02–2020/12 for Hwangji and 2011/07–2020/12 for Toilchun stations) for water quantity and quality items can be directly accessed and collected from official website (http://water.nier.go.kr) of National Institute of Environmental Research (NIER), South Korea. The full data file consisted of training and validation samples. The training sample involved 80% (data = 398 from Hwangji and data = 294 from Toilchun stations) and the validation sample applied the last 20% (data = 99 from Hwangji and data = 74 from Toilchun stations) of full data file.

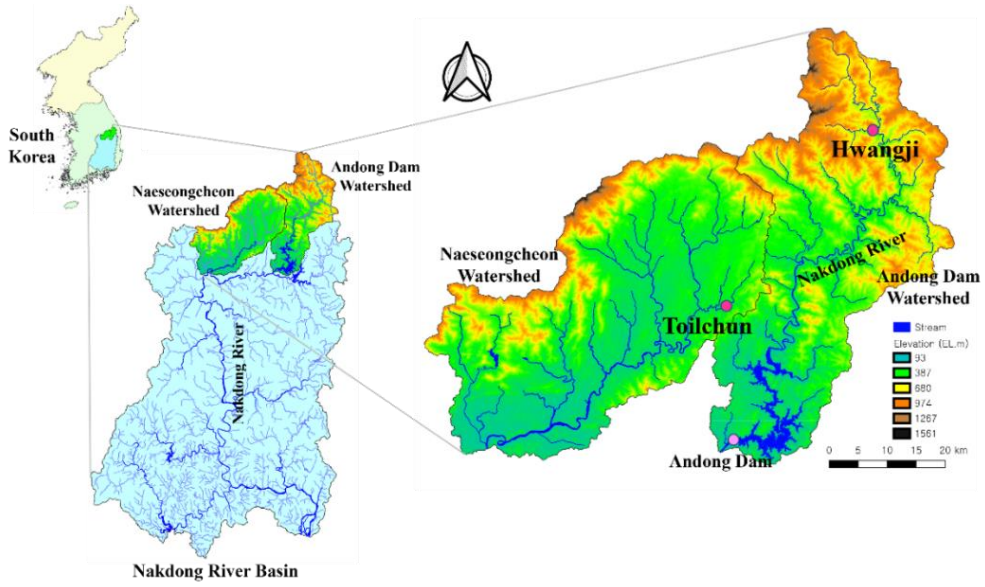


Figure 7. Illustrative map of Hwangji and Toilchun stations.

Recognizing the source code and software of machine learning and deep learning paradigms, the ELM model was evolved by employing the R (https://www.r-project.org, a free software environment for statistical computing and graphics) package and the *elmNNRcpp* (https://cran.r-project.org/web/packages/elmNNRcpp/index.html). In case of SVR model, it was implemented by the DTREG predictive modeling software (www.dtreg.com). In addition, the Deep ESN model was developed by utilizing the MATLAB programing language (https://www.mathworks.com), which is a freely available MATLAB toolbox for the Deep ESN (https://it.mathworks.com/matlabcentral/fileexchange/69402-deepesn).

The optimal number of hidden nodes for the ELM and Wavelet-ELM models was determined using a trial and error process. As the number of hidden nodes was changed from 1 to 5m (where, m is the number of input indicators), the number of hidden nodes with the minimum RMSE value was decided as the optimal value. The logistic sigmoid function and linear function were used for activating hidden and output nodes, respectively. In addition, epsilon type of SVR model with radial basis function (RBF) kernel was employed for predicting BOD_5 concentration using the SVR and Wavelet-SVR models. Also, the V-fold cross-validation were applied to validate the SVR and Wavelet-SVR models, and the grid search algorithm found the optimal parameters by minimizing total errors. Finally, the optimal number of layers and reservoirs units were decided based on the trial and error process for the Deep ESN and Wavelet-Deep ESN models.

Table 1 explains the computed results for the correlation coefficients and P values between individual input indicators and BOD_5 concentration. It can be judged from Table 1 that TOC (e.g., CC = 0.721, P value = 0.0001 at Hwangji and CC = 0.563, P value = 0.0001 at Toilchun stations) and COD (e.g., CC=0.721, P value =0.0001 at Hwangji and CC = 0.626, P value = 0.0001 at Toilchun stations) items exhibited high correlation and statistically significant with BOD_5 concentration among various input indicators. In the addressed article, all indicators can be categorized as class 1 (i.e., in situ-measurement items (pH, EC, DO, and WT), class 2 (i.e., lab-measurement items (SS, COD, T-N, T-P, and TOC) and incubated-measurement item (BOD_5)), and class 3 (i.e., water quantity item (river discharge)), respectively.

Table 1. Correlation coefficients and P values between corresponding input indicators and BOD_5 concentration.

Class	Input indicators	BOD_5 concentration			
		Hwangji		Toilchun	
		CC	P-value	CC	P-value
1	pH	-0.003	0.9491	0.073	0.1024
	EC	0.088	0.0494	-0.262	0.0001
	DO	0.036	0.4185	-0.064	0.1565
	WT	-0.074	0.0974	0.123	0.0058
2	SS	0.120	0.0069	0.462	0.0001
	COD	0.721	0.0001	0.626	0.0001
	T-N	0.163	0.0003	-0.195	0.0001
	T-P	0.349	0.0001	0.479	0.0001
	TOC	0.721	0.0001	0.563	0.0001
3	DIS	-0.042	0.3494	0.184	0.0001

4.2. Mathematical assessment criteria of standalone and double-stage synthesis models

To assess the performance of standalone (ELM, SVR, and Deep ESN) and double-stage synthesis (Wavelet-ELM, Wavelet-SVR, and Wavelet-Deep ESN) models, three mathematical formulae, which have been recognized and utilized worldwide, were employed. The coefficient of determination (R^2) criterion [37,38] is clarified as the square of correlation between surveyed and estimated BOD_5 concentrations (see formula (11)). The Nash-Sutcliffe (NS) efficiency criterion [39] can resolve the models' effectiveness between surveyed and estimated BOD_5 concentrations (see formula (12)). Also, the disparity between surveyed and estimated BOD_5 concentrations can be referred by handling the root mean square error (RMSE) criterion [40]. The RMSE criterion can be computed by employing formula (13).

$$R^2 = \left(\frac{\frac{1}{n}\sum_{i=1}^{n}(BOD_{sur} - \overline{BOD}_{sur})(BOD_{est} - \overline{BOD}_{est})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(BOD_{sur} - \overline{BOD}_{sur})^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(BOD_{est} - \overline{BOD}_{est})^2}} \right)^2 \tag{11}$$

$$NS = 1 - \frac{\sum_{i=1}^{n}[BOD_{sur} - BOD_{est}]^2}{\sum_{i=1}^{n}[BOD_{sur} - \overline{BOD}_{sur}]^2} \tag{12}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}[BOD_{sur} - BOD_{est}]^2} \tag{13}$$

where BOD_{sur} and BOD_{est} = surveyed and estimated BOD_5 concentrations; $\overline{BOD}_{sur}$ and $\overline{BOD}_{est}$ = surveyed and estimated mean BOD_5 concentrations; and n = the number of full data available.

5. Case study

The addressed article utilized the miscellaneous water quantity and quality items to predict BOD_5 concentration in Hwangji and Toilchun stations, South Korea. As defined formerly, the assessment of standalone and double-stage synthesis models to estimate BOD_5 concentration was the essential view of this article.

Among water quality items, some indicators, including pH, EC, DO, and WT, were directly surveyed by utilizing a commercial mechanical tool. Also, the indicators including SS, COD, T-N, T-P, and TOC were surveyed through the laboratory assistant system indirectly. BOD_5 concentration, however, can be indirectly surveyed via the incubation system, based on the 20 ℃ for the 5-day period [41]. Since the plan of addressed article was scheduled to predict BOD_5 concentration utilizing the standalone and double-stage synthesis models, this behavior could save and protect the time and effort to estimate and incubate BOD_5 concentration.

From the correlation coefficients and P values of water quantity and quality items (refer to Table 1), divergent organizations were provided to choose the best input association for given stations. To employ the same input indicators on both stations among them, some input indicators with positive (e.g., COD, TOC, T-P, and SS) and negative (e.g., WT, pH, and DIS) correlations were selected for diverse input associations in Hwangji station because input indicators with negative correlations can also contribute for predicting BOD_5 concentration. However, many input indicators based on positive (e.g., COD, TOC, T-P, SS, WT, pH, and DIS) correlation were implemented for different input associations in Toilchun station.

Hence, the standalone and double-stage synthesis models were evolved for predicting BOD_5 concentration, based on five input associations (so called, 1^{st}–5^{th} distributions). Because TOC and COD items were picked out as the underlying water quality items for given stations, the addressed article determined the consolidation of TOC and COD items as the 1^{st} distribution. Table 2 presents the diverse input associations of water quantity and quality items to predict BOD_5 concentration. All developed models in Table 2 can be categorized into five distributions.

5.1. Predicting BOD_5 concentration at Hwangji station

5.1.1. Application of standalone models

The results of three mathematical formulae (R^2, NS, and RMSE) for the standalone models are summed up in Table 3 for Hwangji station. Table 3 explains that the outcomes of SVR1 ($R^2 = 0.905$, NS = 0.891, and RMSE = 0.299 mg/L) are more excellent than the ELM1 and Deep ESN1 in the validation phase dependent on the 1^{st} distribution. In the 2^{nd} distribution, the SVR2 ($R^2 = 0.908$, NS = 0.905, and RMSE = 0.279 mg/L) performs more excellent than the ELM2 and Deep ESN2. And, the SVR3 ($R^2 = 0.925$, NS = 0.915, and RMSE = 0.264 mg/L) surpasses the ELM3 and Deep ESN3 clearly in the validation phase for the 3^{rd} distribution. The contrast of standalone models in the 4^{th} distribution, furthermore, indicates that the ELM4 ($R^2 = 0.902$, NS = 0.893, and RMSE = 0.295 mg/L) dominates the SVR4 and Deep ESN4 in the validation phase. In the end, the SVR5 ($R^2 = 0.905$, NS = 0.884, and RMSE = 0.309 mg/L) is more accurate than ELM5 and Deep ESN5 in the validation phase for the 5^{th} distribution.

Recognizing the impressive models from the 1^{st}–5^{th} distributions, the best accomplishment of standalone models can be found from the ELM (the 4^{th} distribution), SVR (the 3^{rd} distribution), and Deep ESN (the 2^{nd} distribution) among diverse input associations in the validation phase. Table 3 tells

us that the desirable performance of SVR3 gives more accurate than ELM4 and Deep ESN2 in the validation phase. Therefore, it can be said that the SVR3 is the most accurate for predicting BOD_5 concentration among the desirable standalone models at Hwangji station.

Table 2. Diverse input associations of developed models for predicting BOD_5 concentration.

Classification	Division	Model	Distribution	Input association
Standalone	Machine learning	ELM	ELM1	COD+TOC
			ELM2	COD+TOC+T-P+SS
			ELM3	COD+TOC+WT+pH
			ELM4	COD+TOC+T-P+SS+WT+pH
			ELM5	COD+TOC+T-P+SS+WT+pH+DIS
		SVR	SVR1	COD+TOC
			SVR2	COD+TOC+T-P+SS
			SVR3	COD+TOC+WT+pH
			SVR4	COD+TOC+T-P+SS+WT+pH
			SVR5	COD+TOC+T-P+SS+WT+pH+DIS
	Deep learning	Deep ESN	Deep ESN1	COD+TOC
			Deep ESN2	COD+TOC+T-P+SS
			Deep ESN3	COD+TOC+WT+pH
			Deep ESN4	COD+TOC+T-P+SS+WT+pH
			Deep ESN5	COD+TOC+T-P+SS+WT+pH+DIS
Double-stage synthesis	Machine learning	Wavelet-ELM	Wavelet-ELM1	COD+TOC
			Wavelet-ELM2	COD+TOC+T-P+SS
			Wavelet-ELM3	COD+TOC+WT+pH
			Wavelet-ELM4	COD+TOC+T-P+SS+WT+pH
			Wavelet-ELM5	COD+TOC+T-P+SS+WT+pH+DIS
		Wavelet-SVR	Wavelet-SVR1	COD+TOC
			Wavelet-SVR2	COD+TOC+T-P+SS
			Wavelet-SVR3	COD+TOC+WT+pH
			Wavelet-SVR4	COD+TOC+T-P+SS+WT+pH
			Wavelet-SVR5	COD+TOC+T-P+SS+WT+pH+DIS
	Deep learning	Wavelet-Deep ESN	Wavelet-Deep ESN1	COD+TOC
			Wavelet-Deep ESN2	COD+TOC+T-P+SS
			Wavelet-Deep ESN3	COD+TOC+WT+pH
			Wavelet-Deep ESN4	COD+TOC+T-P+SS+WT+pH
			Wavelet-Deep ESN5	COD+TOC+T-P+SS+WT+pH+DIS

Table 3. Results of three mathematical criteria using the standalone models at Hwangji station.

Classification	Distribution	Validation phase		
		R^2	NS	RMSE (mg/L)
Standalone	ELM1	0.900	0.835	0.368
	ELM2	0.895	0.879	0.315
	ELM3	0.898	0.837	0.365
	ELM4	0.902	0.893	0.295
	ELM5	0.879	0.855	0.344
	SVR1	0.905	0.891	0.299
	SVR2	0.908	0.905	0.279
	SVR3	0.925	0.915	0.264
	SVR4	0.908	0.882	0.310
	SVR5	0.905	0.884	0.309
	Deep ESN1	0.871	0.806	0.398
	Deep ESN2	0.884	0.831	0.371
	Deep ESN3	0.845	0.805	0.399
	Deep ESN4	0.857	0.769	0.434
	Deep ESN5	0.886	0.809	0.394

5.1.2. Application of double-stage synthesis models

The results of three mathematical criteria for the double-stage synthesis models are also arranged in Table 4 at Hwangji station. From Table 4, it is clear that the outcomes of Wavelet-SVR1 ($R^2 = 0.904$, NS = 0.895, and RMSE = 0.293 mg/L) are more dominant compared to the Wavelet-ELM1 and Wavelet-Deep ESN1 in the validation phase dependent on the 1st distribution. Based on the 2nd distribution, the Wavelet-SVR2 ($R^2 = 0.911$, NS = 0.911, and RMSE = 0.271 mg/L) is more excellent than the Wavelet-ELM2 and Wavelet-Deep ESN2. Also, the Wavelet-SVR3 ($R^2 = 0.920$, NS = 0.912, and RMSE = 0.269 mg/L) outperforms the Wavelet-ELM3 and Wavelet-Deep ESN3 regarding the 3rd distribution in the validation phase. Moreover, the contrast of double-stage synthesis models in the 4th distribution demonstrates that the Wavelet-SVR4 ($R^2 = 0.926$, NS = 0.915, and RMSE = 0.264 mg/L) performs superior to the Wavelet-ELM4 and Wavelet-Deep ESN4 in the validation phase. In the end, Wavelet-SVR4 ($R^2 = 0.919$, NS = 0.914, and RMSE = 0.266 mg/L) is more efficient than the Wavelet-ELM5 and Wavelet-Deep ESN5 in the validation phase for the 5th distribution.

Contemplating the outstanding models from the 1st-5th distributions, the admirable performance of double-stage synthesis models can be judged from the Wavelet-ELM (the 2nd distribution), Wavelet-SVR (the 4th distribution), and Wavelet-Deep ESN (the 3rd distribution) among diverse input associations in the validation phase. It can be noticed from Table 4 that the Wavelet-SVR4 provides more effective outcomes than the Wavelet-ELM2 and Wavelet-Deep ESN3 in the validation phase. For that reason, the Wavelet-SVR4 is more trustworthy than the Wavelet-ELM2 and Wavelet-Deep ESN3 for predicting BOD_5 concentration among the desirable double-stage synthesis models at Hwangji station.

Table 4. Results of three mathematical criteria using the double-stage synthesis models at Hwangji station.

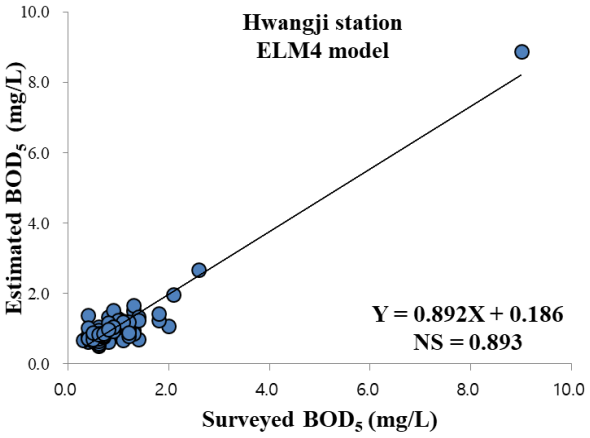
Classification	Distribution	Validation phase		
		R^2	NS	RMSE (mg/L)
Double-stage synthesis	Wavelet-ELM1	0.837	0.776	0.428
	Wavelet-ELM2	0.816	0.812	0.393
	Wavelet-ELM3	0.831	0.772	0.432
	Wavelet-ELM4	0.796	0.777	0.427
	Wavelet-ELM5	0.734	0.717	0.481
	Wavelet-SVR1	0.904	0.895	0.293
	Wavelet-SVR2	0.911	0.911	0.271
	Wavelet-SVR3	0.920	0.912	0.269
	Wavelet-SVR4	0.926	0.915	0.264
	Wavelet-SVR5	0.919	0.914	0.266
	Wavelet-Deep ESN1	0.869	0.826	0.377
	Wavelet-Deep ESN2	0.863	0.832	0.370
	Wavelet-Deep ESN3	0.860	0.833	0.369
	Wavelet-Deep ESN4	0.846	0.815	0.388
	Wavelet-Deep ESN5	0.851	0.817	0.386

5.1.3 Graphical aids of model accomplishment

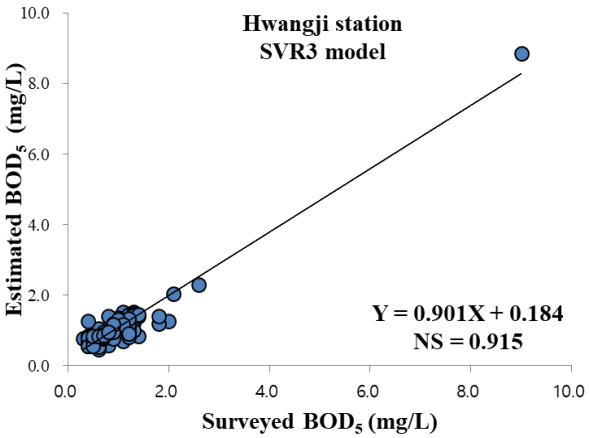
To verify the accuracy of desirable standalone and double-stage synthesis models using graphical aids, Figures 8(a)–(f) present the scatterplots for the surveyed and estimated BOD_5 concentration values at Hwangji station. The linear functions and values of NS efficiency criterion are presented for the corresponding standalone and double-stage synthesis models, respectively. It can be concluded from NS values and the slopes of linear functions that an apparent discrepancy can be followed among the desirable standalone and double-stage synthesis models (ELM4, SVR3, Deep ESN2, Wavelet-ELM2, Wavelet-SVR4, and Wavelet-Deep ESN3). Therefore, the SVR3 and Wavelt-SVR4 performs the most reliable accuracy for predicting BOD_5 concentration values clearly, whereas the Wavelet-ELM2 was the worst among the desirable models at Hwangji station.

Additional portraits can evaluate the performance of standalone and double-stage synthesis models using the boxplot, violin plot [42], and Taylor diagram [43]. Figures 9(a)–(c) show the diverse graphical aids for the desirable standalone and double-stage synthesis models at Hwangji station. It can be found from Figure 9(a) that the estimated BOD_5 concentrations of SVR3 and Wavelet-SVR4 yield more analogous configuration to the surveyed values for median, interquartile ranges and dispersion, adjacent values, and sign of skewness compared to other desirable models. Another graphical aid for the distribution of surveyed and estimated BOD_5 concentration values utilizing the desirable models can be provided with the violin plots (Figure 9(b)). The violin plot can be defined as one of approaches to discern the distribution of assigned numerical values. Figure 9(b) supplies a close shape pattern for the SVR3 and Wavelet-SVR4 concerning the median, interquartile, and distribution of assigned values. In addition, the Taylor diagram (Figure 9(c)) utilizes three statistical indices, including correlation coefficient, normalized standard deviation, and root mean square error. The

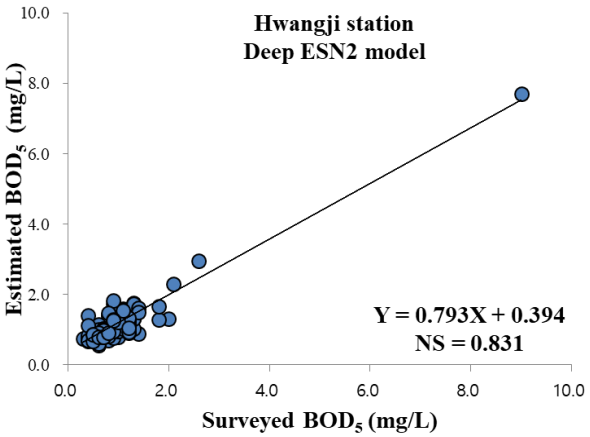
principal approach of Taylor diagram can be explained as to find the closest estimated model with the corresponding surveyed BOD_5 concentration based on standard deviation (polar axis) and correlation coefficient (radial axis). The Taylor diagram, therefore, demonstrates the accuracy and efficiency of SVR3 and Wavelet-SVR4 over the other desirable models (ELM4, Deep ESN2, Wavelet-ELM2, and Wavelet-Deep ESN3).



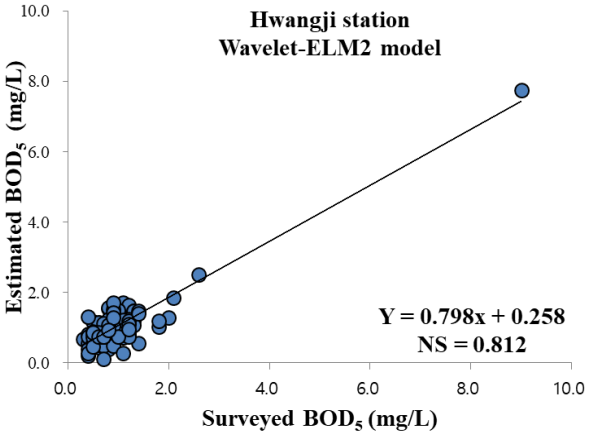
(a) ELM4 model



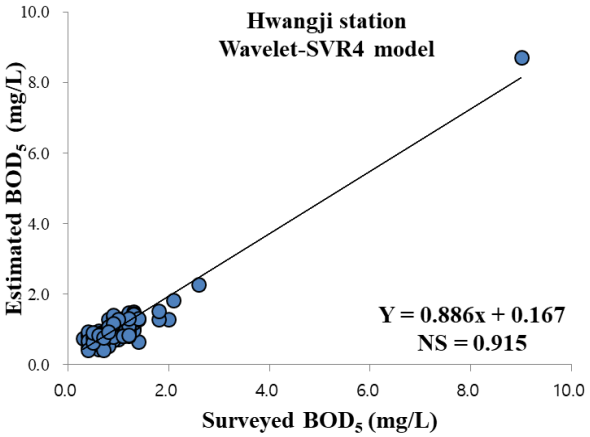
(b) SVR3 model



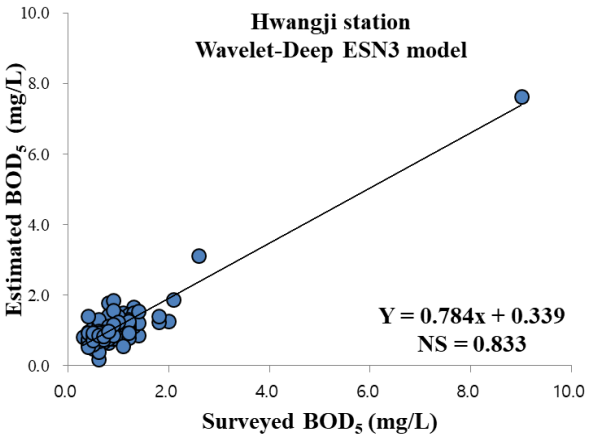
(c) Deep ESN2 model



(d) Wavelet-ELM2 model

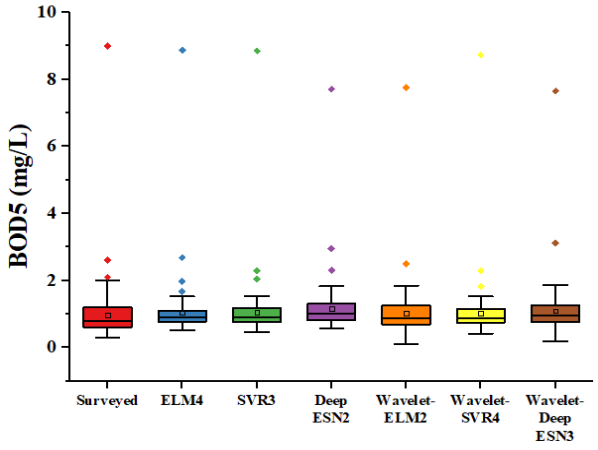


(e) Wavelet-SVR4 model

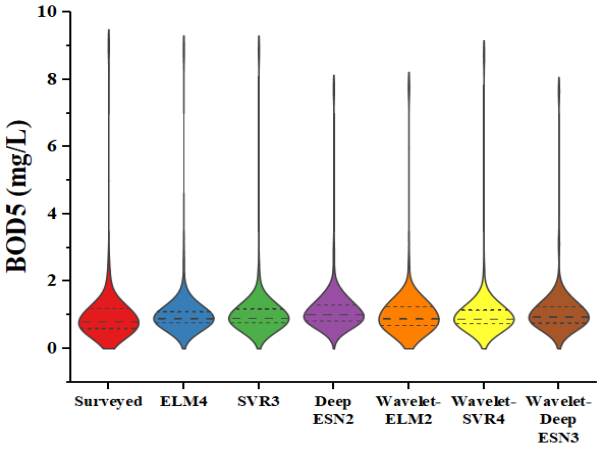


(f) Wavelet-Deep ESN3 model

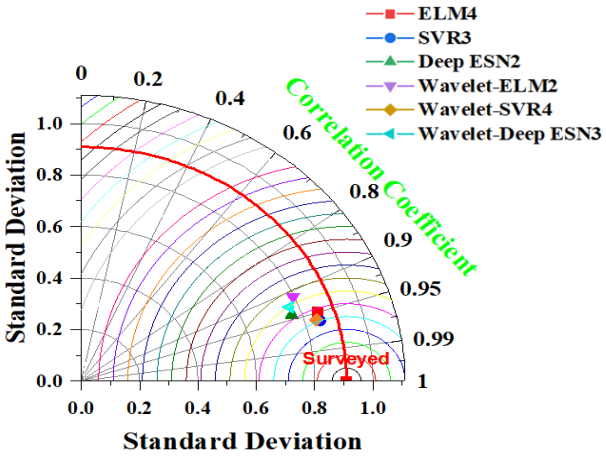
Figure 8. Scatterplots for the desirable standalone and double-stage synthesis models in the validation phase (Hwangji station).



(a) Boxplots



(b) Violin plots



(c) Taylor diagram

Figure 9. Boxplots, violin plots, and Taylor diagram for the desirable standalone and double-stage synthesis models in the validation phase (Hwangji station).

5.2. Predicting BOD_5 concentration at Toilchun station

5.2.1. Utilization of standalone models

The outputs of three mathematical formulae (R^2, NS, and RMSE) for the standalone models are summed up in Table 5 for Toilchun station. Table 5 shows that the estimates of ELM1 (R^2 = 0.571, NS = 0.472, and RMSE = 0.472 mg/L) are preferable to the SVM1 and Deep ESN1 in the validation phase considering the 1st distribution. During the 2nd distribution, the SVR2 model (R^2 = 0.722, NS = 0.691, and RMSE = 0.361 mg/L) are more remarkable compared to the ELM2 and Deep ESN2. Also, the performance of SVR3 (R^2 = 0.701, NS = 0.661, and RMSE = 0.378 mg/L) exceeds the ELM3 and Deep ESN3 obviously regarding the 3rd distribution in the validation phase. The contradiction of standalone models subjected to the 4th distribution, besides, approves that the SVR4 (R^2 = 0.868, NS = 0.854, and RMSE = 0.248 mg/L) outperforms the ELM4 and Deep ESN4 in the validation phase. Eventually, the SVR5 (R^2 = 0.876, NS = 0.870, and RMSE = 0.234 mg/L) is more reliable than ELM5 and Deep ESN5 in the validation phase with the 5th distribution.

Table 5. Results of three mathematical criteria using the standalone models at Toilchun station.

Classification	Distribution	Validation phase		
		R^2	NS	RMSE (mg/L)
Standalone	ELM1	0.571	0.472	0.472
	ELM2	0.671	0.630	0.395
	ELM3	0.677	0.641	0.389
	ELM4	0.739	0.738	0.332
	ELM5	0.808	0.807	0.285
	SVR1	0.534	0.477	0.469
	SVR2	0.722	0.691	0.361
	SVR3	0.701	0.661	0.378
	SVR4	0.868	0.854	0.248
	SVR5	0.876	0.870	0.234
	Deep ESN1	0.376	0.359	0.520
	Deep ESN2	0.637	0.547	0.437
	Deep ESN3	0.491	0.417	0.496
	Deep ESN4	0.491	0.461	0.476
	Deep ESN5	0.608	0.606	0.408

Contemplating the magnificent models among the 1st–5th distributions, the best capability of standalone models can be discovered from the ELM (the 5th distribution), SVR (the 5th distribution), and Deep ESN (the 5th distribution) among diverse input associations in the validation phase. As seen from Table 5, the improved performance of SVR5 contributes better prediction compared to the ELM5 and Deep ESN5 in the validation phase. As a result, the SVR5 is most reliable for predicting BOD_5 concentration among the improved standalone models at Toilchun station.

5.2.2. Utilization of double-stage synthesis models

The outputs of three mathematical criteria for the double-stage synthesis models are still organized as in Table 6 at Toilchun station. As observed from Table 6, the estimates of Wavelet-SVR1 (R^2 = 0.662, NS = 0.647, and RMSE = 0.386 mg/L) are more prevalent than the Wavelet-ELM1 and

Wavelet-Deep ESN1 in the validation phase utilizing the 1st distribution. Favoring the 2nd distribution, the Wavelet-SVR2 (R^2 = 0.866, NS = 0.845, and RMSE = 0.255 mg/L) is more exquisite than the Wavelet-ELM2 and Wavelet-Deep ESN2. Likewise, the performance of Wavelet-SVR3 (R^2 = 0.688, NS = 0.646, and RMSE = 0.386 mg/L) exceeds the Wavelet-ELM3 and Wavelet-Deep ESN3 viewing the 3rd distribution during validation phase. Likewise, the contradiction of double-stage synthesis models in the 4th distribution demonstrates that the Wavelet-SVR4 (R^2 =0.922, NS = 0.917, and RMSE = 0.187 mg/L) surpasses the Wavelet-ELM4 and Wavelet-Deep ESN4 definitely in the validation phase. Eventually, the Wavelet-SVR5 (R^2 = 0.780, NS = 0.775, and RMSE = 0.308 mg/L) is more effective than the Wavelet-ELM5 and Wavelet-Deep ESN5 in the validation phase with the 5th distribution.

Envisaging the eminent models from the 1st–5th distributions, the attractive performance of double-stage synthesis models can be evaluated from the Wavelet-ELM (the 5th distribution), Wavelet-SVR (the 4th distribution), and Wavelet-Deep ESN (the 5th distribution) among various input associations in the validation phase. It can be seen from Table 6 that the Wavelet-SVR4 yields more reliable outcomes compared to the Wavelet-ELM5 and Wavelet-Deep ESN5 in the validation phase. As a result, the Wavelet-SVR4 performs superior to the Wavelet-ELM5 and Wavelet-Deep ESN5 for predicting BOD_5 concentration among the enhanced double-stage synthesis models at Toilchun station.

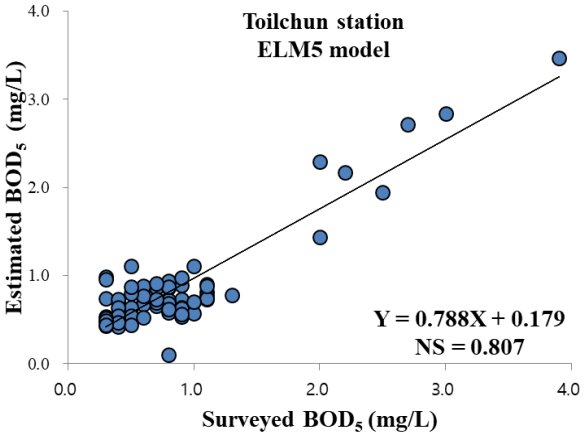
Table 6. Results of three mathematical criteria using the double-stage synthesis models at Toilchun station.

Classification	Distribution	Validation phase		
		R^2	NS	RMSE (mg/L)
Double-stage synthesis	Wavelet-ELM1	0.608	0.565	0.428
	Wavelet-ELM2	0.539	0.530	0.445
	Wavelet-ELM3	0.665	0.627	0.397
	Wavelet-ELM4	0.541	0.522	0.449
	Wavelet-ELM5	0.698	0.689	0.362
	Wavelet-SVR1	0.662	0.647	0.386
	Wavelet-SVR2	0.866	0.845	0.255
	Wavelet-SVR3	0.688	0.646	0.386
	Wavelet-SVR4	0.922	0.917	0.187
	Wavelet-SVR5	0.780	0.775	0.308
	Wavelet-Deep ESN1	0.473	0.471	0.472
	Wavelet-Deep ESN2	0.579	0.550	0.435
	Wavelet-Deep ESN3	0.596	0.570	0.426
	Wavelet-Deep ESN4	0.498	0.477	0.470
	Wavelet-Deep ESN5	0.663	0.660	0.379

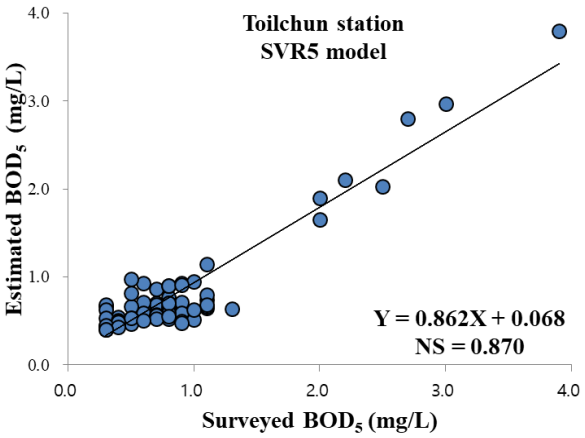
5.2.3. Visual aids of model accomplishment

To validate the precision of desirable standalone and double-stage synthesis models using visual aids, Figures 10(a)–(f) provide the scatterplots for the surveyed and estimated BOD_5 concentration values employing the desirable standalone and double-stage synthesis models at Toilchun station. The linear formulae and values of NS efficiency criterion are inserted for the corresponding standalone and

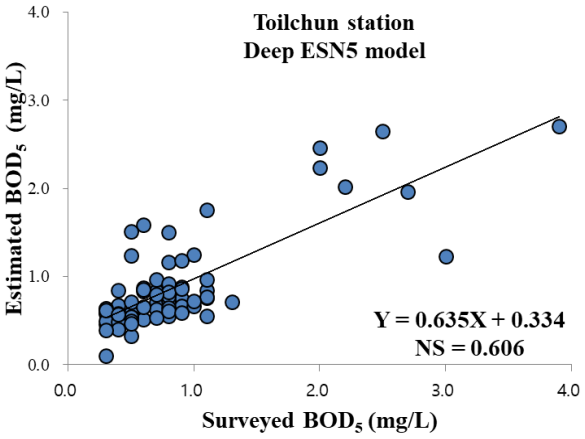
double-stage synthesis models, respectively. It can be inferred from NS values and the slopes of linear formulae that a definite inconsistency can be traced among the desirable standalone and double-stage synthesis models (ELM5, SVR5, Deep ESN5, Wavelet-ELM5, Wavelet-SVR4, and Wavelet-Deep ESN5). Therefore, the Wavelet-SVR4 accomplishes the most reliable precision for predicting BOD_5 concentration values obviously, while the Deep ESN5 yields the least precise among the desirable models at Toilchun station.



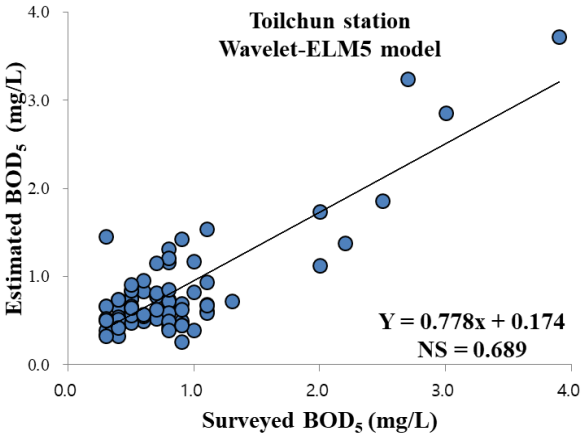
(a) ELM5 model



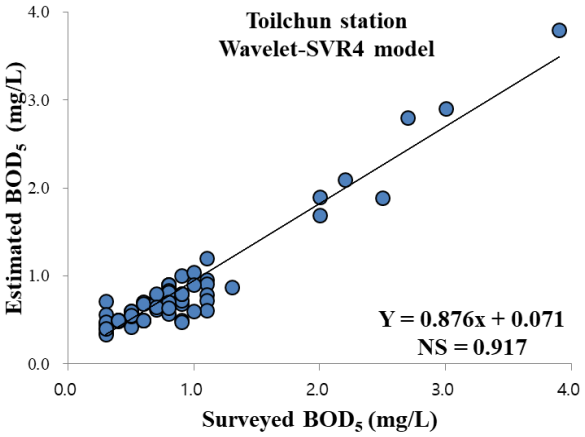
(b) SVR5 model



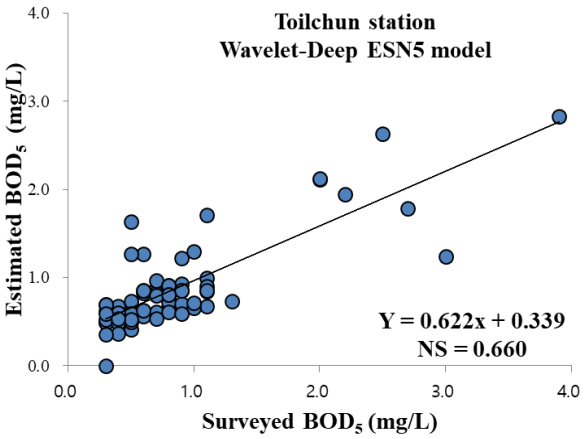
(c) Deep ESN5 model



(d) Wavelet-ELM5 model



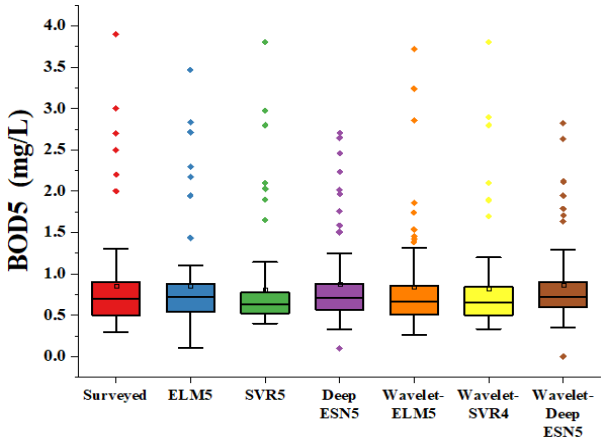
(e) Wavelet-SVR4 model



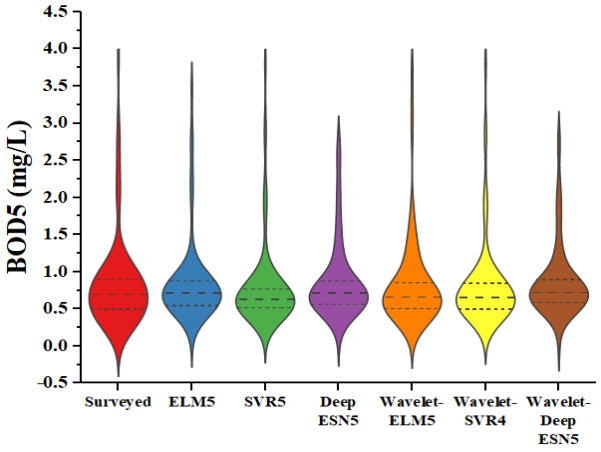
(f) Wavelet-Deep ESN5 model

Figure 10. Scatterplots for the desirable standalone and double-stage synthesis models in the validation phase (Toilchun station).

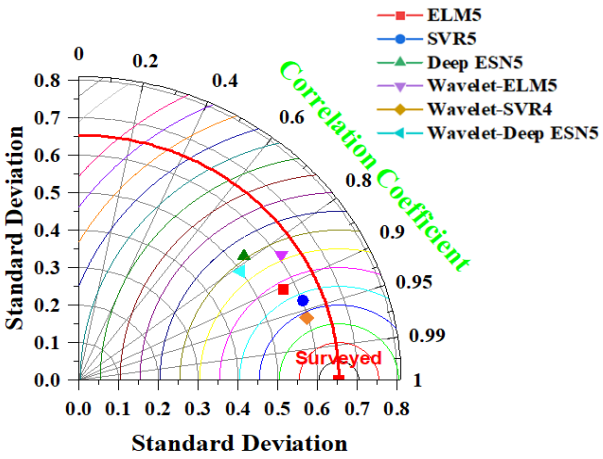
Additional pictures can anticipate the performance of standalone and double-stage synthesis models using the boxplot, violin plot, and Taylor diagram. Figures 11(a)–(c) support the various visual aids for the desirable standalone and double-stage synthesis models at Toilchun station. It can be seen from Figure 11(a) that the boxplots of estimated BOD_5 concentration utilizing the Wavelet-SVR4 can resemble that of surveyed BOD_5 concentration intimately. Another visual aid for the distribution of surveyed and estimated BOD_5 concentration values employing the desirable models can be displayed with the violin plots (Figure 11(b)). The violin plot can be described as one of schematic techniques to reveal the distribution of mandated numerical values. Figure 11(b) demonstrates a similar contour pattern for the Wavelet-SVR4 employing the median, interquartile, and distribution of mandated values. Figure 11(c) supplies the Taylor diagram employing the desirable standalone and double-stage synthesis models for Toilchun station. It can be seen from Figure 11(c) that the point of Wavelet-SVR4 which has the minimal RMSE value visualizes the straight distance from the surveyed one, while the point of Deep ESN5 displays the longest distance from the surveyed point.



(a) Boxplots



(b) Violin plots



(c) Taylor diagram

Figure 11. Boxplots, violin plots, and Taylor diagram for the optimal standalone and double-stage synthesis models during validation phase (Toilchun station).

6. Discussion

The addressed research explored the nonlinear behavior (e.g., hard to predict) of BOD_5 concentration employing standalone and double-stage synthesis models in Hwangji and Toilchun stations, South Korea. Since both stations (i.e., Hwangji and Toilchun) yielded the different high-quality accuracies for the desirable standalone models, it was hard to judge which model predicted BOD_5 concentration with accuracy. Also, the outputs of three mathematical formulae explained that the SVM models with diverse input associations could predict BOD_5 concentration precisely compared to the ELM and Deep ESN models based on the corresponding distribution on both stations. Because all standalone models enforced the various theoretical structures and inference, the accurate prediction was changed for diverse input associations of standalone models.

The main aim for developing the double-stage synthesis models was to enhance the accurate prediction of BOD_5 concentration compared to corresponding standalone models. Unfortunately, the Wavelet-ELM models could not boost the accurate prediction for corresponding ELM models from the perspective of double-stage synthesis models' performance, based on NS values at Hwangji station. Among the Wavelet-SVR models, the Wavelet-SVR1 (0.4% for SVR1), Wavelet-SVR2 (0.7% for SVR2), Wavelet-SVR4 (3.7% for SVR4), and Wavelet-SVR5 (3.4% for SVR5) models slightly enhanced the accurate prediction. Also, all the Wavelet-Deep ESN models increased the predictive accuracy on a small scale, including the Wavelet-Deep ESN1 (2.5% for Deep ESN1), Wavelet-Deep ESN2 (0.1% for Deep ESN2), Wavelet-Deep ESN3 (3.5% for Deep ESN3), Wavelet-Deep ESN4 (6.0% for Deep ESN4), and Wavelet-Deep ESN5 (1.0% for Deep ESN5). Noticing the desirable models' categorization for the standalone and double-stage synthesis models, the Wavelet-SVR4 model, which yielded the best accuracy, improved the accurate prediction by 12.7% (Wavelet-ELM2), 9.8% (Wavelet-Deep ESN3), 2.5% (ELM4), and 10.1% (Deep ESN2), respectively.

Regarding the performance evaluation of double-stage synthesis models by NS values at Toilchun station, only the Wavelet-ELM1 (19.7% for ELM1) model could boost the estimated efficiency obviously among the Wavelet-ELM models. Also, Wavelet-SVR1 (35.6% for SVR1), Wavelet-SVR2 (22.3% for SVR2), and Wavelet-SVR4 (7.4% for SVR4) models increased the accurate prediction

clearly among the Wavelet-SVR models. In addition, all the Wavelet-Deep ESN models, including the Wavelet-Deep ESN1 (31.2% for Deep ESN1), Wavelet-Deep ESN2 (0.5% for Deep ESN2), Wavelet-Deep ESN3 (36.7% for Deep ESN3), Wavelet-Deep ESN4 (3.5% for Deep ESN4), and Wavelet-Deep ESN5 (8.9% for Deep ESN5) models, boosted the precise efficiency, respectively. Considering the desirable models' categorization for the standalone and double-stage synthesis models, the Wavelet-SVR4 model which produced the best accuracy, reinforced the accurate prediction by 32.8% (Wavelet-ELM5), 38.6% (Wavelet-Deep ESN5), 13.4% (ELM5), 5.2% (SVR5), and 51.0% (Deep ESN5), respectively. The double-stage synthesis models, therefore, could not always reinforce the accurate prediction of corresponding standalone models on both stations. This experience pursued the previous works of [2] and [44]. [44] predicted DO concentration employing the single and hybrid machine learning models in Florida, USA. Results demonstrated that the hybrid machine learning models could not regularly improve the predicted accuracy of single machine learning models. Also, [2] implemented the single and combinational paradigm to predict BOD_5 concentration in South Korea. They found that the combinational paradigm could not always increase the predictive accuracy of single models clearly.

Therefore, the process which embeds the different data preprocessing algorithms [45–49] in the diverse standalone (i.e., machine learning and deep learning) models, is required to increase the accurate prediction and efficiency of BOD_5 concentration for the continuous research.

7. Conclusions

The addressed research explored the precision and efficiency of the standalone and double-stage synthesis models for predicting BOD_5 concentration in Hwangji and Toilchun stations, South Korea. Five input associations (1^{st}–5^{th} distributions) were resolved for developing the standalone and double-stage synthesis models based on seven water quantity and quality items. For the modeling and prediction of standalone and double-stage synthesis models, the assembled data were divided into training and validation samples, respectively. Three mathematical formulae (R^2, NS, and RMSE) and four graphical aids (scatter diagram, boxplot, violin plot, and Taylor diagram) were used to evaluate the accurate prediction of addressed models.

Considering the best models from the 1^{st}-5^{th} distributions, the SVR3 (R^2 = 0.925, NS = 0.915, and RMSE = 0.264 mg/L) and Wavelet-SVR4 (R^2 = 0.926, NS = 0.915, and RMSE = 0.264 mg/L) models were the most precise compared to other desirable models (ELM4, Deep ESN2, Wavelet-ELM2, and Wavelet-Deep ESN3) based on standalone and double-stage synthesis models in the validation phase at Hwangji station. Also, the Wavelet-SVR4 model (R^2 = 0.922, NS = 0.917, and RMSE = 0.162 mg/L) provided more precise results than other desirable models (ELM5, SVR5, Deep ESN5, Wavelet-ELM5, and Wavelet-Deep ESN5) for predicting BOD_5 concentration in the validation phase at Toilchun station. However, it was found the addressed research that explained that the precision and efficiency of BOD_5 concentration estimated by the standalone models could not be reinforced by the double-stage synthesis models on both stations. Therefore, using the credible water quantity and quality items from the available data groups can confirm the outputs of the addressed research, and perform the best prediction of BOD_5 concentration employing the different standalone and double-stage synthesis models in river.

Conflict of interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. S. Kim, M. Alizamir, M. Zounemat-Kermani, O. Kisi, V. P. Singh, Assessing the biochemical oxygen demand using neural networks and ensemble tree approaches in South Korea, *J. Environ. Manage.*, **270** (2020), 110834. https://doi.org/10.1016/j.jenvman.2020.110834
2. S. Kim, Y. Seo, M. Zakhrouf, A. Malik, Novel two-stage hybrid paradigm combining data pre-processing approaches to predict biochemical oxygen demand concentration, *J. Korea Water Resour. Assoc.*, **54** (2021), 1037–1051. https://doi.org/10.3741/JKWRA.2021.54.S-1.1037
3. M. Najafzadeh, A. Ghaemi, Prediction of the five-day biochemical oxygen demand and chemical oxygen demand in natural streams using machine learning methods. *Environ. Monit. Assess.*, **191** (2019), 1–21. https://doi.org/10.1007/s10661-019-7446-8
4. S. Jouanneau, L. Recoules, M. J. Durand, A. Boukabache, V. Picot, Y. Primault, et al., Methods for assessing biochemical oxygen demand (BOD): A review. *Water Res.*, **49** (2014), 62–82. https://doi.org/10.1016/j.watres.2013.10.066
5. S. B. H. S. Asadollah, A. Sharafati, D. Motta, Z. M. Yaseen, River water quality index prediction and uncertainty analysis: A comparative study of machine learning models, *J. Environ. Chem. Eng.*, **9** (2021), 104599. https://doi.org/10.1016/j.jece.2020.104599
6. Royal Commission on Sewage Disposal, *Fifth report on methods of treating and disposing of sewage*, United Kingdom, 1908.
7. M. Ay, O. Kisi, Modeling of dissolved oxygen concentration using different neural network techniques in Foundation Creek, El Paso County, Colorado, *J. Environ. Eng.*, **138** (2012), 654–662. https://doi.org/10.1061/(ASCE)EE.1943-7870.0000511
8. B. Chanda, R. Blunck, L. C. Faria, F. E. Schweizer, I. Mody, F. Bezanilla, A hybrid approach to measuring electrical activity in genetically specified neurons, *Nat. Neurosci*., **8** (2005), 1619–1626. https://doi.org/10.1038/nn1558
9. J. Li, H. A. Abdulmohsin, S. S. Hasan, L. Kaiming, B. Al-Khateeb, M. I. Ghareb, et al., Hybrid soft computing approach for determining water quality indicator: Euphrates River, *Neural. Comput. Appl.*, **31** (2019), 827–837. https://doi.org/10.1007/s00521-017-3112-7
10. D.T. Bui, K. Khosravi, J. Tiefenbacher, H. Nguyen, N. Kazakis, Improving prediction of water quality indices using novel hybrid machine-learning algorithms, *Sci. Total Environ*., **721** (2020), 137612. https://doi.org/10.1016/j.scitotenv.2020.137612
11. K. Chen, H. Chen, C. Zhou, Y. Huang, X. Qi, R. Shen, et al., 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.*, **171** (2020), 115454. https://doi.org/10.1016/j.watres.2019.115454
12. V. Sagan, K.T. Peterson, M. Maimaitijiang, P. Sidike, J. Sloan, B. A. Greeling, et al., Monitoring inland water quality using remote sensing: Potential and limitations of spectral indices, bio-optical

simulations, machine learning, and cloud computing, *Earth-Sci. Rev.*, **205** (2020), 103187. https://doi.org/10.1016/j.earscirev.2020.103187

13. M. Alizamir, S. Heddam, S. Kim, A. D. Mehr, On the implementation of a novel data-intelligence model based on extreme learning machine optimized by bat algorithm for estimating daily chlorophyll-a concentration: Case studies of river and lake in USA, *J. Clean. Prod*., **285** (2021), 124868. https://doi.org/10.1016/j.jclepro.2020.124868
14. Y. Jiang, C. Li, L. Sun, D. Guo, Y. Zhang, W. Wang, A deep learning algorithm for multi-source data fusion to predict water quality of urban sewer networks, *J. Clean. Prod.*, **318** (2021), 128533. https://doi.org/10.1016/j.jclepro.2021.128533
15. A. A. M. Ahmed, S. M. A. Shah, Application of adaptive neuro-fuzzy inference system (ANFIS) to estimate the biochemical oxygen demand (BOD) of Surma River, *J. King Saud Univ. Eng. Sci.*, **29** (2017), 237–243. https://doi.org/10.1016/j.jksues.2015.02.001
16. H. Tao, A. M. Bobaker, M. M. Ramal, Z. M. Yaseen, M. S. Hossain, S. Shahid, Determination of biochemical oxygen demand and dissolved oxygen for semi-arid river environment: Application of soft computing models. *Environ. Sci. Pollut. Res.*, **26** (2019), 923–937. https://doi.org /10.1007/s11356-018-3663-x
17. J. Ma, Y. Ding, J. C. Cheng, F. Jiang, Z. Xu, Soft detection of 5-day BOD with sparse matrix in city harbor water using deep learning techniques, *Water Res.*, **170** (2020), 115350. https://doi.org/10.1016/j.watres.2019.115350
18. B. S. Pattnaik, A. S. Pattanayak, S. K. Udgata, A. K. Panda, Machine learning based soft sensor model for BOD estimation using intelligence at edge, *Complex Intell. Syst*., **7** (2021), 961–976. https://doi.org/10.1007/s40747-020-00259-9
19. F. Granata, S. Papirio, G. Esposito, R. Gargano, G. De Marinis, Machine learning algorithms for the forecasting of wastewater quality indicators., *Water*, **9** (2017), 105. https://doi.org/10.3390/w9020105
20. A. Solgi, A. Pourhaghi, R. Bahmani, H. Zarei, Improving SVR and ANFIS performance using wavelet transform and PCA algorithm for modeling and predicting biochemical oxygen demand (BOD), *Ecohydrol. Hydrobiol.*, **17** (2017), 164–175. https://doi.org/10.1016/j.ecohyd.2017.02.002
21. S. Khullar, N. Singh, Water quality assessment of a river using deep learning Bi-LSTM methodology: forecasting and validation, *Environ. Sci. Pollut. Res.*, **29** (2022), 12875–12889. https:// doi.org/10.1007/s11356-021-13875-w
22. N. Nafsin, J. Li, Prediction of 5-day biochemical oxygen demand in the Buriganga River of Bangladesh using novel hybrid machine learning algorithms, *Water Environ. Res.*, **94** (2022), e10718. https://doi.org/10.1002/wer.10718
23. G. B. Huang, Q. Y. Zhu, C. K. Siew, Extreme learning machine: theory and applications, *Neurocomputing*, **70** (2006), 489–501. https://doi.org/10.1016/j.neucom.2005.12.126
24. L. F. Arias-Rodriguez, Z. Duan, J. D. J. Díaz-Torres, M. B. Hazas, J. Huang, B. U. Kumar, et al., Integration of remote sensing and Mexican water quality monitoring system using an extreme learning machine, *Sensors*, **21** (2021), 4118. https://doi.org/10.3390/s21124118
25. S. Tripathi, V. V. Srinivas, R. S. Nanjundish, Downscaling of precipitation for climate change scenarios: A support vector machine approach, *J. Hydrol.*, **330** (2006), 621–640. https://doi.org/10.1016/j.jhydrol.2006.04.030

26. V. N. Vapnik, *The nature of statistical learning theory*, 2nd Edition, Springer-Verlag, New York, 2010.
27. S. Haykin, *Neural networks and learning machines*, 3rd Edition, Prentice Hall, New Jersey, 2009.
28. S. Kim, J. Shiri, O. Kisi, Pan evaporation modeling using neural computing approach for different climatic zones, *Water Resour. Manag*., **26** (2012), 3231–3249. https://doi.org/10.1007/s11269-012-0069-2
29. S. Kim, Y. Seo, V. P. Singh, Assessment of pan evaporation modeling using bootstrap resampling and soft computing methods, *J. Comput. Civ. Eng.*, **29** (2015), 04014063. https://doi.org/10.1061/(ASCE)CP.1943-5487.0000367
30. X. Sun, T. Li, Q. Li, Y. Huang, Y. Li, Deep belief echo-state network and its application to time series prediction, *Knowl. Based Syst.*, **130** (2017), 17–29. https://doi.org/10.1016/j.knosys.2017.05.022
31. M. Alizamir, S. Kim, O. Kisi, M. Zounemat-Kermani, Deep echo state network: a novel machine learning approach to model dew point temperature using meteorological variables, *Hydrol. Sci. J.*, **65** (2020), 1173–1190. https://doi.org/10.1080/02626667.2020.1735639
32. M. H. Yen, D. W. Liu, Y. C. Hsin, C. E. Lin, C. C. Chen, Application of the deep learning for the prediction of rainfall in Southern Taiwan, *Sci. Rep*., **9** (2019), 1–9. https://doi.org/10.1038/s41598-019-49242-6
33. Y. C. Bo, P. Wang, X. Zhang, B. Liu, Modeling data-driven sensor with a novel deep echo state network. *Chemometr. Intell. Lab. Syst*., **206** (2020), 104062. https://doi.org/10.1016/j.chemolab.2020.104062
34. S. G. Mallat, A theory of multiresolution signal decomposition: the wavelet representation, *IEEE Trans. Pattern Anal. Mach. Intell*., **11** (1989), 674–693. https://doi.org/10.1109/34.192463
35. S. Kim, O. Kisi, Y. Seo, V. P. Singh, C. J. Lee, Assessment of rainfall aggregation and disaggregation using data-driven models and wavelet decomposition, *Hydrol. Res*., **48** (2017), 99–116. https://doi.org/10.2166/nh.2016.314
36. M. J. Shensa, The discrete wavelet transform: wedding the a trous and Mallat algorithms, *IEEE Trans. Signal Process*., **40** (1992), 2464–2482. https://doi.org/10.1109/78.157290
37. N. J. Nagelkerke, A note on a general definition of the coefficient of determination, *Biometrika*, **78** (1991), 691–692. https://doi.org/10.1093/biomet/78.3.691
38. P. Krause, D. P. Boyle, F. Bäse, Comparison of different efficiency criteria for hydrological model assessment. *Adv. Geosci*., **5** (2005), 89–97. https://doi.org/10.5194/adgeo-5-89-2005
39. J. E. Nash, J. V. Sutcliffe, River flow forecasting through conceptual models, Part 1 – A discussion of principles, *J. Hydrol.*, **10** (1970), 282–290. https://doi.org/10.1016/0022-1694(70)90255-6
40. J. S. Armstrong, F. Collopy, Error measures for generalizing about forecasting methods: Empirical comparisons, *Int. J. Forecast.*, **8** (1992), 69–80. https://doi.org/10.1016/0169-2070(92)90008-W
41. T. A. Clark, P. H. Dare, M. E. Bruce, Nitrogen fixation in an aerated stabilization basin treating bleached kraft mill wastewater, *Water Environ. Res.*, **69** (1997), 1039–1046. https://doi.org/10.2175/106143097X125740
42. J. L. Hintze, R. D. Nelson, Violin plots: A box plot-density trace synergism, *Am. Stat.*, **52** (1998), 181–184. https://doi.org/10.1080/00031305.1998.10480559
43. K. E. Taylor, Summarizing multiple aspects of model performance in a single diagram, *J. Geophys. Res. Atmos.*, **106** (2001), 7183–7192. https://doi.org/10.1029/2000JD900719

44. M. Zounemat-Kermani, Y. Seo, S. Kim, M. A. Ghorbani, S. Samadianfard, S. Naghshara, et al., Can decomposition approaches always enhance soft computing models? Predicting the dissolved oxygen concentration in the St. Johns River, Florida, *Appl. Sci*., **9** (2019), 2534. https://doi.org/10.3390/app9122534
45. M. Huang, D. Tian, H. Liu, C. Zhang, X. Yi, J. Cai, et al., A hybrid fuzzy wavelet neural network model with self-adapted fuzzy-means clustering and genetic algorithm for water quality prediction in rivers. *Complexity*, **2018** (2018), 8241342. https://doi.org/10.1155/2018/8241342
46. M. Montaseri, S. Z. Z. Ghavidel, H. Sanikhani, Water quality variations in different climates of Iran: toward modeling total dissolved solid using soft computing techniques, *Stoch. Environ. Res. Risk Assess.*, **32** (2018), 2253–2273. https://doi.org/10.1007/s00477-018-1554-9
47. Y. Zhou, Real-time probabilistic forecasting of river water quality under data missing situation: Deep learning plus post-processing techniques, *J. Hydrol*., **589** (2020), 125164. https://doi.org/10.1016/j.jhydrol.2020.125164
48. J. Sha, X. Li, M. Zhang, Z. L. Wang, Comparison of forecasting models for real-time monitoring of water quality parameters based on hybrid deep learning neural networks, *Water*, **13** (2021), 1547. https://doi.org/10.3390/w13111547
49. S. Vijay, K. Kamaraj, Prediction of water quality index in drinking water distribution system using activation functions based ANN, *Water Resour. Manag*., **35** (2021), 535–553. https://doi.org/10.1007/s11269-020-02729-8

AIMS Press