







## Article

# Prediction of Dissolved Oxygen Factor at Oncheon Stream Watershed Using Long Short-Term Memory Algorithm

Heesung Lim <sup>1</sup>, Hyungjin Shin <sup>1</sup>, Jaenam Lee <sup>1</sup>, Jongwon Do <sup>1</sup>, Inhyeok Song <sup>2</sup> and Youngkyu Jin <sup>1,\*</sup>

<sup>1</sup> Rural Research Institute, Korea Rural Community Corporation, Ansan 15634, Republic of Korea; hslim1@ekr.or.kr (H.L.); shjin@ekr.or.kr (H.S.); jnlee@ekr.or.kr (J.L.); jonduru@ekr.or.kr (J.D.)

<sup>2</sup> Department of Agricultural Engineering, Chungnam National University, Daejeon 34134, Republic of Korea; sih4093@cnu.ac.kr

\* Correspondence: accvn75@ekr.or.kr

**Abstract:** Rapid urbanization and industrialization have caused water quality issues in urban rivers. Appropriate measures based on water quality monitoring systems and prediction methods are needed for water quality management. While South Korea has operated a water quality monitoring system that measures various environmental factors and has accumulated water quality data, a water quality prediction system is not in place. This study suggests a water quality prediction method based on a long short-term model using water quality and meteorological monitoring data. Additionally, we present a derived input set of the prediction model that can improve the prediction model performance. The prediction model's performance was evaluated by the coefficient of determination under various conditions, such as the hyperparameters, temporal resolution of input data, and application of upstream and downstream data. As a result, using the temporal resolution of the input data as hourly data improved predictions by an average of 25.6% over three days of the prediction period compared to daily data. Meanwhile, it was analyzed that the hyperparameters and using upstream and downstream data have a minor effect on the model performance. The results of this study underscore the crucial role of the number, duration, and temporal resolution of available monitoring data in water quality management.

**Keywords:** water quality; LSTM; deep learning; monitoring data; urban stream



**Citation:** Lim, H.; Shin, H.; Lee, J.; Do, J.; Song, I.; Jin, Y. Prediction of Dissolved Oxygen Factor at Oncheon Stream Watershed Using Long Short-Term Memory Algorithm. *Water* **2024**, *16*, 2363. <https://doi.org/10.3390/w16172363>

Academic Editors: Roohollah Noori, Tianxiang Wang and Alex Neumann

Received: 1 July 2024

Revised: 20 August 2024

Accepted: 21 August 2024

Published: 23 August 2024



**Copyright:** © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

In South Korea, 60~70% of annual precipitation is concentrated during the wet season (25 June to 25 September), which is unfavorable in terms of water resource management. In addition, rapid industrialization and urbanization are causing environmental and water pollution due to the concentration of the population in large cities and reckless development. To improve polluted water quality, significant costs are necessary for purification, including facility and maintenance expenses. One method to minimize economic losses in water quality management involves establishing realistic standards and enhancing river self-purification activities. To achieve this goal, developing a data-based model by integrating current water quality monitoring data with results obtained from physical models and machine learning can facilitate the prediction and management of water pollution. To this end, it is possible to develop a data-based model by learning the results through a physical-based model or machine learning along with current water quality monitoring data to predict and manage water pollution [1].

Models for water quality prediction can be divided into conceptual and physical models, and methods using data-based machine learning and deep learning. Among these, conceptual models that can predict water quality include the QUAL2E model [2], WASP (water quality analysis simulation program) [3], and W-ARIMA-GRU [4]. HSPF (hydrological simulation program—Fortran) combines a watershed model and a water quality model, which can simulate the runoff process of nonpoint pollutants due to rainfall

with the interaction of hydraulic phenomena, sediments, and chemicals in rivers [5]. As a result of modeling the water quality impact in small watersheds using HSPF, HSPF is known to be effective in relatively small sub-watersheds, and both quantity and quality of water could be modeled [6]. Taheri Tizro et al. [7] performed predictions for nine water quality parameters using the ARIMA model, and, as a result, the prediction accuracy was relatively poor. Zhou et al. [4] performed a prediction of water quality index, dissolved oxygen (DO), and pH using W-ARIMA-GRU, which showed excellent generality and efficiency as a water quality index prediction model, but long-term prediction results were not good. Kim et al. [8] evaluated HSPF, which is widely used for ensemble data assimilation. As a result, the prediction technique was confirmed for various variables, such as observed biochemical oxygen demand (BOD) and chlorophyll a (Chl-a), except for dissolved oxygen (DO), and the improvement for DO was insufficient. However, DO is also used as an indicator of water quality and water pollution and plays a particularly important role in the aquatic environment [9,10]. Paliwal et al. [11] confirmed the effect of each model parameter on the DO and BOD prediction through QUAL2E and found that the consideration of the nitrogen cycle is necessary to improve the DO estimate.

Research is underway to provide useful water quality environment information by analyzing big data for each purpose and using real-time monitoring data to measure accurate water quality information according to recent technological changes [12]. The research using data showed good prediction effects by conducting river water level prediction studies that are relatively easier to secure data for than river water quality studies [13–15]. In most river water level prediction studies, the accuracy was very close to the actual water level when the previous time was close. Dam inflow, not river water level, was studied using a machine learning model [16,17], and the performance of the machine learning model showed excellent results even in the relative prediction of dam inflow. Kala and Vaidyanathan [18] and Mislán et al. [19] reported that high accuracy could be obtained by performing rainfall prediction using an artificial neural network for the accuracy of rainfall prediction. However, many studies have shown differences in prediction performance depending on the value of input data, and limitations in data collection. In research on water resources, research applying AI to fields that could not be handled with traditional technologies and thinking methods such as DO monitoring and water quality monitoring is taking place [20]. Much research has been conducted on water quality prediction using artificial neural networks [21–24], and in particular, many studies are being conducted on DO and BOD concentration prediction. Ref. [25] predicted and estimated DO concentration by applying artificial neural networks, and Ahmed [26] also applied artificial neural networks to DO and BOD prediction. Although many studies are being conducted on water quality prediction using artificial neural networks, water quality prediction also showed problems due to limitations in data collection.

As in the case above, it was confirmed that there are many studies on DO concentration prediction using artificial neural networks. However, research on DO concentration prediction using deep learning algorithms that deepen artificial neural networks is relatively lacking. Deep learning requires a large amount of data, and, in the past, there was a lot of cost and complexity to store and manage new data, so it was difficult to build data. However, with the development of data processing technology over the past 10 years, it has become much easier to store and generate data, and, thus, the amount of data accumulated is increasing. In this study, the correlation between the upstream, middle, and downstream points of the Oncheon stream watershed was analyzed using the data accumulated at the Busan Health and Environment Research Institute. For the analysis of upstream, middle, and downstream points, the LSTM (long short-term memory) algorithm, which is excellent for time series learning, was used for comparative analysis by performing day prediction using DO factor day data and time prediction using time data. To analyze the accuracy of prediction, the coefficient of determination  $R^2$  was calculated and evaluated. Data collected by time through the automatic measurement network were used by linear interpolation because missing data occurred due to calibration, maintenance, non-use, and equipment

power cut-off. Using the constructed time and day data, the LSTM algorithm predicted water quality after several hours and several days and comprehensively reviewed the analysis of upstream, middle, and downstream associations.

## 2. Materials and Methods

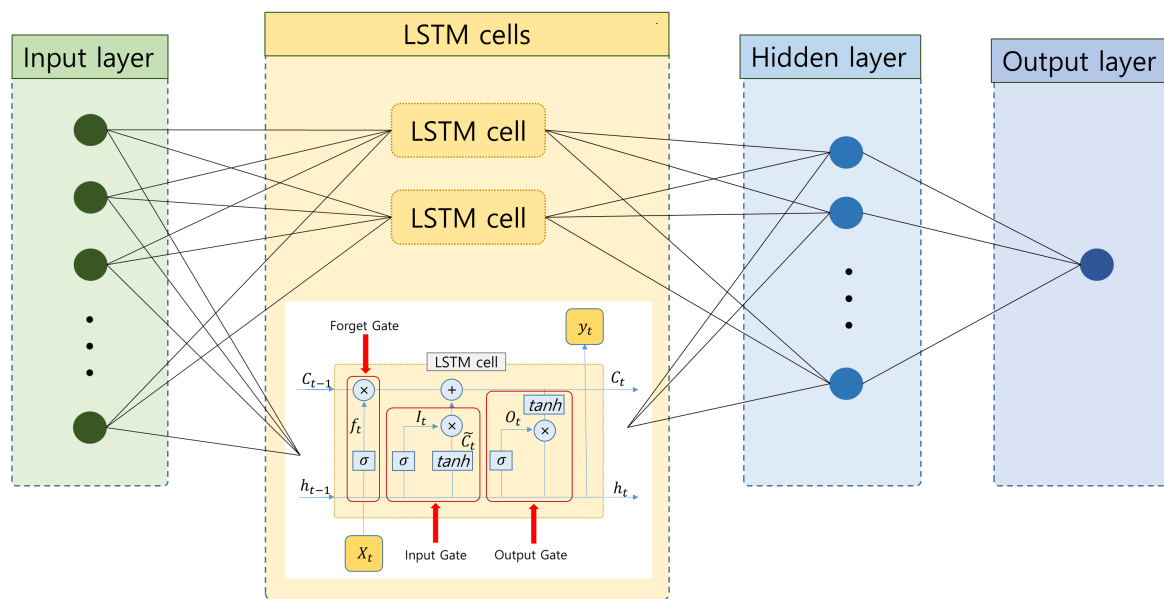
A water quality monitoring network using the water quality measurement system plays an important role in monitoring changes in water pollution and responding to pollution. However, the cost of installing and maintaining the water quality monitoring network constitutes a significant expense. To this end, it is necessary to prepare appropriate countermeasures based on water quality prediction. In particular, the most important thing in the prediction of dissolved oxygen (DO) concentration is the existing concentration data. Existing concentration data can be used for learning using artificial neural networks or used for comparison and verification with prediction results. For artificial neural network learning [15], the water level prediction was performed using the LSTM model, and the prediction result was very close to the actual water level. The previous paper determined that time series prediction simulation based on various factors would be possible due to the continuous improvement of deep learning-related algorithms and computing power. Using this, it was attempted to conduct a time series learning study based on the DO factor. Shim et al. [27] conducted a correlation analysis on the characteristics of monthly water quality changes between the upper and lower streams of tributaries, and, as a result, the correlations between chemical oxygen demand (COD), total organic carbon (TOC), and biochemical oxygen demand (BOD) were high. It was judged that a study on DO concentration prediction through the comparison of upper, middle, and downstream influences and time series learning was necessary, so daily or hourly prediction was performed.

### 2.1. Recurrent Neural Networks

The recurrent neural network (RNN) algorithm applied in this study has a loop-repeating structure in which past data effectively affect future processing for time series data processing. RNNs are often used to process continuous data, remember past information, and use this information to predict the current output values. The hidden layer of the RNN includes not only the current input layer but also the output values of the past hidden layer [28]. An RNN has a very useful advantage in that it utilizes previous information in the current state. However, an RNN exhibits a long-term dependency problem that reflects information close to the present time, but not information from the distant past. Because of this phenomenon, Hochreiter and Schmidhuber [29] proposed the LSTM algorithm to solve the problem of long-term dependence of the recurrent neural network.

### 2.2. LSTM Algorithm Selection

The long short-term memory (LSTM) is a type of recurrent neural network (RNN) capable of long-term dependency learning, and the core of the LSTM is the cell state. Figure 1 shows the structure of the LSTM; the concept of a cell state ( $C_t$ ) is introduced to update the state ( $h_t$ ) at a specific point in time. The LSTM controls the information to be included in the cell state through a total of three gates: the input gate ( $i_t$ ), forget gate ( $f_t$ ), and output gate ( $o_t$ ) inside the cell state, and transfers it to the next state. Next, we used sigmoid ( $\sigma$ ) as a forget gate ( $f_t$ ), which determines whether to discard or use information, to check whether specific information is removed. Equations for each state are presented in Equations (1)–(6) [29,30]. Equation (1) defines the  $f_t$  expression.



**Figure 1.** Structure of recurrent neural network long short-term memory.  $C_t$ , cell state;  $f_t$ , forget gate;  $h_t$ , hidden state;  $i_t$ , input gate;  $O_t$ , output gate.

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \tag{1}$$

where  $W_f$  is the forget gate weight,  $b_f$  is the forget gate bias value, and  $\sigma$  is the sigmoid activation function.

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{2}$$

$$\bar{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \tag{3}$$

where  $W_i$  is the input gate weight and  $b_i$  is the input gate bias value. In this step of updating the previous cell state, new information determined through  $i_t$  is added to the value deleted or used by  $f_t$ . Equation (3) defines the update of the cell state.

$$C_t = f_t \cdot C_{t-1} + i_t \cdot \bar{C}_t \tag{4}$$

Finally, the output gate ( $o_t$ ) can be written as Equation (5), and the state ( $h_t$ ) at a specific point in time can be written as Equation (6).

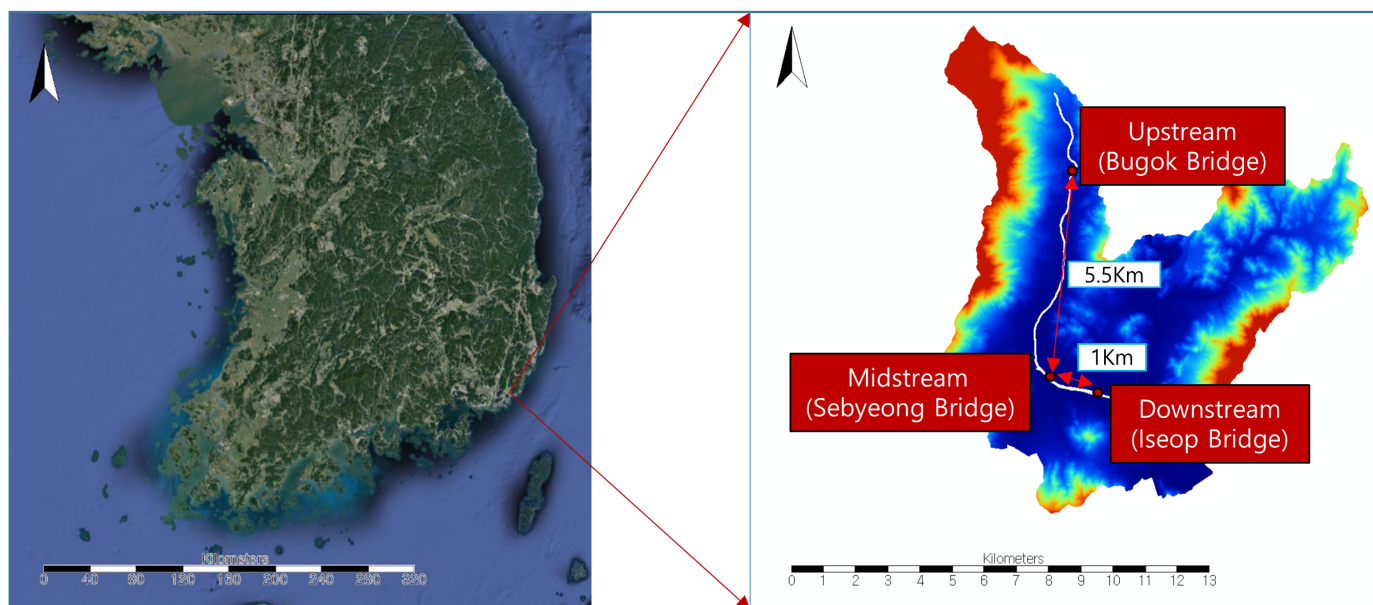
$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \tag{5}$$

$$h_t = o_t \cdot \tanh(C_t) \tag{6}$$

### 2.3. Study Area

The prediction of water pollution factors for the future can establish river water quality management measures and prepare for future water pollution. To predict future water pollution factors, it is necessary to first have the existing data necessary for the study to make predictions using the data [31]. In this study, water quality data of Bugok Bridge, Sebyeong Bridge, and Iseop Bridge in the Oncheon stream watershed in Busan, South Korea, where automatic water quality monitoring networks are installed, were used. The Oncheon stream is located in Busan and, as the first tributary of the Suyeong river, the basin area occupies about 28% of the total area of the Suyeong river of 56.28 km<sup>2</sup> [32]. Data from the automatic water quality monitoring network in the Oncheon stream were provided by the Busan Research Institute of Public Health and Environment (RIPHE) (<https://www.busan.go.kr/ihe/index>, accessed on 5 May 2024). Bugok Bridge, Sebyeong

Bridge, and Iseop Bridge Observatories are located at Bugok Bridge ( $35^{\circ}14'32''$ ,  $129^{\circ}05'23''$ ), Sebyeong Bridge ( $35^{\circ}11'48''$ ,  $129^{\circ}04'57''$ ), and ISeop Bridge ( $35^{\circ}11'35''$ ,  $129^{\circ}05'33''$ ). The location of each point is as shown in Figure 2, and the difference in straight line distance is 5.5 km between Bugok Bridge and Sebyeong Bridge, and 1 km between Sebyeong Bridge and Iseop Bridge.



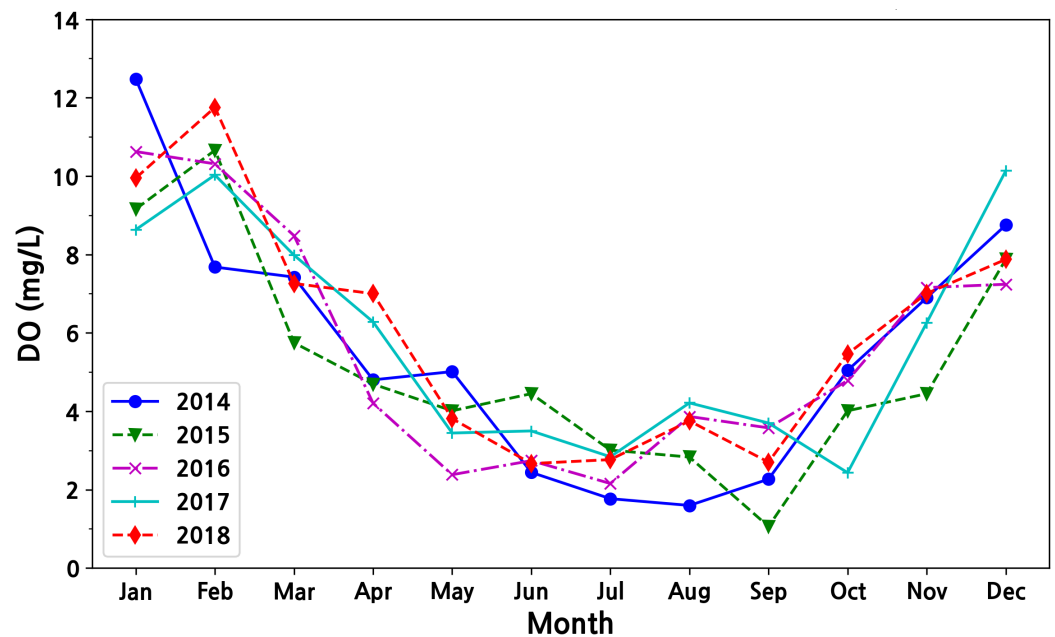
**Figure 2.** The study area of Oncheon stream watershed (Busan, South Korea).

#### 2.4. Construction of DO Data and Meteorological Data

In this study, the dissolved oxygen (DO) concentration prediction was performed in the Oncheon Stream watershed in Busan by using the long short-term memory (LSTM) algorithm, which is a type of deep learning. The DO concentration data from 1 January 2014~31 December 2018 at the Bugok Bridge, Sebyeong Bridge, and Iseop Bridge branches were collected through the Busan Research Institute of Public Health and Environment (RIPHE) System and used for learning. The monthly graph of DO concentration from 2016 to 2018 at the point of the Iseop Bridge located downstream of the Oncheon watershed is shown in Figure 3.

In the hourly data received from the Busan Research Institute of Public Health and Environment (RIPHE) System, 1732 data points at Bugok Bridge, 1814 data points at Sebyeong Bridge, and 1379 data points at ISeop Bridge were missing due to correction, maintenance, non-use, and equipment power cut-off. To check whether there was any abnormality in the data before using them for research, the abnormal data were removed after checking all the data. For the time unit DO concentration prediction study, the missing data were replaced by linear interpolation. For the daily DO concentration prediction study, the time unit data before linear interpolation were averaged and converted into daily data, and then the data converted into daily data were linearly interpolated and used for the study. In previous studies, Lim et al. [31] and Hyo-Joon Jeong [33] suggested that a study linking meteorological data is needed rather than making predictions using only water quality data in water quality prediction research. Accordingly, the meteorological data used for this study were hourly meteorological data (temperature, relative humidity, wind speed, and precipitation) collected from the National Water Resources Management Comprehensive Information System. The missing data from the hourly meteorological data were replaced by linear interpolation, and the meteorological data were converted into daily data by averaging the hourly meteorological data in the same way as the DO concentration data.





**Figure 3.** Monthly averaged DO concentration at the point of the Iseop Bridge in Oncheon stream watershed.

### 2.5. Study Conditions

In this study, as shown in Figure 4, the dissolved oxygen (DO) concentration data and meteorological data (temperature, wind speed, relative humidity, and precipitation) from 2014 to 2017 at the Bugok, Sebyeong, and Iseop points were used as training data. The test data predicted the DO concentration in 2018 and evaluated the actual data. As the experimental environment, the tensorflow library developed by Google was used. As the environmental condition of the long short-term memory (LSTM) algorithm, the mean square error was applied as the loss function. The Adam optimizer was applied as an optimization function applied when learning the weights of the model, and a hyperbolic tangent was applied to the LSTM cell activation function. To compare the learning of the model, the study was conducted while changing the sequence length to 3, 5, and 7, and iteration to 3000, 5000, and 10,000.

We used the coefficient of determination  $R^2$  as the performance measure to evaluate the performance of the prediction model under different conditions.  $R^2$  is widely used for performance evaluation in hydrological and water quality modeling studies [34,35].  $R^2$ , a model of the goodness of fit, is a statistical measure of how closely the model results approximate the observed data.  $R^2$  ranges from 0 to 1, and a value closer to 1 means that the model simulates the actual phenomenon.

In the case when  $R^2$  has a value between 0 and 1, the value requires criteria to evaluate appropriateness to the results of the model. Moriasi et al. [36] suggested the guidelines for model performance evaluation criteria of various performance measures (e.g., coefficient of determination,  $R^2$ ; Nash–Sutcliffe efficiency, NSE; root mean square error, RMSE; and percent bias, PBIAS) based on the synthesis and results of the meta-analysis. Meanwhile, the performance evaluation criteria should be adjusted considering the temporal–spatial scales and measurement data of the simulation model, which can affect the value of the performance measure [36]. Table 1 presents the performance evaluation criteria of  $R^2$  for a watershed scale. As shown in Table 1, the criteria are the strictest for the output response of Flow. In this study, the model predicts the DO concentration discharged in the watershed scale, and the temporal scale of predicted data is hourly and daily. Thus, we referred to the strictest criteria, the output response of Flow in Table 1, as performance evaluation criteria for the prediction model of DO concentration.

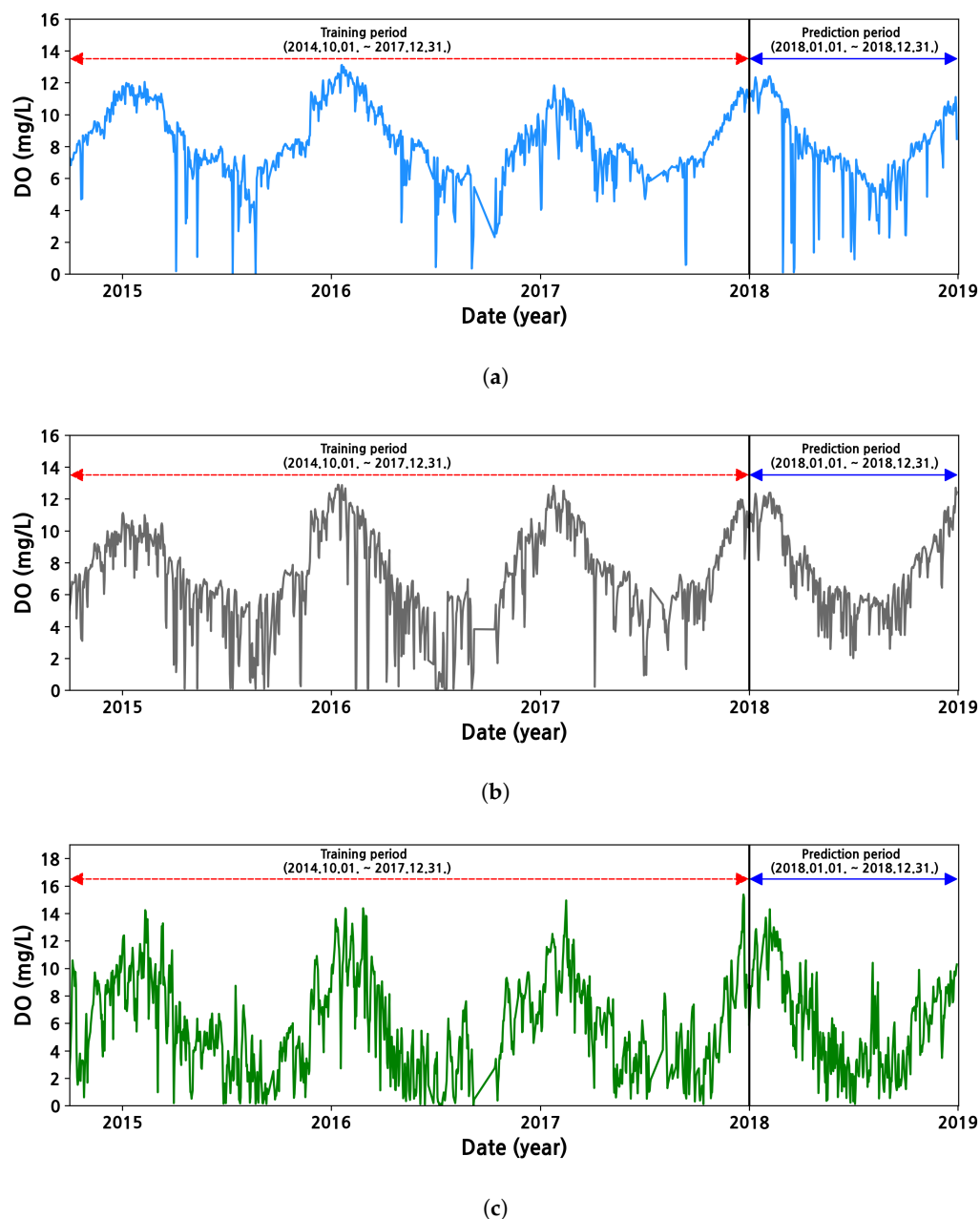


Figure 4. Time series data sets for DO concentration data: (a) Bugok Bridge, (b) Sebyeong Bridge, and (c) Iseop Bridge.

Table 1. Evaluation criteria for the recommended statistical performance measure for watershed scale models.

Measure	Output Response	Temporal Scale *	Performance Evaluation Criteria			
			Very Good	Good	Satisfactory	Not Satisfactory
$R^2$	Flow	D-M-A	$R^2 > 0.85$	$0.75 < R^2 < 0.85$	$0.60 < R^2 < 0.75$	$0.60 > R^2$
	Sediment	M	$R^2 > 0.80$	$0.65 < R^2 < 0.80$	$0.40 < R^2 < 0.65$	$0.40 > R^2$
	N/P	M	$R^2 > 0.70$	$0.60 < R^2 < 0.70$	$0.30 < R^2 < 0.60$	$0.30 > R^2$

Note: \* D, M, and A denote daily, monthly, and annual temporal scales, respectively.

### 3. Results and Discussion

The results calculated by each condition were analyzed through three comparisons. First, the daily prediction and time prediction of the dissolved oxygen (DO) concentration

downstream of the Oncheon stream watershed were analyzed; Table 2 performs an overall analysis of  $R^2$  at the point of the Iseop Bridge (downstream). As a daily prediction result, when using time data, the highest  $R^2$  is 0.8461, and, when forecasting using daily data, the highest  $R^2$  is 0.8008, indicating that time prediction has higher prediction accuracy than a daily prediction, and the lowest  $R^2$  is the  $R^2$  of temporal data prediction and daily data prediction, which are 0.8346 and 0.7388, which show a lot of difference. As a result of the  $R^2$  evaluation, time prediction and day prediction analysis results show that  $R^2$  is 0.8 or higher and both time prediction and day prediction are analyzed as excellent results. In the analysis,  $R^2$  is less than 0.75, indicating satisfactory results. As for the 2-day prediction results, the highest  $R^2$  values are 0.7519 and 0.6931, showing excellent results when using time data at the highest values, but satisfactory results are obtained when using day data. In addition, the lowest  $R^2$  values are 0.7328 and 0.5113, which show satisfactory results when using time data, but are analyzed as unsatisfactory results when using day data. In the 3-day prediction results, most of the results show satisfactory results when time data is used, but most of the results are unsatisfactory when using day data. When the prediction results using time data and the prediction results using daily data are compared, it is confirmed that most of the prediction results using time data are better than the prediction results using daily data. Figure 5 is a scatter graph of the measured values and predicted values for 24 h and 1 day at the Iseop Bridge point when the sequence length is 3 and iteration is 3000. However, it shows a tendency not to show linearity in time prediction.

**Table 2.** Model performance results of the DO concentration (Iseop Bridge point).

Sequence Length	Iterations	$R^2$					
		24 h	1 Day	48 h	2 Day	72 h	3 Day
3	3000	0.8461	0.8008	0.7499	0.6931	0.7039	0.6718
	5000	0.8445	0.796	0.7496	0.6854	0.6987	0.6288
	10,000	0.8429	0.7634	0.7519	0.6716	0.6997	0.5871
5	3000	0.837	0.7889	0.74	0.6489	0.695	0.5868
	5000	0.841	0.79	0.7382	0.6182	0.6921	0.5494
	10,000	0.8346	0.79	0.7439	0.6182	0.6832	0.5494
7	3000	0.8377	0.7639	0.7515	0.5757	0.7044	0.4931
	5000	0.8402	0.7576	0.745	0.535	0.6979	0.4646
	10,000	0.8406	0.7388	0.7328	0.5113	0.6991	0.4272

Second, an analysis of prediction results using upper, middle, and downstream time data and daily data was conducted. In Table 3, the maximum and minimum values of  $R^2$  and the average of the total  $R^2$  are calculated and analyzed for comparison according to the relationship between upper, middle, and downstream time prediction and daily prediction. As a result of forecasting using time data, all of them show excellent results with an  $R^2$  of 0.75 or higher up to the 48 h forecast. One-day prediction results using daily data show excellent results with an  $R^2$  of 0.80 or more, but a 2-day prediction result shows satisfactory results with an  $R^2$  of 0.75 or less. As a result of using time data for 3-day prediction, satisfactory results are shown with an  $R^2$  of 0.60 or more, and, as a result of using daily data, an unsatisfactory result is shown with an average  $R^2$  of 0.60 or less. As for the comparison according to the use of upper, middle, and downstream data, it is analyzed that the prediction is good when only the middle and downstream data are used when time data are used for one-day prediction. On the other hand, when daily data are used, it is analyzed that prediction is good when data from upstream and downstream are used. When using time data, the 2-day prediction is analyzed as the best prediction when only the upstream and downstream data are used, and, when using the day data, the best prediction is performed when the middle and downstream data are used. For the 3-day prediction, it



is analyzed that the best prediction is achieved when upstream and downstream data are used when time data are used, and, when only upstream and downstream data are used, even when day data are used. As a result, the comprehensive analysis does not show much difference when time data are used, but it is analyzed that prediction is good when only upstream and downstream data are used, and, when only upstream and downstream data are used, the prediction is good when day data are used.

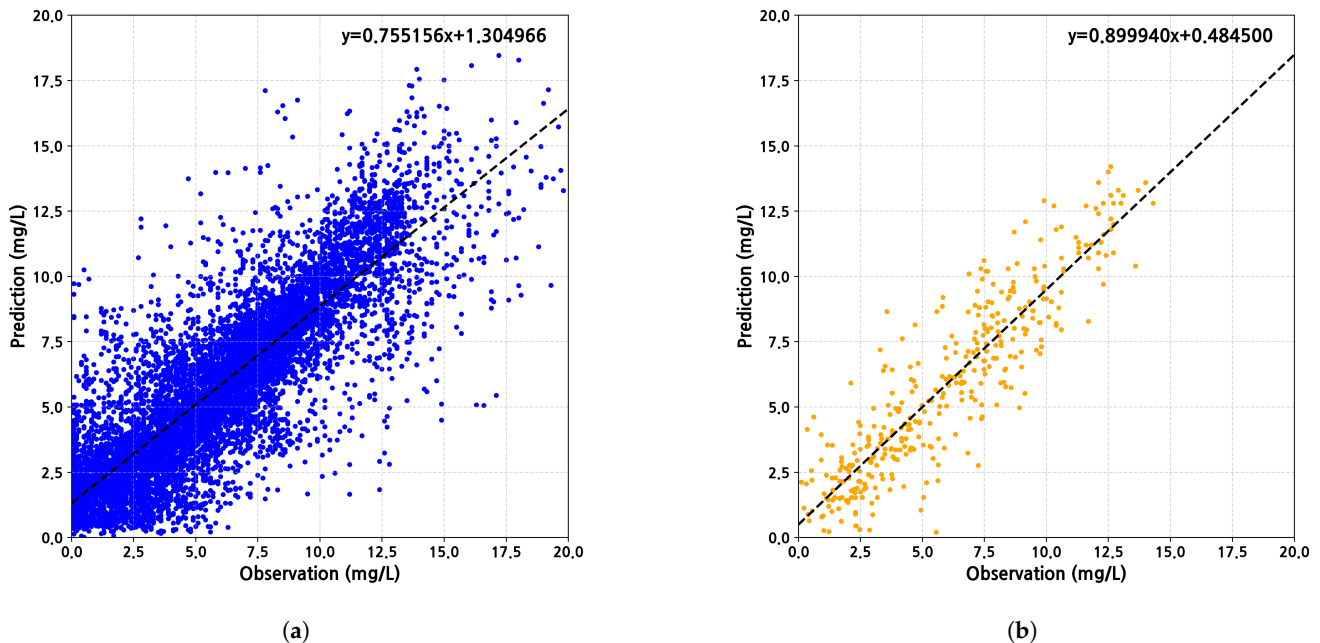


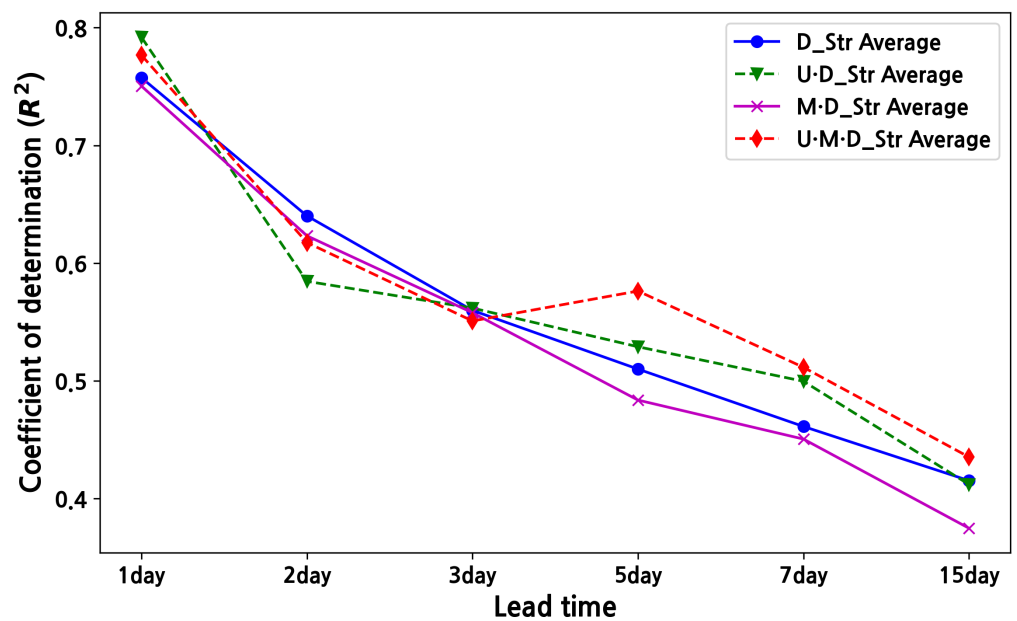
Figure 5. DO concentration scatter plots for lead time of 1 day vs. 24 h: (a) 24 h; (b) 1 day.

Third, analysis was conducted using time data of upper, middle, and downstream. To comprehensively analyze the results of DO concentration prediction using the LSTM algorithm, when upper, middle, and downstream data were used, the  $R^2$  values for lengths 3, 5, and 7 of the sequence length were averaged and summarized in Table 4. As a result of the 1-h, 2-h, and 6-h prediction, all comparison points were analyzed with a very good result with an  $R^2$  of 0.9 or higher. In the 1-hour prediction, the difference between the highest  $R^2$  and the lowest  $R^2$  was 0.0002, and there was little difference. The difference between the highest  $R^2$  and the lowest  $R^2$  in the 2-h and 6-h predictions was 0.0026 and 0.0184, and there was little difference, and it was analyzed that the best results were obtained when only middle and downstream data were used. As a result of 12-h and 24-h prediction, all comparison points were analyzed as excellent results with an  $R^2$  of 0.8 or higher. The results of the other comparison points did not show much difference, except that the prediction results when only the midstream and downstream data were used for 12 h were unusually slightly higher. As a result of the 48-h and 72-h prediction, it was analyzed as a satisfactory result with an  $R^2$  of 0.65 or higher, and there was no significant difference according to the use of upper, middle, and downstream data. Figure 6 shows the values in Table 4 as a graph, and it was confirmed that the accuracy of the prediction rapidly decreased from the 2-h prediction to the 12-h prediction. It was confirmed that the accuracy of the 24-h prediction improved slightly, and then the accuracy of the prediction decreased again.

**Table 3.** Model performance results of the DO concentration (Prediction of hour vs. day).

U·D Str <sup>1</sup>	$R^2$					
	24 h	1 day	48 h	2 day	72 h	3 day
Max	0.8421	0.8156	0.7574	0.6688	0.7145	0.6182
Min	0.8338	0.773	0.7355	0.5405	0.6969	0.4774
Average	0.8388	0.7919	0.7465	0.5846	0.703	0.5618
M·D Str <sup>2</sup>	$R^2$					
	24 h	1 day	48 h	2 day	72 h	3 day
Max	0.8474	0.7776	0.7524	0.6717	0.7028	0.6137
Min	0.8379	0.7273	0.7394	0.5514	0.6932	0.4856
Average	0.8445	0.7504	0.7452	0.6232	0.6983	0.5579
U·M·D Str <sup>3</sup>	$R^2$					
	24 h	1 day	48 h	2 day	72 h	3 day
Max	0.8461	0.8008	0.7519	0.6931	0.7044	0.6718
Min	0.8346	0.7388	0.7328	0.5113	0.6832	0.4272
Average	0.8405	0.7766	0.7447	0.6175	0.6971	0.5509

Notes: <sup>1</sup> U·D Str is an abbreviation for upstream and downstream, which means predicting the DO concentration downstream using the DO concentration data of the up- and downstream, and meteorological data as input data; <sup>2</sup> M·D Str is an abbreviation for midstream and downstream; <sup>3</sup> U·M·D Str is an abbreviation for upstream, midstream, and downstream.



**Figure 6.** DO concentration prediction result of hourly data.

**Table 4.** Model performance results of the DO concentration (Prediction of hour).

Comparison Area	$R^2$						
	1 h	2 h	6 h	12 h	24 h	48 h	72 h
D Str	0.9987	0.9938	0.9285	0.8278	0.8468	0.7443	0.696
U·D Str	0.9986	0.9914	0.9185	0.8263	0.8388	0.7465	0.703
M·D Str	0.9986	0.994	0.9369	0.8531	0.8445	0.7452	0.6983
U·M·D Str	0.9985	0.9923	0.9319	0.8283	0.8405	0.7447	0.6971

Fourth, analysis was conducted using daily data from the upper, middle, and downstream points. Table 5 summarizes the  $R^2$  values for lengths 3, 5, and 7 of the sequence length when using upper, middle, and downstream data to comprehensively analyze the results of daily data prediction. As a result of the 1-day prediction, both  $R^2$  are analyzed as excellent results with a value of 0.75 or higher. The difference between the highest  $R^2$  and the lowest  $R^2$  is 0.0415, which is higher than the 24 h prediction difference of 0.008 in the time prediction. As a result of the 2-day prediction,  $R^2$  is mostly analyzed as a satisfactory result of 0.6 or more, but the results of the upstream and downstream data are less than 0.6, indicating that the accuracy of the prediction is low. The 3-day, 5-day, 7-day, and 15-day predictions are all analyzed as having an  $R^2$  of less than 0.6, indicating that the accuracy of the prediction is low, and there is a difference in the prediction accuracy according to the use of daily data. The difference between the highest  $R^2$  and the lowest  $R^2$  in the 2-day and 3-day prediction results do not show much difference. However, in the 5-day, 7-day, and 15-day forecasts, the difference between the highest  $R^2$  and the lowest  $R^2$  values according to the use of data appear to be large. Figure 7 shows the values in Table 5 as a graph, and it is confirmed that most of the predictions fall sharply. When all of the upper, middle, and downstream data are used, it is confirmed that the accuracy of the forecast increases and then decreases again in the 5-day forecast.

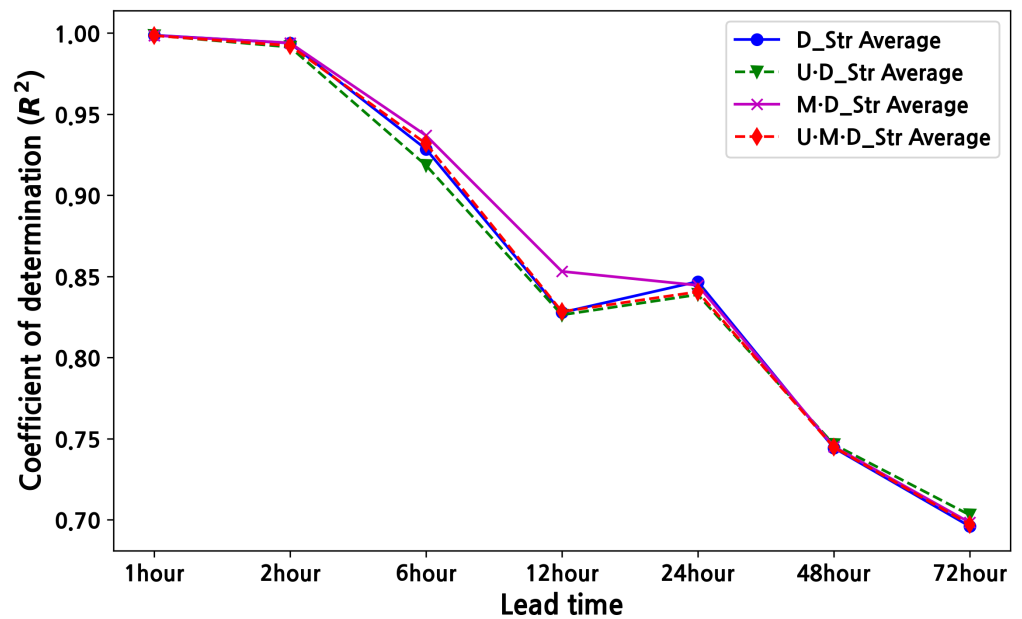


Figure 7. DO concentration prediction result of daily data.

Table 5. Model performance results of the DO concentration (prediction by day).

Comparison Area	$R^2$					
	1 Day	2 Day	3 Day	5 Day	7 Day	15 Day
D Str	0.7575	0.6401	0.5599	0.5101	0.4613	0.4153
U·D Str	0.7919	0.5846	0.5618	0.529	0.4997	0.4121
M·D Str	0.7504	0.6232	0.5579	0.4837	0.4506	0.3749
U·M·D Str	0.7766	0.6175	0.5509	0.5763	0.5115	0.4354

#### 4. Conclusions

In this study, using the long short-term memory (LSTM) algorithm, correlation analysis was conducted between the upper, middle, and lower streams in the prediction of the dissolved oxygen (DO) concentration temporal data and daily data in the Oncheon stream

watershed in Busan. To judge the accuracy of prediction results, meteorological data (average wind speed, average temperature, relative humidity, and rainfall) were collected from 16 January 2014 to 31 December 2018. The study was conducted by dividing the DO concentration data into a training data set and a test data set that does not include the DO concentration data. To quantify the prediction accuracy, the  $R^2$  evaluation index was used for comparison and analysis, and the following conclusions were drawn as a result of the analysis when only downstream data were used and when upstream and midstream data were included:

1. As a result of the prediction of the downstream point using the LSTM algorithm, the change in sequence length and iteration did not show much difference. It was confirmed that the result of the study using time data had slightly higher prediction accuracy than the study using day data and showed a lot of difference. In the daily prediction, the difference between the prediction using time data and the prediction using daily data was not large. However, as the 2-day and 3-day forecasting times increased, the prediction using time data showed higher prediction performance than the prediction using daily data.
2. It was confirmed that the prediction accuracy using the time data was higher than the prediction using the daily data in all of the prediction results using the upper, middle, and downstream time data and daily data. In the prediction of the DO concentration at the downstream point, the data values of the upstream and midstream DO concentration did not seem to affect the prediction of the DO concentration at the downstream point.
3. In the correlation analysis of upper, middle, and downstream data using time data, it appeared that the DO concentration data values of the upstream and middle stream did not affect the DO concentration prediction of the downstream point in the prediction of the DO concentration at the downstream point. In the correlation analysis of upper, middle, and downstream data using daily data, the DO concentration data values of the upstream and middle stream did not affect the prediction of the DO concentration at the downstream point.

The results of analyzing the DO concentration prediction using the LSTM algorithm proposed in this study are as follows. Predictions using time data simulates better than predictions using day data, so it is judged that predictions using time data are more applicable in predicting DO concentration. It was confirmed that the upstream and midstream data did not help much in predicting the DO concentration in the downstream area in the correlation analysis of upper, middle, and downstream temporal data. Furthermore, it was confirmed that they were of little help in the correlation analysis of work data, but not much. Therefore, data from upper, middle, and lower streams and time data rather than daily data are very helpful in predicting DO concentration. It is judged that a long-term forecast for practical application is needed rather than a short-term forecast of three days or less. It is judged that a study using time data is necessary rather than a study using day data, and it is judged that additional studies on external factors that can improve prediction accuracy are needed.

In conclusion, water quality prediction using the LSTM algorithm was only possible for short-term prediction, and limitations of the study were revealed in long-term predictions. However, in this study, the importance of big data was examined through the comparative analysis of daily data and hourly data. There was no correlation between the upstream, midstream, and downstream areas, and additional analysis will be conducted through site-by-site predictions. In addition to predicting DO concentration, it is also considered necessary to predict various water quality factors such as BOD, COD, and SS when additional data are collected.

**Author Contributions:** Conceptualization, H.L. and I.S.; methodology, H.L. and J.L.; software, H.L. and Y.J.; validation, H.S., J.L., and J.D.; investigation, I.S.; data curation, H.S. and J.L.; writing—original draft preparation, H.L. and Y.J.; writing—review and editing, I.S. and Y.J.; visualization, H.L. and Y.J.; supervision, H.L.; project administration, H.L.; funding acquisition, H.L. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work was supported by Korea Environment Industry & Technology Institute (KEITI) through Water Management Program for Drought Project, funded by Korea Ministry of Environment (MOE) (2022003610002). This research was funded by the Institute of Planning and Evaluation for Technology in Food, Agriculture, and Forestry (IPET), grant number 320046053HD020.

**Data Availability Statement:** The data that support the findings of this study are available from the first and corresponding authors upon reasonable request.

**Conflicts of Interest:** The authors declare no conflicts of interest.

## Abbreviations

The following abbreviations are used in this manuscript:

ARIMA	Autoregressive integrated moving average
BOD	Biochemical oxygen demand
Chl-a	Chlorophyll a
COD	Chemical oxygen demand
DO	Dissolved oxygen
D Str	Downstream
HSPF	Hydrological simulation program–Fortran
LSTM	Long short-term model
M Str	Midstream
RNN	Recurrent neural network
TOC	Total organic carbon
U Str	Upstream
WASP	Water quality analysis simulation program
W-ARIMA-GRU	Wavelet decomposition-autoregressive integrated moving average-gated recurrent unit

## References

- Lee, E.; Kim, T. Predicting BOD under Various Hydrological Conditions in the Dongjin River Basin Using Physics-Based and Data-Driven Models. *Water* **2021**, *13*, 1383. [\[CrossRef\]](#)
- Brown, L.C.; Barwell, T.O. *The Enhanced Stream Water Quality Models QUAL2E and QUAL2E-UNCAS: Documentation and User Manual*; EPA: Athens, GA, USA, 1987.
- Ambrose, R.B.; Wool, T.A.; Martin, J.L. *The Water Quality Analysis Simulation Program, WASP5, Part A: Model Documentation*; Environmental Research Laboratory, US Environmental Protection Agency: Washington, DC, USA, 1993.
- Zhou, S.; Song, C.; Zhang, J.; Chang, W.; Hou, W.; Yang, L. A hybrid prediction framework for water quality with integrated W-ARIMA-GRU and LightGBM methods. *Water* **2022**, *14*, 1322. [\[CrossRef\]](#)
- Bicknell, B.R.; Imhoff, J.C.; Kittle, J.L., Jr.; Donigian, A.S., Jr.; Johanson, R.C. *Hydrological Simulation Program Fortran, User's Manual for Release 12*; EPA: Athens, GA, USA, 2001.
- Liu, Z.; Tong, S. Using HSPF to Model the Hydrologic and Water Quality Impacts of Riparian Land-Use Change in a Small Watershed. *J. Environ. Inform.* **2011**, *17*, 1. [\[CrossRef\]](#)
- Taheri Tizro, A.; Ghashghaie, M.; Georgiou, P.; Voudouris, K. Time series analysis of water quality parameters. *J. Appl. Res. Water Wastewater* **2014**, *1*, 40–50.
- Kim, S.; Seo, D.J.; Riazi, H.; Shin, C. Improving water quality forecasting via data assimilation—Application of maximum likelihood ensemble filter to HSPF. *J. Hydrol.* **2014**, *519*, 2797–2809. [\[CrossRef\]](#)
- Kisi, O.; Alizamir, M.; Docheshmeh Gorgij, A. Dissolved oxygen prediction using a new ensemble method. *Environ. Sci. Pollut. Res.* **2020**, *27*, 9589–9603. [\[CrossRef\]](#)
- Kim, H.I.; Kim, D.; Mahdian, M.; Salamattalab, M.M.; Bateni, S.M.; Noori, R. Incorporation of water quality index models with machine learning-based techniques for real-time assessment of aquatic ecosystems. *Environ. Pollut.* **2024**, *355*, 124242. [\[CrossRef\]](#)
- Paliwal, R.; Sharma, P.; Kansal, A. Water quality modelling of the river Yamuna (India) using QUAL2E-UNCAS. *J. Environ. Manag.* **2007**, *83*, 131–144. [\[CrossRef\]](#)
- Park, S.H.; Seo, Y.C.; Kim, Y.H.; Pang, S.P. Big Data-based Monitoring System Design for Water Quality Analysis that Affects Human Life Quality. *J. Korea Entertain. Ind. Assoc.* **2021**, *15*, 289–295. [\[CrossRef\]](#)



13. Panyadee, P.; Champrasert, P.; Aryupong, C. Water level prediction using artificial neural network with particle swarm optimization model. In Proceedings of the 2017 5th International Conference on Information and Communication Technology (ICoICT7), Melaka, Malaysia, 17–19 May 2017; pp. 1–6. [CrossRef]
14. Piasecki, A.; Jurasz, J.; Skowron, R. Application of Artificial Neural Networks (ANN) in Lake Drwęckie Water Level Modelling. *Limnol. Rev.* **2015**, *15*, 21–29. [CrossRef]
15. Jung, S.; Cho, H.; Kim, J.; Lee, G. Prediction of water level in a tidal river using a deep-learning based LSTM model. *J. Korea Water Resour. Assoc.* **2018**, *51*, 1207–1216. [CrossRef]
16. Hong, J.; Lee, S.; Bae, J.H.; Lee, J.; Park, W.J.; Lee, D.; Kim, J.; Lim, K.J. Development and Evaluation of the Combined Machine Learning Models for the Prediction of Dam Inflow. *Water* **2020**, *12*, 2927. [CrossRef]
17. Mok, J.Y.; Choi, J.H.; Moon, Y.I. Prediction of multipurpose dam inflow using deep learning. *J. Korea Water Resour. Assoc.* **2020**, *53*, 97–105. [CrossRef]
18. Kala, A.; Vaidyanathan, S. Prediction of Rainfall Using Artificial Neural Network. In Proceedings of the 2018 International Conference on Inventive Research in Computing Applications (ICIRCA), Coimbatore, India, 11–12 July 2019; pp. 339–342. [CrossRef]
19. Mislan; Haviluddin; Hardwinarto, S.; Sumaryono; Aipassa, M. Rainfall Monthly Prediction Based on Artificial Neural Network: A Case Study in Tenggara Station, East Kalimantan—Indonesia. *Procedia Comput. Sci.* **2015**, *59*, 142–151. [CrossRef]
20. Lowe, M.; Qin, R.; Mao, X. A review on machine learning, artificial intelligence, and smart technology in water treatment and monitoring. *Water* **2022**, *14*, 1384. [CrossRef]
21. Zhao, Z.; Fan, B.; Zhou, Y. An Efficient Water Quality Prediction and Assessment Method Based on the Improved Deep Belief Network—Long Short-Term Memory Model. *Water* **2024**, *16*, 1362. [CrossRef]
22. Kim, Y.; Kwak, S.; Lee, M.; Jeong, M.; Park, M.; Park, Y.G. Determination of Optimal Water Intake Layer Using Deep Learning-Based Water Quality Monitoring and Prediction. *Water* **2024**, *16*, 15. [CrossRef]
23. Wu, J.; Wang, Z. A Hybrid Model for Water Quality Prediction Based on an Artificial Neural Network, Wavelet Transform, and Long Short-Term Memory. *Water* **2022**, *14*, 610. [CrossRef]
24. Wu, X.; Zhang, Q.; Wen, F.; Qi, Y. A Water Quality Prediction Model Based on Multi-Task Deep Learning: A Case Study of the Yellow River, China. *Water* **2022**, *14*, 3408. [CrossRef]
25. Ay, M.; Kisi, O. Modeling of Dissolved Oxygen Concentration Using Different Neural Network Techniques in Foundation Creek, El Paso County, Colorado. *J. Environ. Eng.* **2012**, *138*, 654–662. [CrossRef]
26. Ahmed, A.M. Prediction of dissolved oxygen in Surma River by biochemical oxygen demand and chemical oxygen demand using the artificial neural networks (ANNs). *J. King Saud Univ.-Eng. Sci.* **2017**, *29*, 151–158. [CrossRef]
27. Shim, K.; Gyeonghoon, K.; Seongmin, K.; Youngseok, K.; Jin-pil, K. Comparison of Changes in Upstream and Downstream Water Quality of Tributary Rivers: Gyeseong-stream and Hwapo-stream in Nakdongmiryang Watershed. *J. Korean Soc. Water Environ.* **2020**, *36*, 445–452. [CrossRef]
28. Chun, H.J.; Yang, H.S. A Study on Prediction of Housing Price Using Deep Learning. *J. Resid. Environ. Institue Korea* **2019**, *17*, 37–49. [CrossRef]
29. Hochreiter, S.; Schmidhuber, J. Long Short-term Memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef]
30. Olah, C. Available online: <https://colah.github.io/posts/2015-08-Understanding-LSTMs/> (accessed on 5 May 2024).
31. Lim, H.; An, H.; Choi, E.; Kim, Y. Prediction of the DO concentration using the machine learning algorithm: Case study in Oncheoncheon, Republic of Korea. *Korean J. Agric. Sci.* **2020**, *47*, 1029–1037. [CrossRef]
32. Choi, C.; Kim, E.; Kim, K.; Kim, S. Application of Detention and Infiltration-based Retention Hybrid Design Technique to Oncheon Stream. *KSCE J. Civ. Environ. Eng. Res.* **2011**, *31*, 99–108. [CrossRef]
33. Jeong, H.J.; Lee, S.J.; Lee, H.K. Water Quality Forecasting of Chungju Lake Using Artificial Neural Network Algorithm. *J. Environ. Sci. Int.* **2002**, *11*, 201–207.
34. Kim, J.; Chae, S.K.; Kim, B.S. Evaluation of Water Quality Prediction Models at Intake Station by Data Mining Techniques. *J. Environ. Impact Assess.* **2011**, *20*, 705–716.
35. Kim, M.; Shin, H. Study on Establishing Algal Bloom Forecasting Models Using the Artificial Neural Network. *J. Korea Water Resour. Assoc.* **2013**, *46*, 697–706. [CrossRef]
36. Moriasi, D.N.; Gitau, M.W.; Pai, N.; Daggupati, P. Hydrologic and Water Quality Models: Performance Measures and Evaluation Criteria. *Trans. ASABE* **2015**, *58*, 1763–1785. [CrossRef]

**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.