Journal of
Marine Science and Engineering

MDPI

Article

Temporal Prediction of Coastal Water Quality Based on Environmental Factors with Machine Learning

Junan Lin [1], Qianqian Liu [2,3,*], Yang Song [4], Jiting Liu [5], Yixue Yin [6] and Nathan S. Hall [7]

[1] Department of Mechanical and Process Engineering, Swiss Federal Institute of Technology in Zurich, 8092 Zurich, Switzerland; linjun@student.ethz.ch
[2] Department of Physics and Physical Oceanography, University of North Carolina Wilmington, Wilmington, NC 28403, USA
[3] Center for Marine Science, University of North Carolina Wilmington, Wilmington, NC 28409, USA
[4] Department of Computer Science, University of North Carolina Wilmington, Wilmington, NC 28403, USA; songy@uncw.edu
[5] Department of Computer Science, Columbia University, New York, NY 10027, USA; jl6247@columbia.edu
[6] Department of Information Networking Institute, Carnegie Mellon University, Pittsburgh, PA 15213, USA; yixuey@andrew.cmu.edu
[7] Institute of Marine Sciences, University of North Carolina Chapel Hill, Morehead City, NC 28557, USA; nshall@email.unc.edu
* Correspondence: liuq@uncw.edu

Abstract: The accurate forecast of algal blooms can provide helpful information for water resource management. However, the complex relationship between environmental variables and blooms makes the forecast challenging. In this study, we build a pipeline incorporating four commonly used machine learning models, Support Vector Regression (SVR), Random Forest Regression (RFR), Wavelet Analysis (WA)-Back Propagation Neural Network (BPNN) and WA-Long Short-Term Memory (LSTM), to predict chlorophyll-a in coastal waters. Two areas with distinct environmental features, the Neuse River Estuary, NC, USA—where machine learning models are applied for short-term algal bloom forecast at single stations for the first time—and the Scripps Pier, CA, USA, are selected. Applying the pipeline, we can easily switch from the NRE forecast to the Scripps Pier forecast with minimum model tuning. The pipeline successfully predicts the occurrence of algal blooms in both regions, with more robustness using WA-LSTM and WA-BPNN than SVR and RFR. The pipeline allows us to find the best results by trying different numbers of neuron hidden layers. The pipeline is easily adaptable to other coastal areas. Experience with the two study regions demonstrated that enrichment of the dataset by including dominant physical processes is necessary to improve chlorophyll prediction when applying it to other aquatic systems.

Keywords: water quality forecast; coastal ocean; algal blooms; machine learning models

check for updates

Citation: Lin, J.; Liu, Q.; Song, Y.; Liu, J.; Yin, Y.; Hall, N.S. Temporal Prediction of Coastal Water Quality Based on Environmental Factors with Machine Learning. *J. Mar. Sci. Eng.* **2023**, *11*, 1608. https://doi.org/10.3390/jmse11081608

Academic Editors: Wei Qin and Georgios Sylaios

Received: 17 July 2023
Revised: 2 August 2023
Accepted: 14 August 2023
Published: 17 August 2023

1. Introduction

Chlorophyll *a* (Chl-a), a pigment that absorbs the light needed for plants to photosynthesize, is a measure of the amount of phytoplankton in a water body. It has been used as an indicator of algal blooms and the state of water quality in coastal and estuarine oceans. With harmful algal blooms (HABs) and an overgrowth of algae in water becoming a major environmental problem in aquatic systems, algal bloom monitoring and prediction are crucial for water management [1–4]. The accurate Chl-a simulation and forecast, even just a few days in advance of bloom occurrence, can provide useful early warning information to water managers and the public for decision making [2,5–8]. Different approaches have been taken to predict Chl-a, which in general are categorized into two groups: mechanistic models [9–12] and data-driven models [13–16].

The mechanistic models depend on a mechanistic understanding of the relationship between physiological processes and environmental factors, because the phytoplankton's growth is driven by multiple factors, including temperature, daily duration of sunlight, sunlight intensity, nutrients concentrations (e.g., nitrate, ammonium, phosphate, silicate), etc. Such models use predefined characteristics and assumptions about how the primary and secondary producers interact with the environment, as well as predator–prey interactions and nutrient cycling [17,18]. Although powerful, the wide implementation of mechanistic models is limited by several factors. Firstly, a mechanistic ecological model usually needs to be coupled with a complex hydrodynamic model. Often, hydrodynamic simulations have a large uncertainty, generating errors that propagate to water quality models [19]. Secondly, mechanistic models often simplify the phytoplankton community to include only a few functional groups. Information on the rates/parameters of processes impacting the phytoplankton is usually estimated based on laboratory studies or empirical observations from similar water bodies found in the literature, which can also cause errors. Thirdly, as the mechanistic models get more complex, they require more resources in computational time, energy and sampling to calibrate and validate the model.

With the extensive availability and quick accumulation of in situ and remotely sensed data and the rapid progress in the development of complex data-driven models such as piecewise regressions, wavelet analysis (WA) and deep neural networks to fill the spatial and temporal gaps in field observations, a growing number of data-driven models have been developed to forecast water quality [20–23]. However, the complexity of stressors, including solar radiation, nutrients, temperature and salinity and stressor interactions make it challenging to predict Chl-a [24] with a standard system/process that can be applied to multiple areas. In this paper, we build a pipeline solution that can potentially automate the model building and prediction process and help predict blooms in different aquatic systems. Using the HABs-prone Neuse River Estuary (NRE) in NC, U.S., and the Scripps Pier station, CA, U.S., as examples, we build a pipeline composed of data processing, feature engineering and deep neural networks. This automated pipeline can allow researchers to enrich a dataset using any features that could be helpful for Chl-a prediction and produce the optimal model for a given aquatic system.

The NRE is a sub-estuary to the Pamlico Sound estuarine system, which is the second largest estuarine complex in the U.S (Figure 1a). Along the U.S. east coast, approximately half of the nursery area for commercially fished species occurs within the Pamlico Sound system [25]. The NRE is a particularly important spawning area for many commercially and recreationally important fish including shad, striped bass and red drum, and an important harvest area for shrimp and blue crabs [26]. Additionally, the Pamlico Sound system supports a $3.7 billion outdoor recreation industry, much of which is related to water activities including swimming, sailing, paddling, waterfowl hunting and recreational fishing [27]. Therefore, the sustainable management of NRE-Pamlico Sound coastal ecosystems is essential for the environment and benefits local economies by maintaining fish populations and adequate water quality for recreation.

The NRE has a history of eutrophication symptoms, including hypoxia and HABs [28–30]. Serving as an important fisheries habitat and recreational area, accurate forecasts of water quality in NREs are critical for decision making for various stakeholders, including water resource and fisheries managers and anglers. The monitoring and prediction of Chl-a, an important indicator of water quality, is a critical step in sustainable ecological management. Long-term (1994—present) monitoring of Chl-a and other biogeochemical and ecological parameters has been conducted by the Neuse River Estuary Modeling and Monitoring Program (ModMon; [31]) and remotely via satellite observations (e.g., by NOAA VIIRS-SNPP).

Red tides and other HABs have also been a burden on the coastal ecosystems of California, resulting in fish and shellfish mortality, poisoning in marine mammal/bird populations, illness including respiratory irritation and failure and indirect impacts on the economy [6,7,32]. Programs including the Harmful Algal Bloom Monitoring and Alert Program (HABMAP) have been implemented to monitor HABs and facilitate information exchange among scientists (https://calhabmap.org (accessed on 1 August 2022)). The Scripps Pier is on the coast of California, USA, in La Jolla (Figure 1b). It is one of the monitoring sites offering regular and frequent samplings of biological and environmental variables, now a part of the Southern California Coastal Ocean Observing System (SCCOOS; http://www.sccoos.org (accessed on 12 January 2023)). Different approaches including machine learning attempts have been used to predict water quality in this area [7,22,33]. In this study, we apply our machine learning pipeline to the Scripps Pier station to predict Chl-a and examine the pipeline's performance.

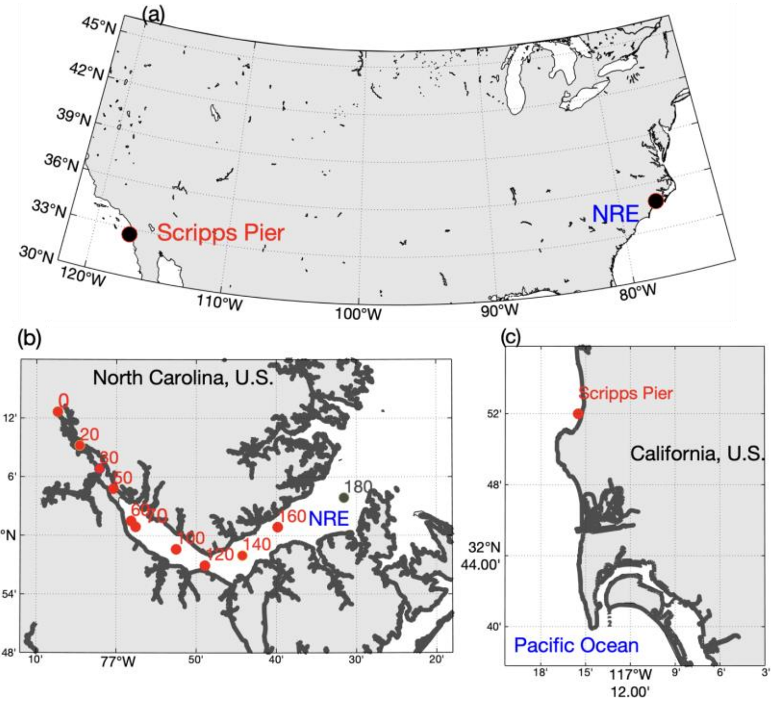

Figure 1. Locations of study areas, the NRE system and Scripps Pier, in the USA (**a**), and locations of sampling stations at the (**b**) NRE and (**c**) Scripps Pier. Red dots in both study areas represent sampling stations, with Station 180 (**b**) being excluded due to the lack of data from 1994 to 1998.

In this paper, we build a pipeline incorporating four commonly used machine learning models, Support Vector Regression (SVR), Random Forest Regression (RFR), Wavelet Analysis (WA)-Back Propagation Neural Network (BPNN) and WA-Long Short-Term Memory (LSTM), to find the optimal model for Chl-a prediction in NRE, NC and the Scripps Pier, CA.

2. Materials and Methods

2.1. Datasets

To predict water quality at the NRE and Scripps Pier, we collected observational data from various reliable and publicly available sources to build data-driven models and assess the models' performance through statistical matrices.

For the NRE system, we used 24 years (1994–2017) of historical hydrologic and water quality data from the ModMon program. The ModMon dataset comprises 11 mid-river water quality sampling stations along the NRE from the river head to its mouth at Pamlico Sound (Figure 1a). The program sampled hydrographic, chemical and ecological parameters from the surface (with sampling depths < 1 m from the surface, of which 99% are within 0.5 m from the surface) and bottom (0.5 m above bottom) depths on an approximate bi-weekly basis throughout the year [28]. To develop the machine learning models, we built a dataset to include non-equal (irregular) interval surface sampling of 10 NRE stations (Station 0, 20, 30, 50, 60, 70, 100, 120, 140, 160) from 1994 to 2017. Station 180 data were excluded from the dataset because the station 180 dataset lacked data from 1994 to 1998. For the accuracy of the data-driven model, we kept 10 ModMon features that are related to Chl-a in aquatic systems as input forcings, which are date, the distance (km) of the station downstream from Station 0, water temperature, salinity, dissolved oxygen (DO), pH value, particulate organic carbon (POC), nitrate/nitrite (NO_3/NO_2), ammonium (NH_4) and orthophosphate (PO_4) (see Table A1 in Appendix A for their units). Chl-a was included in the dataset for model calibration and assessment.

Our datasets also included daily averaged weather data, including winds, air pressure and air temperature at NOAA station near Cape Lookout Bight, NC (station CLKN7; https://www.ndbc.noaa.gov (accessed on 1 August 2022)), and daily averaged river flow from the Neuse River at the US Geological Survey (USGS) site 02091814 near Fort Barnwell, NC (https://waterdata.usgs.gov/nwis/ (accessed on 1 August 2022)). To capture the weather's cumulative effect on Chl-a, we used the accumulative form of wind speed, pressure, air temperature and discharge. It is derived by adding data from the observations over continuous n days; the n value selection will be discussed in Section 3.1. In the contents below, the dataset that combines ModMon data, weather data and the Neuse River discharge is referred to as the "NRE dataset".

The dataset for Scripps Pier water quality prediction (Figure 1b) was from the Southern California Coastal Ocean Observing System (SCCOOS, https://erddap.sccoos.org/erddap/tabledap/HABs-ScrippsPier.html (accessed on 1 August 2022)) [22]. This dataset included NH_4, Chl-a, NO_3, NO_2, phaeophytin, PO_4, silicate (SiO_4) and water temperature. Chl-a was used for model assessment. Unlike the NRE dataset, samplings of Scripps Pier are nearly equally spaced (shown in the Appendix A) and collected at a single station. Like the NRE dataset, we combined the SCCOOS Scripps Pier data with wind data from NCEP North American Regional Reanalysis (NARR). The combined dataset is called the "Scripps Pier dataset" below.

Different variables were included in the modeling datasets for these two areas, which were determined by data availability in each area and their distinct environment features. Both datasets included nutrient features and meteorological data, considering their strong correlation with phytoplankton growth and water mixing. In contrast, the NRE, an estuary with large freshwater input, is more susceptible to river discharge and seawater intrusion compared with the Scripps Pier station. Therefore, river discharge and offshore distance from the river mouth station were included in the NRE dataset. The variables are further selected as input features for the machine learning models (Section 2.3).

2.2. Methods

The pipeline we implemented for NRE and Scripps Pier water quality prediction used models that have shown reliable performance in water quality prediction. It was run on Google Colaboratory with the free plan (a 2-core CPU with 12.7 GB RAM). It included two benchmark models, Support Vector Regression (SVR) and Random Forest Regression (RFR) [34,35] and another two coupled models that are more commonly adopted for water quality data processing and prediction, WA-Back Propagation Neural Network (BPNN) and WA-Long Short-Term Memory (LSTM) [23,35,36].

2.2.1. Support Vector Regression (SVR)

SVR is a supervised learning algorithm used for prediction [37], based on the statistical learning theory by Cortes and Vapnik [38]. It determines two flexible hyperplanes with the same distance, ε, around the predicted function symmetrically. The model only penalizes the samples outside the boundaries with a penalization coefficient C, reducing prediction error by minimizing the sum of the squared weight (w) and the penalization terms (ξ and ξ^*), $||w||^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i{}^*)$ (Figure 2).

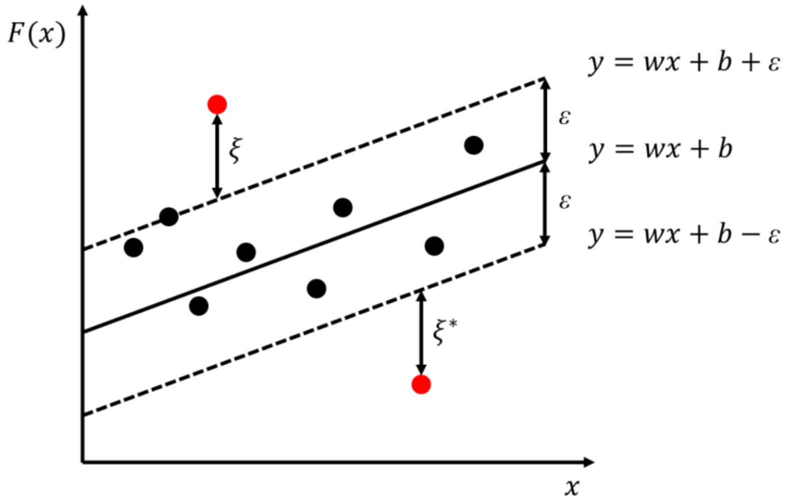

Figure 2. Support vector regression scheme, in which C is the penalization coefficient and ε is the distance between the boundary and the predicted function, both of which are hyperparameters that are pre-determined. ξ is the distance from a sample satisfying $y > wx + b + \varepsilon$ to the upper boundary $y = wx + b + \varepsilon$, and ξ^* is the distance from a sample satisfying $y < wx + b - \varepsilon$ to the lower boundary $y = wx + b - \varepsilon$. For a single penalized sample (represented by red dots), depending on where it lies, either ξ or ξ^* is larger than 0. When the sample satisfies $y > wx + b + \varepsilon$, $\xi > 0$ and ξ^* should be treated as 0; when it satisfies $y < wx + b - \varepsilon$, $\xi^* > 0$ and ξ should be treated as 0.

2.2.2. Random Forest Regression (RFR)

RFR is a supervised learning algorithm that uses an ensemble learning method for regression, which combines several decision trees and aggregates multiple machine learning models to make better predictions than a single model [39]. The decision trees are constructed in parallel, with no interaction among them. The maximum number of features used at each tree and the number of trees to be grown are two user-defined parameters. At each node, one sample feature among the selected features is searched for the best splitting. The random forest regression consists of k trees, where k is the user-defined number of trees to be grown [39–41]. The diagram in Figure 3 shows the structure of a Random Forest with 200 decision trees.

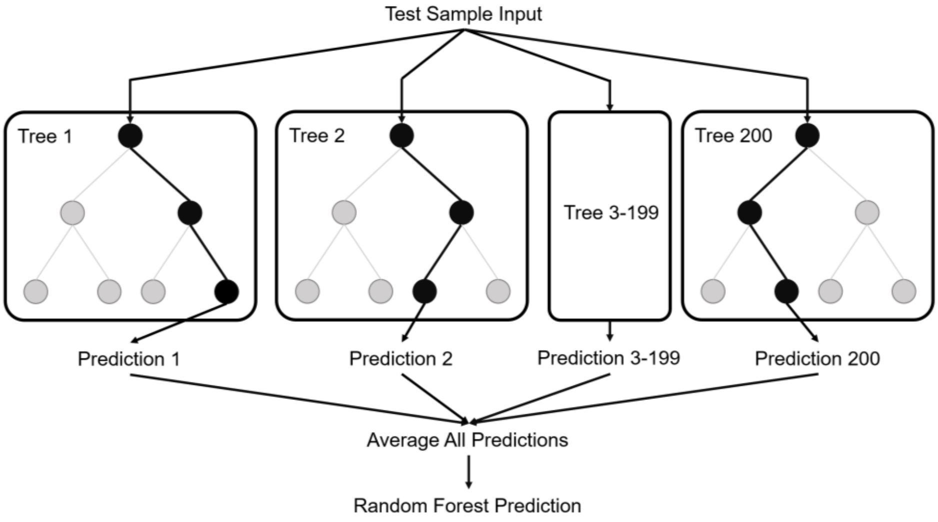

Figure 3. Random forest regressor scheme with 200 decision trees.

2.2.3. Wavelet Analysis (WA)

WA is a method that can decompose a signal into elementary forms at different scales and positions and then reconstruct the signal with high precision. WA has been widely applied in noise filtering, data compression and signal analysis [42]. Compared with the classic Fourier transform which only provides information on signals in frequency space, WA can capture both time and frequency information [42,43]. Therefore, WA can better resolve non-stationary signals and has demonstrated reliable performance in numerical applications in different fields.

Multiple wavelet-related transforms exist for different applications. In this study, we used discrete wavelet transform (DWT) to decompose the Chl-a concentration into detailed or approximate components. The DWT is discrete in shift and scale parameters, but continuous in time. Consequently, WA analysis of the NRE and Scripps datasets required interpolation of the datasets to a constant time interval as described below. $\psi_{m,\,n}(t)$ is the set of child wavelets, which is scaled and shifted from the given mother wavelet $\psi(t)$. Following Akansu et al. [44], we defined the child wavelet as:

$$\psi_{m,\,n}(t) = {a_0}^{-\frac{m}{2}}\psi\left({a_0}^{-m}t - nb_0\right) \tag{1}$$

where a_0, b_0, m, n are constant integers, with m being the scale parameter and n being the shift parameter. As a dyadic discrete wavelet, we set a_0 as 2 and b_0 as 1 [45]. Equation (1) can be rewritten as

$$\psi_{m,\,n}(t) = 2^{-\frac{m}{2}}\psi\left(2^{-m}t - n\right) \tag{2}$$

Then, we used the child wavelet to define the wavelet transform of a signal $f(t)$ [44].

$$d_{m,\,n}(t) = 2^{-\frac{m}{2}}\int f(t)\psi\left(2^{-m}t - n\right)dt \tag{3}$$

By using DWT, a time series can be divided into two different types of coefficients: approximation and detail coefficients at every decomposition level. The result can be shown as a binary-like tree (Figure 4). The coefficients' frequency and vector length gradually decrease to half of the upper level as we move down the decomposition tree. For example, the detail coefficient, cD1, in the first level should be half the length of the original signal, *S*. If the original signal, *S*, has a frequency ranging from 0 to f_n, the cD1 frequency should range from $f_n/2$ to f_n. Generally, only approximation coefficients are decomposed continuously [46], so our model used detail components in all levels and the approximation component in the deepest level only as features.

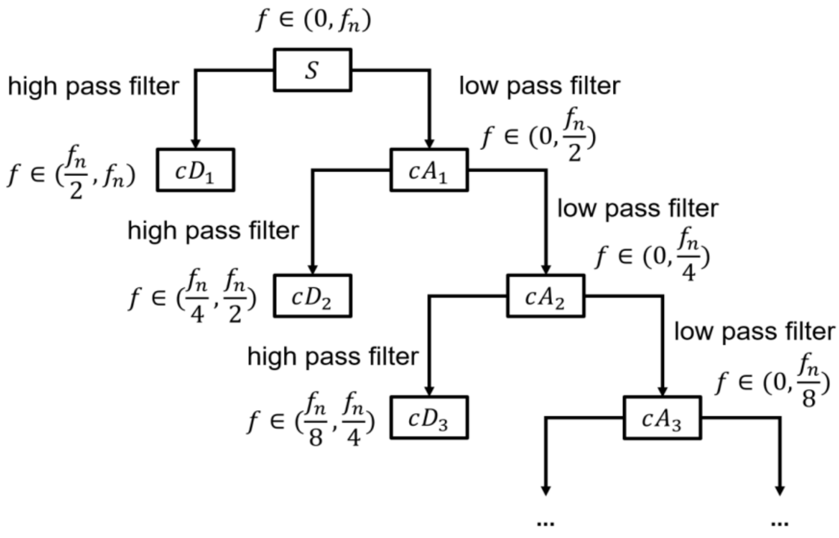
$f \in (0, f_n)$
S
high pass filter
low pass filter
$f \in (\frac{f_n}{2}, f_n)$
cD_1
cA_1
$f \in (0, \frac{f_n}{2})$
high pass filter
low pass filter
$f \in (\frac{f_n}{4}, \frac{f_n}{2})$
cD_2
cA_2
$f \in (0, \frac{f_n}{4})$
high pass filter
low pass filter
$f \in (\frac{f_n}{8}, \frac{f_n}{4})$
cD_3
cA_3
$f \in (0, \frac{f_n}{8})$
...
...

Figure 4. The structure of wavelet decomposition. S represents the original signal; *cDx* and *cAx* represent the detail component and approximation component coefficients in each decomposition level.

2.2.4. Back Propagation Neural Network (BPNN)

BPNN is one of the most typical neural network models [47], which is also known as Artificial Neural Network (ANN). It consists of an input layer, several hidden layers and an output layer (Figure 5). For the forward process, each node (except that in the input layer) receives the linear combination of outputs from the previous layer as its input and generates output to the next layer by applying an activation function ($f[\cdot]$ in Figure 5) to the input. Classic activation functions include sigmoid, tanh, ReLU, etc. To minimize error, back propagation should be carried out to properly adjust the weight of the linear combination mentioned above, in which the partial derivative of the loss function with respect to all weights is calculated, and the weights are changed in the direction of the steepest descent of the loss function. Compared with traditional machine learning models, a significant advantage of BPNN is that it can simulate all functions, theoretically. In this study, we used the Pytorch package to construct the network with one hidden layer and chose the ReLU as the activation function. The loss was calculated by the mean squared error (MSE) function.

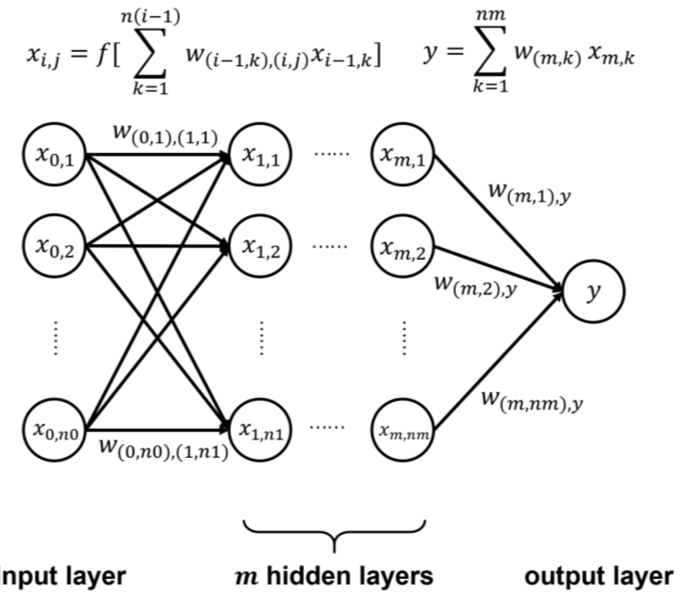

Figure 5. A general structure of BPNN. In the flowchart, $x_{0,j}$ represents the j$^{\text{th}}$ input feature; $x_{i,j}$ $(i = 1, 2, \ldots, m)$ represents the j$^{\text{th}}$ value in the i$^{\text{th}}$ hidden layer; $w_{(i-1,k),(i,j)}$ represents the weight of $x_{i-1,k}$ in the linear combination of $x_{i,j}$; $w_{(m,k),y}$ represents the weight of $x_{m,k}$ in the linear combination of the result y; and $f[\cdot]$ is the activation function.

2.2.5. Long Short-Term Memory Networks (LSTM)

LSTM is a well-known neural network that has been used in time-series prediction [48], including successful prediction of temporal-based water quality of Chl-a concentration [21], total phosphorus [49] and dissolved oxygen [34,49]. Therefore, we chose LSTM as another machine learning model to predict Chl-a.

A standard LSTM unit comprises three key structures (Figure 6): a forget gate, the input gate layer and the output gate layer. The first step for an LSTM model design is to decide what should be discarded by the forget gate layer. This gate maps the output, h_{t-1}, from the last cell and the input, x_t, to a feature vector within [0, 1] through a sigmoid layer. Next, the input gate layer will decide what will be updated and stored in the new cell state, C_t, by using it and $\widetilde{C}_t$. Finally, the output information, h_t, will be filtered after the output gate layer based on our cell state.

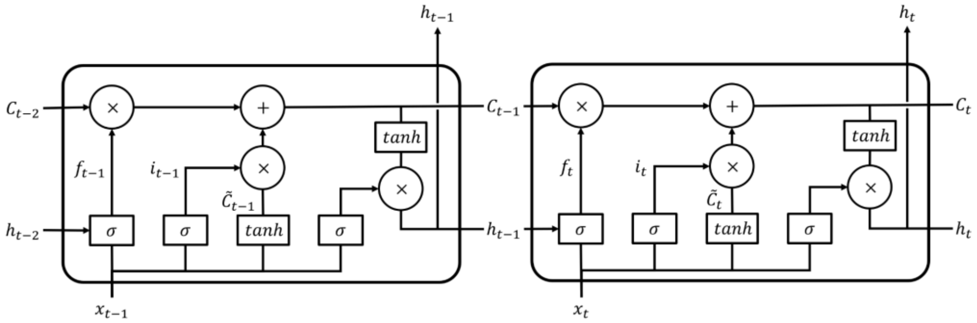

Figure 6. Scheme of LSTM neurons. For each time step, t, the forget gate first helps the cell state, C_{t-1}, to forget unnecessary information by multiplying it with f_t, the sigmoid-processed information provided by output, h_{t-1}, and current time step state, x_t. Then, the input gate helps the cell state, C_{t-1}, to obtain new information by adding it with the multiplication of i_t (information about which values should be updated) and $\widetilde{C}_t$ (candidate values that can be added to the cell state). In the end, the current output, h_t, is derived from tanh-processed cell state, C_t, with the help of the multiplication of sigmoid-processed x_t so that we only obtain part of the state as the output as we wish.

2.3. Feature Engineering

Water quality change can lag behind its predictors, and the lag is determined by domain features [50,51]. To better capture the time sequence of water quality information, we aggregated features of wind speed, water pressure and air temperature over several days preceding the prediction time with window sizes of 1 to 7 days and 30 days, then applied an input feature selection method to evaluate these aggregated features.

Feature selection is a necessary step in modeling nonlinear systems [52]. Selecting the combination of the best subset of features and ignoring irrelevant features can effectively improve model accuracy/reliability, lead to shorter training time and reduce both input dimensionality and unnecessary model complexity. Many features have been selected for Chl-a prediction, of which certain features are more useful than others [21,34]. In this study, we applied the chi-squared (CHI) algorithm to identify the most relevant features for Chl-a prediction. CHI is a non-parametric algorithm that evaluates correlations between variables and assesses whether the independents are positively correlated or not. It performs best for multi-class data [53]. The higher the CHI score, the more relevant the feature is, and it can be selected for model training. To reduce errors in model training and allow models to converge faster, we normalize the input features into [0, 1] for prediction [54].

2.4. Model Assessment

The models' performance in predicting Chl-a concentrations was measured by Root Mean Square Errors (RMSE) and R^2 score. In addition, we used skill metrics for binary event forecasts [8,55] to assess the model's performance in predicting the occurrence of algal bloom, with Chl-a exceeding a threshold of 40 μg/L [56]. The metrics include the probability of detection (*POD*), the probability of false detection (*POFD*), frequency bias (*B*) and the Pierce skill score (*PSS*). They are defined as,

$$POD = \frac{a}{a+c} \tag{4}$$

$$POFD = \frac{b}{b+d} \tag{5}$$

$$B = \frac{a+b}{a+c} \tag{6}$$

$$PSS = POD - POFD = \frac{ad - bc}{(b+d)(a+c)} \tag{7}$$

where a represents the number of correctly predicted occurrence of blooms (hits); b, incorrectly predicted blooms (false alarms); c, false negatives (misses); and d, correctly predicted absence of blooms. Higher PODs represent better performance with values in the range [0, 1], while higher $POFD$s represent worse performance in the range of [0, 1]. B is the ratio of predicted events to observed events, in the range of $[0, \infty]$, with 1 representing an unbiased forecast. PSS is in the range of $[-1, 1]$, with larger values representing a better performance.

3. Data Engineering and Model Building

3.1. Data Processing

Data cleaning is the process of improving data quality by fixing and removing incorrect, corrupted, duplicate and incomplete data within a dataset, and it should be performed before model training. For the NRE dataset, river discharge and atmospheric data are available at regular intervals. However, the ModMon dataset sampling has irregular intervals (Table A2 in Appendix A). Because WA requires strictly equally spaced temporal data, we linearly interpolated water quality data using an interval of 14 days. For all 10 stations used, we chose 5 January 1999 as the first day of linear interpolation and 31 October 2017 as the last day. With 14 days as the forced interval, there are 474 samples for each depth at each station. Among them, data earlier than 24 March 2014 (80% of all data) were used as training data, while the rest of the data were considered as testing data.

The Scripps Pier dataset in [22] has a regular sampling interval of ~7 days (Table A3 in Appendix A), so the original dataset was used for prediction. Similarly for the NRE dataset, the first 80% of Scripps Pier data were selected as training data and the rest as testing data. The numbers of observations in the original (before-interpolation) dataset used in the training data and testing data for both areas are shown in Table A4 in Appendix A.

Then, we constructed useful accumulative features using the aggregation approach mentioned in Section 2.3 and evaluated their relevance by CHI algorithm feature selection. For the NRE dataset, the results of feature importance for different window sizes are shown in Figure 7. The features with the first highest local feature importance were selected, including accumulative wind speed over 2 days preceding the time at which the forecast is made for (referred to as prediction date below), air pressure over 3 days preceding the prediction date and air temperature and river discharge over 5 days preceding the prediction date. For the Scripps Pier dataset, through the feature selection process, we found that the weather data on the prediction date showed the highest feature importance.

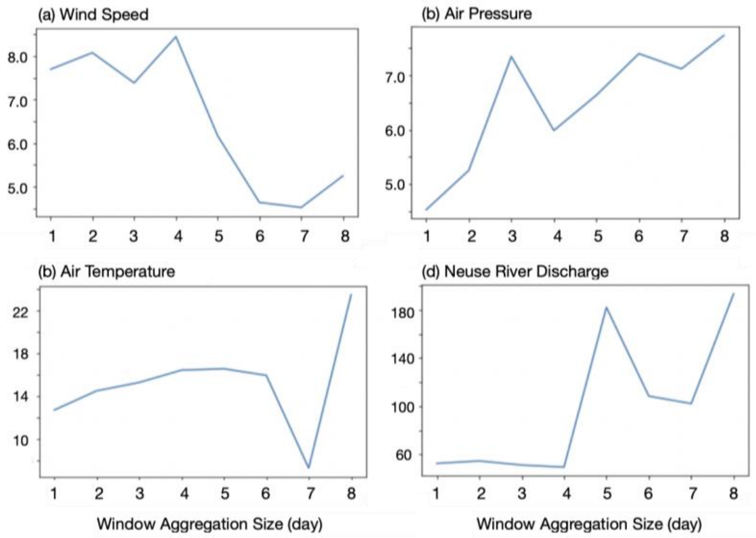

Figure 7. Feature importance (CHI score) for 1–8 days aggregated (**a**) Wind speed, (**b**) air pressure, (**c**) air temperature and (**d**) the Neuse River discharge.

3.2. Model Implementation

To tune the baseline models of RFR and SVR, we utilized the grid search method to find the optimal hyperparameters of a model. Grid search, as one of the widely used global optimization approaches, searches exhaustively to find the best parameter combination out of all the possible parameter choices [57,58].

When tuning RFR, we set the parameters "max depth", "max features", "min samples split" and "n estimators" as objects of the grid search. When tuning SVR, we also set parameters such as C, gamma and Kernel types (e.g., polynomial kernel, linear kernel and Gaussian radial basis kernel).

For the WA-LSTM and WA-BPNN models, WA was applied to the Chl-a feature before model training. We chose Wavelet Daubechies 4 (db4; [59]) as our wavelet type, and constructed a 4-level WA model to decompose the Chl-a features (Figure 8). When building the neural network models of LSTM and BPNN, CHI feature importance was implemented again to select the most important features in the NRE and Scripps Pier datasets, including features from the original dataset and the additional accumulative features constructed above by aggregating over a window size. Together, we had 30 input features for the NRE model and 22 for the Scripps Pier. We ranked the features by their feature importance from highest to lowest and chose the first *P* features for model training, in which the optimal *P* was determined by traversals.

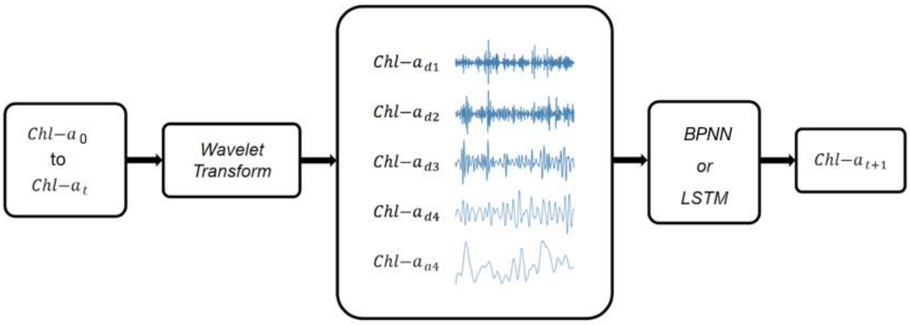

Figure 8. Construction of the WA process for LSTM and BPNN models. $Chl - a_t$ represents Chl-a at time t. $Chl - a_{dx}$ represents the detail component at the x$^{\text{th}}$ decomposition level and $Chl - a_{a4}$ represents the approximation component at the 4$^{\text{th}}$ decomposition level.

This data processing was carried out for both BPNN and LSTM. Finally, we searched for the best subset of features with an increase of three features from half of the features to the number of all the features and the number of hidden layers with an increase of two layers from half of the input dimensions to twice the input dimensions.

4. Results

Our models used historical observations at each station to execute a one-step forecast, predicting the next time step in a sequence. The time step is 1 week for the Scripps Pier and 2 weeks for NRE. Parameters for the best-performing RFR and SVR models and their overall (station-averaged) performance in modeling Chl-a concentration measured by RMSE are shown, respectively, in Tables 1 and 2, and the performance in binary event forecasts (modeling the occurrence of blooms) in Table 3. According to Tables 1–3, the RFR and SVR models had large RMSEs (>10 μg/L) in predicting Chl-a concentrations in NRE. Although their performance in false detection was high (with low *POFD*), they failed to capture any algal blooms with poor bias (*B*) and *PSS* scores.

Table 1. Parameter and overall (station-averaged) performance of baseline models (RFR).

Best Performance Parameters	Max Depth	Max Features	Min Samples Split	N Estimators	RMSE
NRE	20	sqrt [1]	2	300	11.0
Scripps Pier	20	sqrt	2	300	0.5

[1]. If there are n features originally, then $\sqrt{n}$ features will be considered for each best split by RFR.

Table 2. Parameter and overall (station-averaged) performance of baseline models (SVR).

Best Performance Parameters	C	Gamma	Kernel Type	RMSE
NRE	100	0.1	Gaussian	10.9
Scripps Pier	10	0.1	Gaussian	0.5

Table 3. Skill assessment for prediction of algal bloom occurrence using 940 pairs of predictions and observations in the training dataset for baseline models and WA models for all NRE stations (0–160). Skill metrics for binary predictions include the bias (*B*) and Pierce skill score (PSS). We also include *a*, the number of hits; *b*, false alarms; *c*, false negatives (misses); and *d*, correct negatives, probability of detection (*POD*) and probability of false detection (*POFD*).

Models	*a*	*b*	*c*	*d*	*POD*	*POFD*	*B*	*PSS*
RFR	0	0	31	909	0.0	0.0	0.0	0.0
SVR	0	0	31	909	0.0	0.0	0.0	0.0
WA-LSTM	9	5	16	910	0.4	0.005	0.6	0.4
WA-BPNN0	16	7	9	908	0.6	0.008	0.9	0.6
WA-BPNN	16	6	9	909	0.6	0.007	0.9	0.6

Figure 9 compares observations and model results from baseline models (RFR and SVR) and the neural network model of WA-LSTM for NRE. The WA-LSTM model utilized the top *P* features among the 30 features on each station. Here we presented the results of the best-performing model on each station with *P* shown in Table 4. The best models on each station did not necessarily use the same subset of features. For example, the "aggregated discharge" feature was more important for stations closer to the river mouth than the offshore stations, potentially related to the river discharge's dilution and advection effects.

Table 4. Performance skills (RMSE and R^2) and model parameters from baseline models and WA models for individual NRE stations (0–160) and the Scripps Pier station.

Site		0	20	30	50	60	70	100	120	140	160	Scripps
RFR	RMSE	2.24	5.79	23.14	15.02	10.14	9.97	11.23	22.95	14.38	8.97	0.51
SVR	RMSE	3.56	6.17	23.33	17.02	10.09	11.86	12.80	24.54	15.05	9.70	0.53
WA-LSTM	RMSE	1.27	2.55	7.40	7.96	3.68	3.96	4.86	12.05	9.51	6.48	0.43
	R^2	0.70	0.70	0.75	0.70	0.63	0.79	0.67	0.48	0.41	0.53	0.70
	Neuron	59	53	53	59	29	53	35	59	40	59	32
	P	30/30	27/30	27/30	30/30	15/30	27/30	18/30	30/30	21/30	30/30	17/22
WA-BPNN0 [1]	RMSE	1.21	2.66	4.97	6.38	4.48	4.11	5.69	8.51	8.42	3.84	0.47
	R^2	0.72	0.70	0.89	0.81	0.69	0.78	0.65	0.74	0.57	0.83	0.65
	neuron	19	19	19	15	19	17	15	19	19	13	19
WA-BPNN [2]	RMSE	1.14	2.49	5.00	5.87	4.01	4.05	5.27	8.45	7.82	3.79	0.41
	R^2	0.76	0.73	0.89	0.84	0.75	0.78	0.70	0.75	0.63	0.83	0.73
	neuron	53	27	53	59	59	37	51	29	35	59	32
	P	30/30	24/30	30/30	30/30	30/30	27/30	27/30	18/30	27/30	30/30	17/22
Obs.	std	2.30	4.83	15.27	14.52	8.05	8.70	9.58	16.75	12.88	9.43	0.79

[1]. The BPNN models that used only the Chl-a feature from WA. [2]. The BPNN models that used feature selection among (a) the Chl-a feature from WA, (b) the other features from the original dataset (such as dissolved oxygen and ammonium) and (c) accumulative features generated (such as the accumulative air pressure from 3 days preceding the prediction date).

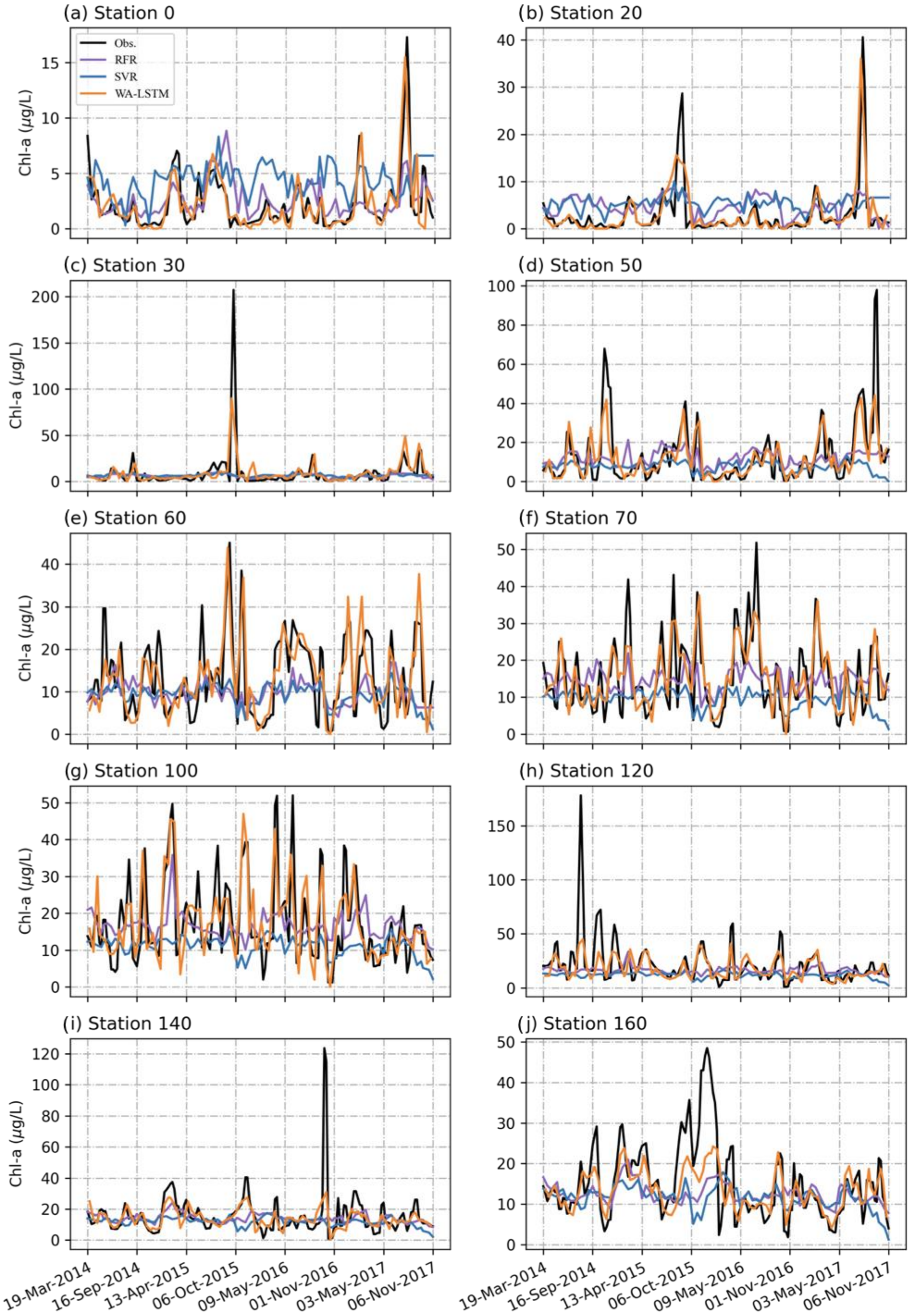

Figure 9. Comparison of the Chl-a (μg/L) prediction using RFR, SVR and WA-LSTM with in situ observations (Obs.) at NRE stations 0, 20, 30, 50, 60, 70, 100, 120, 140 and 160.

Figure 9 and Table 3 show that WA-LSTM can capture the occurrence of most algal blooms in NRE, especially for stations inshore of station 120, although it tends to underestimate the peak magnitude. More specifically, WA-LSTM predicted the occurrence of an algal bloom in October 2015 at station 30, with a peak concentration ~100 μg/L (137 μg/L on September 22 and 92 μg/L on 6 October). Although much smaller than the actual observation of 207 μg/L, it captured the occurrence of the bloom event. In contrast, RFR and SVR models predicted much smaller fluctuations in Chl-a and failed to capture any major bloom. To quantitatively compare the model's performance, we calculated the RMSEs and R^2 for all the models (Table 4). The RMSEs for WA-LSTM were 43–69% smaller than for RFR and 51–69% smaller than for SVR for stations inshore of 120. For the most downstream stations 120 and 160, WA-LSTM also underestimated the magnitude of major blooms but still captured the bloom timing and performed better than RFR and SVR with larger bias and *PSS* (Table 3). The R^2 for RFR and SVR ranged from −0.88 to 0.26 (not shown in the table), significantly smaller than for WA-LSTM.

Figure 10 shows comparisons between WA-LSTM and WA-BPNN with enriched features including accumulative wind, air pressure and river discharge. WA-BPNN better performed at predicting the bloom magnitude than WA-LSTM, even for station 120 and its offshore stations (Figure 10h–j). The average RMSE for WA-BPNN was 13% smaller than for WA-LSTM, with R^2 increasing by 0.12. WA-BPNN also captured more blooms than WA-LSTM with larger *B* and *PSS* (Table 3).

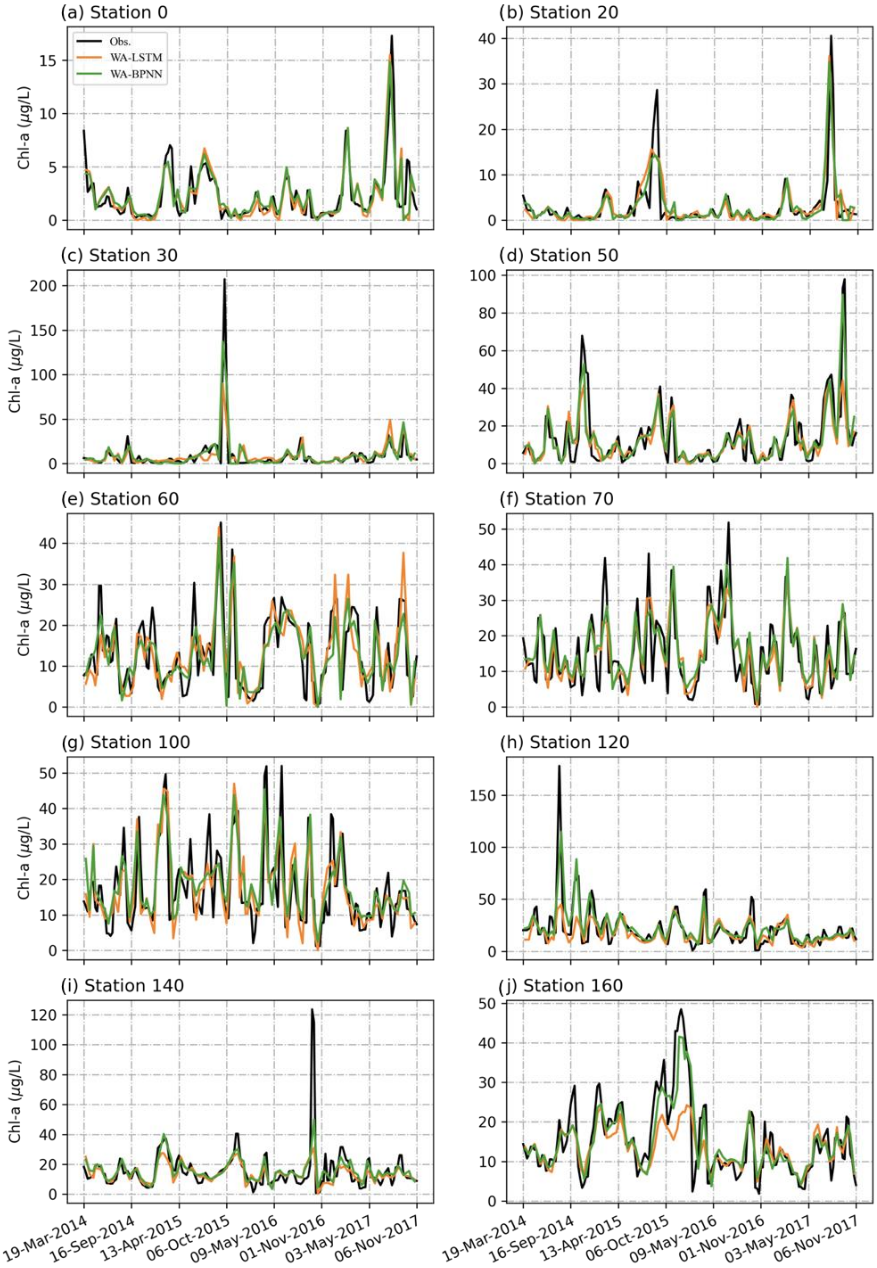

Figure 10. Comparison of the Chl-a (μg/L) prediction using WA-LSTM and WA-BPNN with in situ observations at NRE stations 0, 20, 30, 50, 60, 70, 100, 120, 140 and 160.

For the WA-BPNN results shown in Figure 10, we used the same enriched dataset, including discharge, winds and air temperature. We know that solely using the features from WA may also produce decent results, following the method in [60]. However, to examine the contribution of the additional nutrients and other environmental features, we compared the performance of models that consider all the possible features (WA-BPNN) versus the models that take only Chl-a wavelets analysis results as the model input (referred to as WA-BPNN0 below); then, we used the results from the best-so-far BPNN model for this comparison. We also tuned the models for the best results by trying different numbers of hidden layers of neurons. The performance of such models and the final selected parameters are shown in Table 4.

The results demonstrate that adding more potentially useful features will improve the performance of neural network models. By adding additional features, the RMSE dropped by 5.5%, and the R^2 increased by 0.03 on average compared with the models that used only the features from WA. Although the increase in R^2 is small, it is consistent throughout the stations. The performance in binary event forecasts using WA-BPNN0 and WA-BPNN was comparable (Table 3).

We applied the same pipeline to the Scripps Pier Chl-a prediction and assessed the models' performance in predicating Chl-a concentration. The performance in binary event forecasts was not evaluated because it is more traditional to use a Pseudo-nitzschia cell count as a bloom indicator in the Pacific Northwest [61]. The results showed a smaller RMSE using WA-LSTM than RFR and SVR, and an even smaller RMSE and larger R^2 using WA-BPNN with the enriched dataset. Only the WA-LSTM and WA-BPNN models captured the significant bloom event that occurred in late May to early June where Chl-a increased to about 5 times the time series average (Figure 11).

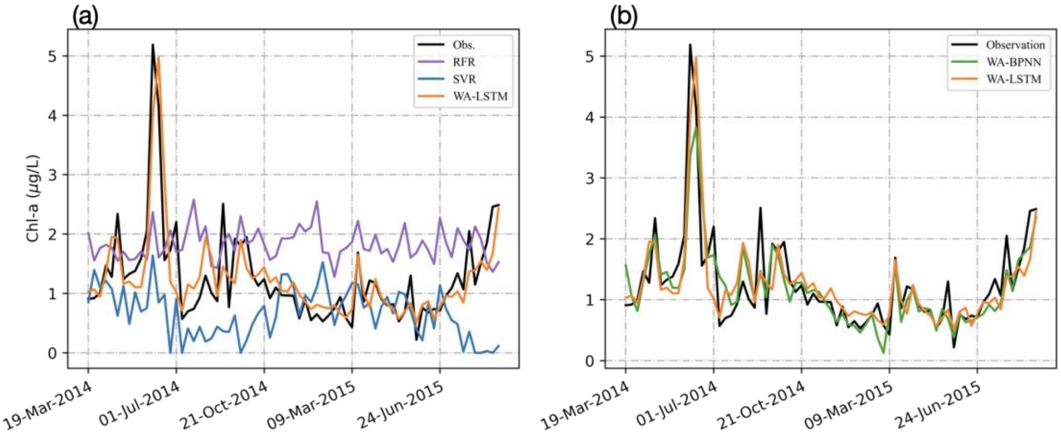

Figure 11. Comparison of the Chl-a (μg/L) prediction using (**a**) RFR, SVR and WA-LSTM and using (**b**) WA-BPNN and WA-LSTM with observations at Scripps Pier station.

5. Discussion and Conclusions

Predicting chlorophyll is critical to water quality management across aquatic systems, as chlorophyll has been used to indicate algal blooms and the state of water quality in coastal and estuarine waters. In addition, the enhanced ability to predict blooms will also help scientists to understand how blooms develop. Knowing that a bloom is about to happen, researchers can adjust their sampling plan to better understand the drivers of bloom formation, considering that the development of a bloom may destroy the conditions that spawn it. In this study, we created a pipeline to predict water quality using four machine learning models that are commonly used in water quality prediction. The pipeline includes two benchmark models, SVR and RFR and another two models incorporating WA, which are WA-LSTM and WA-BPNN. The same pipeline was applied to two bloom-prone locations, NRE, NC—where machine learning models were applied for short-term (2-week forward) algal bloom forecasts at single stations for the first time—and Scripps Pier, CA.

We found that while SVR and RFR failed to capture bloom events at most NRE stations, WA-LSTM and WA-BPNN successfully predicted blooms' occurrence and magnitude, especially for stations closer to the river mouth for the NRE area. The pipeline can also find the best results by trying different numbers of hidden layers neurons. Through the experiments, we found that the addition of meteorological and/or river flow data—determined by the environmental features of the study area—in addition to water quality data is a necessary step to further improve chlorophyll prediction.

Different variables were included in the dataset for each area, which was determined by data availability and environmental features of the study area. While both datasets included nutrients and meteorological data, only the NRE dataset included river discharge because of its direct influence by the Neuse River. Since the stations closer to the river mouth are more susceptible to river discharge, we also included distance from the river mouth station as an input feature for the NRE dataset. Although minimum tuning was needed for the switch from the NRE Chl-a forecast to Scripps Pier forecast, the WA-BPNN model works well for both regions, with an R^2 of 0.73 for the Scripps Pier and 0.77 for the NRE.

Considering that the pipeline works well for chlorophyll prediction for two distinct regions with different hydrodynamics and geometry, the pipeline can be potentially applied to other aquatic systems. Our experience with the NRE and Scripps Pier demonstrates that when applying the pipeline to a different aquatic system, in addition to water quality features that are considered essential growth limiting factors for phytoplankton, the inclusion of dominant physical features may improve the model's accuracy. We provide the source code of our pipeline on GitHub (https://github.com/QianqianLiu/WaterQuality (accessed on 15 August 2023)) to support future applications in other areas.

Author Contributions: Conceptualization, Q.L. and Y.S.; methodology, Y.S.; software, J.L. (Junan Lin), J.L. (Jiting Liu) and Y.Y.; validation, J.L. (Junan Lin), J.L. (Jiting Liu) and Y.Y.; formal analysis, J.L. (Junan Lin), J.L. (Jiting Liu) and Y.Y.; investigation, Q.L., Y.S., J.L. (Junan Lin), J.L. (Jiting Liu) and Y.Y.; resources, Q.L. and Y.S.; data curation, N.S.H. and Q.L.; writing—original draft preparation, J.L. (Junan Lin), Q.L., J.L. (Jiting Liu) and Y.Y.; writing—review and editing, Y.S. and N.S.H.; visualization, J.L. (Junan Lin), J.L. (Jiting Liu) and Y.Y.; supervision, Q.L.; project administration, Q.L.; funding acquisition, Q.L. and Y.S. All authors have read and agreed to the published version of the manuscript.

Funding: Liu and Song were funded by National Science Foundation Office of Advance Cyberinfrastructure 2230046. Liu was also funded by the award to the University of North Carolina at Wilmington through North Carolina Sea Grant R/22-SFA-3.

Data Availability Statement: The code and data used in this paper are available on GitHub (https://github.com/QianqianLiu/WaterQuality (accessed on 15 August 2023)). Data collected by the Neuse River Modeling and Monitoring (ModMon) program can be accessed through the Southeast Coastal Ocean Observing Regional Association's data portal at https://portal.secoora.org/ (accessed on 12 January 2023).

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

Appendix A

Table A1. ModMon features in the NRE dataset.

Feature	Unit
date	/
depth level of measurement	/
distance from Station 0	km
water temperature	°C
salinity	ppt
dissolved oxygen (DO)	%
pH	/
particulate organic carbon (POC)	µg/L
nitrate/nitrite (NO_3/NO_2)	µg/L
ammonium (NH_4)	µg/L
orthophosphate (PO_4)	µg/L

Table A2. Time interval of samplings in Station 0 of the ModMon dataset in days.

Interval of Sampling (Days)	Sampling Number	Proportion
5~11	13	4.64%
12	20	7.14%
13	46	16.43%
14	56	20.00%
15	37	13.21%
16	17	6.07%
>=17	91	32.50%
In total	280	100%

Table A3. Time intervals of samplings for the Scripps Pier dataset in days. This dataset has a relatively uniform sampling interval.

Interval of Sampling (Days)	Sampling Number	Proportion
6	29	8.06%
7	302	83.89%
8	28	7.78%
9	1	0.28%
In total	360	100%

Table A4. Number of original (before-interpolation) observations used in training and testing dataset for NRE and Scripps pier machine learning models.

Dataset	Station	Number of Obs. in Training Data	Number of Obs. in Testing Data	Total
NRE	0	342	79	421
	20	342	79	421
	30	358	80	438
	50	358	80	438
	60	356	80	436
	70	358	80	438
	100	356	80	436
	120	358	80	438
	140	350	80	430
	160	351	80	431
Scripps pier		300	61	361

References

1. Obenour, D.R.; Gronewold, A.D.; Stow, C.A.; Scavia, D. Using a Bayesian hierarchical model to improve Lake Erie cyanobacteria bloom forecasts. *Water Resour. Res.* **2014**, *50*, 7847–7860. [CrossRef]
2. Rowe, M.D.; Anderson, E.J.; Wynne, T.T.; Stumpf, R.P.; Fanslow, D.L.; Kijanka, K.; Vanderploeg, H.A.; Strickler, J.R.; Davis, T.W. Vertical distribution of buoyant Microcystis blooms in a Lagrangian particle tracking model for short-term forecasts in Lake Erie. *J. Geophys. Res. Oceans* **2016**, *175*, 238. [CrossRef]
3. Stumpf, R.P.; Wynne, T.T.; Baker, D.B.; Fahnenstiel, G.L. Interannual variability of cyanobacterial blooms in Lake Erie. *PLoS ONE* **2012**, *7*, 42444. [CrossRef]
4. Stumpf, R.P.; Davis, T.W.; Wynne, T.T.; Graham, J.L.; Loftin, K.A.; Johengen, T.H.; Gossiaux, D.; Palladino, D.; Burtner, A. Challenges for mapping cyanotoxin patterns from remote sensing of cyanobacteria. *Harmful Algae* **2016**, *54*, 160–173. [CrossRef] [PubMed]
5. Caron, D.A.; Garneau, M.-È.; Seubert, E.; Howard, M.D.A.; Darjany, L.; Schnetzer, A.; Cetinić, I.; Filteau, G.; Lauri, P.; Jones, B. Harmful algae and their potential impacts on desalination operations off southern California. *Water Res.* **2010**, *44*, 385–416. [CrossRef] [PubMed]
6. Lewitus, A.J.; Horner, R.A.; Caron, D.A.; Garcia-Mendoza, E.; Hickey, B.M.; Hunter, M.; Huppert, D.D.; Kudela, R.M.; Langlois, G.W.; Largier, J.L.; et al. Harmful algal blooms along the North American west coast region: History, trends, causes, and impacts. *Harmful Algae* **2012**, *19*, 133–159. [CrossRef]
7. McGowan, J.A.; Deyle, E.R.; Ye, H.; Carter, M.L.; Perretti, C.T.; Seger, K.D.; Verneil, A.; Sugihara, G. Predicting coastal algal blooms in southern California. *Ecology* **2017**, *98*, 1419–1433. [CrossRef]

8. Liu, Q.; Rowe, M.D.; Anderson, E.J.; Stow, C.A.; Stumpf, R.P.; Johengen, T.H. Probabilistic forecast of microcystin toxin using satellite remote sensing, in situ observations and numerical modeling. *Environ. Model. Softw.* **2020**, *128*, 104705. [CrossRef]
9. Powell, T.M.; Lewis, C.V.W.; Curchitser, E.N.; Haidvogel, D.B.; Hermann, A.J.; Dobbins, E.L. Results from a three-dimensional, nested biological-physical model of the California Current System and comparisons with statistics from satellite imagery. *J. Geophys. Res.* **2006**, *111*, C07018. [CrossRef]
10. Fennel, K.; Wilkin, J.; Levin, J.; Moisan, J.; O'Reilly, J.; Haidvogel, D. Nitrogen cycling in the Middle Atlantic Bight: Results from a three-dimensional model and implications for the North Atlantic nitrogen budget. *Glob. Biogeochem. Cycles* **2006**, *20*. [CrossRef]
11. Fennel, K.; Gehlen, M.; Brasseur, P.; Brown, C.W.; Ciavatta, S.; Cossarini, G.; Crise, A.; Edwards, C.A.; Ford, D.; Friedrichs, M.A.M.; et al. Advancing Marine Biogeochemical and Ecosystem Reanalyses and Forecasts as Tools for Monitoring and Managing Ecosystem Health. *Front. Mar. Sci.* **2019**, *6*, unsp 89. [CrossRef]
12. Faugeras, B.; Bernard, O.; Sciandra, A.; Lévy, M. A mechanistic modelling and data assimilation approach to estimate the carbon/chlorophyll and carbon/nitrogen ratios in a coupled hydrodynamical-biological model. *Nonlinear Process. Geophys.* **2004**, *11*, 515–533. [CrossRef]
13. Anderson, C.R.; Sapiano, M.R.P.; Prasad, M.B.K.; Long, W.; Tango, P.J.; Brown, C.W.; Murtugudde, R. Predicting potentially toxigenic Pseudo-nitzschia blooms in the Chesapeake Bay. *J. Mar. Syst.* **2010**, *83*, 127–140. [CrossRef]
14. Yin, S.; Ding, S.X.; Xie, X.; Luo, H. A Review on Basic Data-Driven Approaches for Industrial Process Monitoring. *IEEE Trans. Ind. Electron.* **2014**, *61*, 6418–6428. [CrossRef]
15. Jin, D.; Lee, E.; Kwon, K.; Kim, T. A Deep Learning Model Using Satellite Ocean Color and Hydrodynamic Model to Estimate Chlorophyll-a Concentration. *Remote Sens.* **2021**, *13*, 2003. [CrossRef]
16. Yu, X.; Shen, J. A data-driven approach to simulate the spatiotemporal variations of chlorophyll-a in Chesapeake Bay. *Ocean Model.* **2021**, *159*, 101748. [CrossRef]
17. Chai, F.; Dugdale, R.C.; Peng, T.-H.; Wilkerson, F.P.; Barber, R.T. One-dimensional ecosystem model of the equatorial Pacific upwelling system. Part I: Model development and silicon and nitrogen cycle. *Deep Sea Res. Part II Top. Stud. Oceanogr.* **2002**, *49*, 2713–2745. [CrossRef]
18. Liu, Q.; Anderson, E.J.; Zhang, Y.; Weinke, A.D.; Knapp, K.L.; Biddanda, B.A. Modeling reveals the role of coastal upwelling and hydrologic inputs on biologically distinct water exchanges in a Great Lakes estuary. *Estuar. Coast. Shelf Sci.* **2018**, *209*, 41–55. [CrossRef]
19. Beck, M.B. Water quality modeling: A review of the analysis of uncertainty. *Water Resour. Res.* **1987**, *23*, 1393–1442. [CrossRef]
20. Wang, Z.; Chai, F.; Xue, H.; Wang, X.H.; Zhang, Y.J.; Dugdale, R.; Wilkerson, F. Light Regulation of Phytoplankton Growth in San Francisco Bay Studied Using a 3D Sediment Transport Model. *Front. Mar. Sci.* **2021**, *8*, 633707. [CrossRef]
21. Yu, X.; Shen, J.; Du, J. A Machine-Learning-Based Model for Water Quality in Coastal Waters, Taking Dissolved Oxygen and Hypoxia in Chesapeake Bay as an Example. *Water Resour. Res.* **2020**, *56*, e2020WR027227. [CrossRef]
22. Yu, P.; Gao, R.; Zhang, D.; Liu, Z.-P. Predicting coastal algal blooms with environmental factors by machine learning methods. *Ecol. Indic.* **2021**, *123*, 107334. [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. Cloern, J.E. The relative importance of light and nutrient limitation of phytoplankton growth: A simple index of coastal ecosystem sensitivity to nutrient enrichment. *Aquat. Ecol.* **1999**, *33*, 3–15. [CrossRef]
25. Burkholder, J.; Eggleston, D.; Glasgow, H.; Brownie, C.; Reed, R.; Janowitz, G.; Posey, M.; Melia, G.; Kinder, C.; Corbett, R.; et al. Comparative impacts of two major hurricane seasons on the Neuse River and western Pamlico Sound ecosystems. *Proc. Natl. Acad. Sci. USA* **2004**, *101*, 9291–9296. [CrossRef]
26. *North Carolina Coastal Habitat Prediction Plan*; North Carolina Division of Marine Fisheries: Morehead City, NC, USA, 2010.
27. Van, G.; Christina, H.; Winkle, V.; O'neil, M.; Matthews, K.; Sinha, P. *Economic Valuation of the Albemarle-Pamlico Watershed's Natural Resources Final Report*; RTI International: Research Triangle Park, NC, USA, 2016.
28. Paerl, H.; Pinckney, J.; Fear, J.; Peierls, B. Ecosystem responses to internal and watershed organic matter loading:consequences for hypoxia in the eutrophying Neuse River Estuary, North Carolina, USA. *Mar. Ecol. Prog. Ser.* **1998**, *166*, 17–25. [CrossRef]
29. Wool, T.A.; Davie, S.R.; Rodriguez, H.N. Development of Three-Dimensional Hydrodynamic and Water Quality Models to Support Total Maximum Daily Load Decision Process for the Neuse River Estuary, North Carolina. *J. Water Resour. Plan. Manag.* **2003**, *129*, 295–306. [CrossRef]
30. Katin, A.; Del Giudice, D.; Obenour, D.R. Modeling biophysical controls on hypoxia in a shallow estuary using a Bayesian mechanistic approach. *Environ. Model. Softw.* **2019**, *120*, 104491. [CrossRef]
31. Paerl, H.W.; Rossignol, K.L.; Hall, S.N.; Peierls, B.L.; Wetz, M.S. Phytoplankton Community Indicators of Short- and Long-term Ecological Change in the Anthropogenically and Climatically Impacted Neuse River Estuary, North Carolina, USA. *Estuaries Coasts* **2010**, *33*, 485–497. [CrossRef]
32. Anderson, C.R.; Kudela, R.M.; Kahru, M.; Chao, Y.; Rosenfeld, L.K.; Bahr, F.L.; Anderson, D.M.; Norris, T.A. Initial skill assessment of the California Harmful Algae Risk Mapping (C-HARM) system. *Harmful Algae* **2016**, *59*, 1–18. [CrossRef]
33. Kim, H.-J.; Miller, A.J.; McGowan, J.; Carter, M.L. Coastal phytoplankton blooms in the Southern California Bight. *Prog. Oceanogr.* **2009**, *82*, 137–147. [CrossRef]

34. Li, Z.; Peng, F.; Niu, B.; Li, G.; Wu, J.; Miao, Z. Water Quality Prediction Model Combining Sparse Auto-encoder and LSTM Network. *IFAC-Pap.* **2018**, *51*, 831–836. [CrossRef]
35. Shi, B.; Wang, P.; Jiang, J.; Liu, R. Applying high-frequency surrogate measurements and a wavelet-ANN model to provide early warnings of rapid surface water quality anomalies. *Sci. Total Environ.* **2018**, *610–611*, 1390–1399. [CrossRef]
36. Xu, L.; Liu, S. Study of short-term water quality prediction model based on wavelet neural network. *Math. Comput. Model.* **2013**, *58*, 807–813. [CrossRef]
37. McCulloch, W.S.; Pitts, W. A logical calculus of the ideas immanent in nervous activity. *Bull. Math. Biophys.* **1943**, *5*, 115–133. [CrossRef]
38. Cortes, C.; Vapnik, V. Support-vector networks. *Mach. Learn.* **1995**, *20*, 273–297. [CrossRef]
39. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
40. Biau, G.; Scornet, E. A random forest guided tour. *TEST* **2016**, *25*, 197–227. [CrossRef]
41. Liaw, A.; Wiener, M. Classification and regression by randomForest. *R News* **2002**, *2*, 18–22.
42. Sifuzzaman, M.; Islam, M.R.; Ali, M.Z. Application of Wavelet Transform and its Advantages Compared to Fourier Transform. *J. Phys. Sci.* **2009**, *13*, 121–134.
43. Ghaderpour, E.; Pagiatakis, S.D.; Hassan, Q.K. A Survey on Change Detection and Time Series Analysis with Applications. *Appl. Sci.* **2021**, *11*, 6141. [CrossRef]
44. Akansu, A.N.; Serdijn, W.A.; Selesnick, I.W. Emerging applications of wavelets: A review. *Phys. Commun.* **2010**, *3*, 1–18. [CrossRef]
45. Cohen, A. Ten Lectures on Wavelets, CBMS-NSF Regional Conference Series in Applied Mathematics, Vol. 61, I. Daubechies, SIAM, 1992, xix + 357 pp. *J. Approx. Theory* **1994**, *78*, 460–461. [CrossRef]
46. Yang, M.; Sang, Y.-F.; Liu, C.; Wang, Z. Discussion on the Choice of Decomposition Level for Wavelet Based Hydrological Time Series Modeling. *Water* **2016**, *8*, 197. [CrossRef]
47. Hecht-Nielsen, R. Theory of the Backpropagation Neural Network. In *Neural Networks for Perception*; Elsevier: Amsterdam, The Netherlands, 1992; pp. 65–93. [CrossRef]
48. Siami-Namini, S.; Tavakoli, N.; Siami Namin, A. A Comparison of ARIMA and LSTM in Forecasting Time Series. In Proceedings of the 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA), Orlando, FL, USA, 17–20 December 2018; pp. 1394–1401. [CrossRef]
49. Wang, Y.; Zhou, J.; Chen, K.; Wang, Y.; Liu, L. Water quality prediction method based on LSTM neural network. In Proceedings of the 2017 12th International Conference on Intelligent Systems and Knowledge Engineering (ISKE), Nanjing, China, 24–26 November 2017; pp. 1–5. [CrossRef]
50. Sokolova, E.; Ivarsson, O.; Lillieström, A.; Speicher, N.K.; Rydberg, H.; Bondelind, M. Data-driven models for predicting microbial water quality in the drinking water source using E. coli monitoring and hydrometeorological data. *Sci. Total Environ.* **2022**, *802*, 149798. [CrossRef]
51. Tornevi, A.; Bergstedt, O.; Forsberg, B. Precipitation Effects on Microbial Pollution in a River: Lag Structures and Seasonal Effect Modification. *PLoS ONE* **2014**, *9*, e98546. [CrossRef]
52. Hejazi, M.I.; Cai, X. Input variable selection for water resources systems using a modified minimum redundancy maximum relevance (mMRMR) algorithm. *Adv. Water Resour.* **2009**, *32*, 582–593. [CrossRef]
53. Yang, Y.; Pedersen, J. A Comparative Study on Feature Selection in Text Categorization. *Icml* **1997**, *97*, 35.
54. Aksu, G.; Güzeller, C.O.; Eser, M.T. The Effect of the Normalization Method Used in Different Sample Sizes on the Success of Artificial Neural Network Model. *Int. J. Assess. Tools Educ.* **2019**, *6*, 170–192. [CrossRef]
55. Hogan, R.J.; Mason, I.B. Deterministic Forecasts of Binary Events. In *Forecast Verification*; Jolliffe, I.T., Stephenson, D.B., Eds.; Wiley: Hoboken, NJ, USA, 2011; pp. 31–59. [CrossRef]
56. North Carolina Administrative Code. *Classification and Water Quality Standards Applicable to Surface Waters and Wetlands of North Carolina. Raleigh, North Carolina. 2022*; 15A NCAC 02B.0200; North Carolina Department of Environmental Quality: Raleigh, NC, USA, 2022.
57. Chan, S.; Treleaven, P. Continuous Model Selection for Large-Scale Recommender Systems. In *Handbook of Statistics*; Elsevier: Amsterdam, The Netherlands, 2015; Volume 33, pp. 107–124. [CrossRef]
58. Bergstra, J.; Bengio, Y. Random Search for Hyper-Parameter Optimization. *J. Mach. Learn. Res.* **2012**, *13*, 281–305.
59. Daubechies, I. *Ten Lectures on Wavelets*; SIAM: Philadelphia, PA, USA, 1992.
60. Ravansalar, M.; Rajaee, T.; Ergil, M. Prediction of dissolved oxygen in River Calder by noise elimination time series using wavelet transform. *J. Exp. Theor. Artif. Intell.* **2016**, *28*, 689–706. [CrossRef]
61. Stone, H.B.; Banas, N.S.; MacCready, P.; Trainer, V.L.; Ayres, D.L.; Hunter, M.V. Assessing a model of Pacific Northwest harmful algal bloom transport as a decision-support tool. *Harmful Algae* **2022**, *119*, 102334. [CrossRef] [PubMed]

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.