Water Research 186 (2020) 116349

Contents lists available at ScienceDirect

Water Research

journal homepage: www.elsevier.com/locate/watres

Using convolutional neural network for predicting cyanobacteria concentrations in river water

JongCheol Pyo[a], Lan Joo Park[b], Yakov Pachepsky[c], Sang-Soo Baek[a], Kyunghyun Kim[d,**], Kyung Hwa Cho[a,*]

[a] School of Urban and Environmental Engineering, Ulsan National Institute of Science and Technology, 50 UNIST-gil, Eonyang-eup, Ulju-gun, Ulsan 689-798, Republic of Korea
[b] Water Quality Assessment Research Division, National Institute of Environmental Research, Hwangyeong-ro 42, Seogu, Incheon 22689, Republic of Korea
[c] Environmental Microbial and Food Safety Laboratory, USDA-ARS, Beltsville, MD, USA
[d] Watershed and Total Load Management Research Division, National Institute of Environmental Research, Hwangyeong-ro 42, Seogu, Incheon 22689, Republic of Korea

ARTICLE INFO

Article history:
Received 6 February 2020
Revised 14 July 2020
Accepted 26 August 2020
Available online 26 August 2020

Keywords:
Convolutional neural network
EFDC
Synthetic data
Microcystis
Prediction

ABSTRACT

Machine learning modeling techniques have emerged as a potential means for predicting algal blooms. In this study, synthetic spatio-temporal water quality data for a river section were generated with a 3D water quality model and used to investigate the capability of a convolutional neural network (CNN) for predicting harmful cyanobacterial blooms. The CNN model displayed a reasonable capacity for short-term predictions of cyanobacteria (Microcystis) biomass. In the nowcasting of Microcystis, the CNN performance had a Nash-Sutcliffe Efficiency (NSE) of 0.87. An increase in the forecast lead time resulted in a decrease in the prediction accuracy, reducing the NSE from 0.87 to 0.58. As the spatial observation density increased from 20% to 100% of the input image grids, the CNN prediction NSE had improved from 0.70 to 0.84. Adding noise to the data resulted in accuracy deterioration, but even at the noise amplitude of 10%, the accuracy was acceptable for some applications, with NSE = 0.76. Visualization of the CNN results characterized its performance variations across the studied river reach. Overall, this study successfully demonstrated the capability of the CNN model for cyanobacterial bloom prediction using high temporal frequency images.

1. Introduction

Predicting water quality changes is a challenging issue in environmental research due to the complexity and uncertainty of biochemical processes in natural waters (Beck et al., 2012; Summers et al., 1993). In particular, the accurate prediction of algal dynamics is complicated due to substantial uncertainties associated with growth, nutrient uptake, grazing, and photosynthesis, among other factors (Li et al., 2013; Xie et al., 2012).

With the rapid development of computational power, numerous efforts have been made to transition water quality models from lumped steady-state models to three-dimensional dynamic models (Wang et al., 2015). Especially, models have been improved continuously for algae-related predictions. Clark and Jaworski (1972) utilized a one-dimensional dynamic estuary model to predict algae concentrations. However, the one-dimensional approach ignored significant physical phenomena, such as substantial vertical differences. Additionally, the prediction accuracy was hampered by the use of the simplified equations that averaged the vertical and lateral variations (Ulanowicz, 1976). Later studies developed two-dimensional hydrodynamic simulations to consider longitudinal and vertical advection. Martin (1988) applied a two-dimensional water quality model coupled with hydrodynamic algorithms to predict algal dynamics. However, the two-dimensional model had limited applicability to water bodies with pronounced lateral differences. Recently, with the improvement in computational capability, three-dimensional hydrodynamic models have been actively utilized to predict algal concentrations. Wu and Xu (2011) and Kim et al. (2017) predicted algal blooms with reasonable accuracy by applying the Environmental Fluid Dynamics Code (EFDC). Although this model has been developed for predicting algal bloom trends when sufficient observed data is available, the prediction of the algal dynamics still requires improvements. Interactions among the major processes, uncertainty in the kinetic rate parameter val-

* Corresponding author.
** Co-corresponding author.
E-mail addresses: matthias@korea.kr (K. Kim), khcho@unist.ac.kr (K.H. Cho).

https://doi.org/10.1016/j.watres.2020.116349
0043-1354/

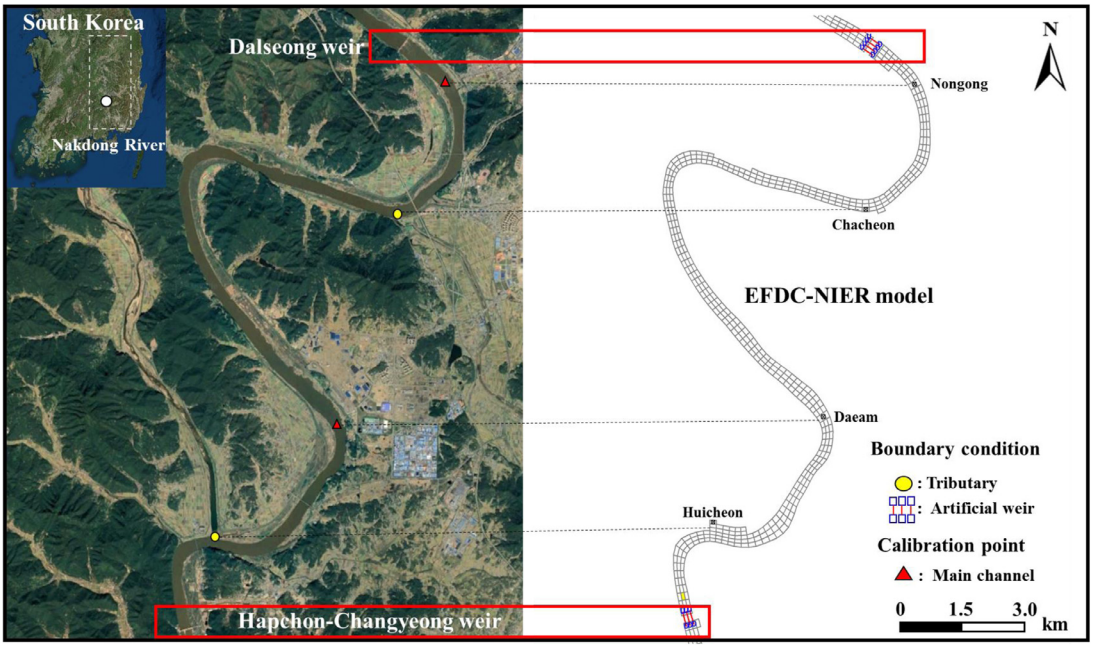

Fig. 1. Hapchon-Changyeong weir region and EFDC–NIER setup. Two tributaries (Huicheon and Chacheon) and two weirs (Hapchon-Changyeong weir and Dalseong weir) were assigned as boundary condition of EFDC–NIER model. Two points (Daeam and Nongong) in main channel of the weir were utilized as model calibration points.

ues, and complexity of the three-dimensional hydrodynamic simulation hinder the efficient application of the model for forecasting algal blooms (Li et al., 2013; Xie et al., 2012).

Data-driven models have been spotlighted as an alternative approach to estimating water quality parameters by reflecting the underlying features of water constituent dynamics. Teles et al. (2006) effectively implemented a time series forecasting of cyanobacteria using artificial neural networks (ANN). Cho et al. (2011) successfully predicted arsenic concentrations in groundwater using ANNs, and Park et al. (2015) showed satisfactory performance of support vector machines (SVM) in chlorophyll-a predictions. Conventional data-driven models with relatively low-dimensional data, such as the time series of point-measured constituents, have shown reasonable accuracy. However, such models contain a large number of internal units, internal parameters, and prohibitively high training times for accurate predictions using high-dimensional data, such as multi-dimensional imagery, that have become an essential output of modern monitoring programs (Wójcik and Kurdziel, 2019).

Deep learning techniques such as convolutional neural networks (CNNs) appear to be suitable to process the multi-dimensional imagery for classification and regression tasks because CNNs can use the high-dimensional images as inputs and extract sophisticated features in the imagery data, thus radically increasing the explanatory and predictive capacity of the neural network (Chen et al., 2016; Yu et al., 2017). Therefore, CNNs can be utilized to estimate algal concentrations with multi-dimensional data. Pyo et al. (2019) has introduced a CNN regression with hyperspectral imagery for estimating cyanobacteria biomass.

The applicability of CNNs can be expanded to the prediction tasks when images are available with high temporal frequency (Poliyapram et al., 2019). To date, CNN predictions for algal blooms have been rarely implemented because it is challenging to obtain imagery data with high temporal frequency from airborne platforms or satellites due to high costs and variable weather conditions (Yang et al., 2010). Nevertheless, the algal prediction performance of the CNN model can be evaluated using synthetic datasets obtained as outputs of hydrodynamics-based water quality models, which can generate high temporal frequency imagery. Additionally, a CNN trained with the hydrodynamics-based model outputs can become a surrogate model or metamodel (Hong et al., 2017). This model enables sensitivity and scenario analyses that are hard to implement by running the hydrodynamics-based models when there are a large number of scenarios.

This study aims to evaluate the applicability of CNNs for algal bloom predictions with model-generated high temporal frequency image (gridded) data and investigate the performance mechanisms through scenario analyses, which we believe have never been addressed previously. We used a calibrated EFDC to generate model output images of the concentrations of Microcystis in grid cells for a river section. These images were used to develop the CNN models applied for the scenario analyses. The differences between the scenarios included variances in the forecast lead-time, spatial observation density, and amplitude of noise added to the input images.

2. Materials and methods

2.1. Study area

The reach of the Nakdong River, the longest river in South Korea, was selected for this study. Two large multi-purpose weirs, the Dalseong (DS) and Hapcheon-Changyeong (HC) weirs, define the upstream and downstream boundaries of the reach (referred to as the HC reach), which has the average water depth of 8.4 m and the length of 29.4 km (Fig. 1).

Two main tributaries, the Huichan and Chacheon, bring pollutants from domestic and industrial sewage treatment plants, livestock facilities, and agricultural fields to the reach (Lee and Kim, 2017). These point and non-point source pollutants, including those from the upstream boundary, cause eutrophication in the HC reach, with several algal blooms (Lee et al., 2014). Additionally, the weir increases the flow retention time and stability, enhancing harmful cyanobacterial blooms in the summer (Lee et al., 2014). Cyanobacterial blooms were observed in the reach a total of 129 days during the four years from 2012 to 2016. Furthermore, the cell counts of cyanobacteria during 61% of the bloom period (i.e.

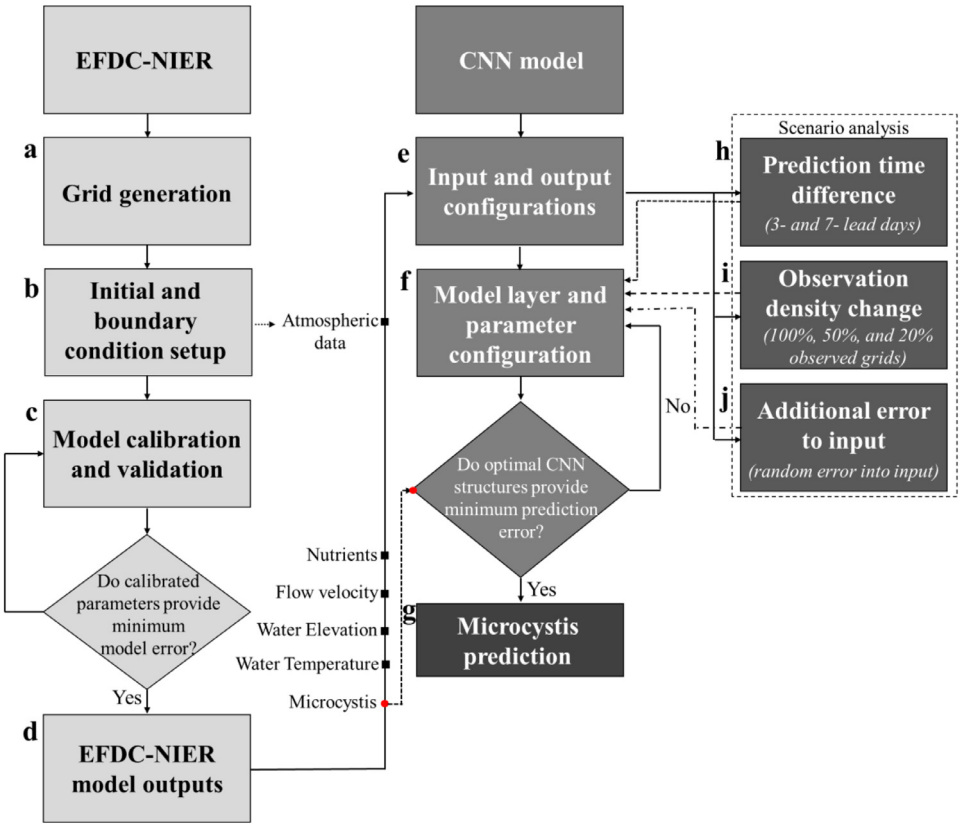

Fig. 2. Flowchart for predicting Microcystis concentration using CNN model; a-d is for EFDC–NIER model simulation including model construction (a-b), model evaluation (c), and model prediction (d); e-g is for CNN model operation including input data configuration from EFDC–NIER outputs (e), internal structure and parameter configuration (f), and CNN model prediction (g); h-j is for the scenario including 3- an 7-day lead prediction (h), observed grid variation of 100%, 50%, and 20% (i), and uniform, normal, and gamma distribution error adding into the CNN inputs (j).

79 days) exceeded the warning criteria set for early algal alerts in South Korea (1000 cells·mL^{-1}). (Lee and Kim, 2017).

2.2. EFDC model

The layout of this study is presented in Fig. 2. The first part contains simulations by the hydrodynamic model (Fig 2a–d). This study selected the EFDC model, which is a three-dimensional hydrodynamic and water quality model developed at the Virginia Institute of Marine Science and approved by the US EPA (Hamrick, 1992). The EFDC model includes hydrodynamics, water quality, sediment transport, and toxic contaminant transport and fate modules, which are capable of simulating flow, transport and fate, and biochemical processes in rivers, lakes, estuaries, and coastal areas. The hydrodynamic module solves the three-dimensional shallow water equations coupled with salinity and temperature transport (Hamrick and Mills, 2000). Additionally, the EFDC solves the vertically hydrostatic and free-surface mean turbulent fluid motion equation.

The specific mass balance equations reflect the changes in the concentrations of water quality constituents such as algal groups, nitrogen, phosphorus, dissolved oxygen, silica, and carbon:

$$\frac{\partial}{\partial t}(m_x m_y HC) + \frac{\partial}{\partial x}(m_y HuC) + \frac{\partial}{\partial y}(m_x HvC) + \frac{\partial}{\partial z}(m_x m_y wC) = \frac{\partial}{\partial x}\left(\frac{m_y HA_x}{m_x}\frac{\partial C}{\partial x}\right) + \frac{\partial}{\partial y}\left(\frac{m_x HA_y}{m_y}\frac{\partial C}{\partial y}\right) + \frac{\partial}{\partial z}\left(\frac{m_x m_y A_z}{H}\frac{\partial C}{\partial z}\right) + m_x m_y HS_c \quad (1)$$

where u, v, and w are the velocities in the x-, y-, and z- directions, respectively; C is the concentration of the water-quality-related constituent; A_x, A_y, and A_z are the turbulent diffusivities in the x-, y-, and z- directions; H is the water depth; m_x and m_y are scale factors; and S_c includes internal and external sources and sinks per unit volume.

The water quality module can simulate three algal groups, including cyanobacteria, diatoms, and green algae. The equation of the module describes algal growth, basal metabolism, predation, settling, and external loading for the algal groups:

$$\frac{\partial B_x}{\partial t} = (P_x - BM_x - PR_x)B_x + \frac{\partial}{\partial z}(WS_x B_x) + \frac{WB_x}{V} \quad (2)$$

where B_x is algal biomass of each algal group x in carbon concentration (gC·m^{-3}); P_x is the production rate (day^{-1}); BM_x is the basal metabolism rate (day^{-1}); PR_x is the predation rate (day^{-1}); WS_x is the positive settling velocity (m·day^{-1}); WB_x is the external load (gC·day^{-1}); t is time (day); and V is the cell volume (m^3).

2.2.1. EFDC–NIER model

The original EFDC model has a limitation in the simulation of cyanobacteria because it does not mimic vertical cyanobacterial migration under buoyancy control, which can be critical in terms of competition with other algal species for light and nutrients. The National Institute of Environmental Research (NIER) in the Republic of Korea modified the EFDC to consider this functionality by allowing algal density to be changed depending on the amount of light the algae receive. The modified version, EFDC–NIER, provides additional capabilities such as simulations for several algal species. Different cyanobacteria species, such as the Microcystis and Anabaena species, can be simulated concurrently by the EFDC–NIER. Structures of the original control files (i.e., EFDC.INP and WQ3DWC.INP) were preserved, and new control files were

built for the initial conditions and parameters of the additional algal species. The cyanobacterial density and vertical movement velocity were adopted from Kromkamp and Walsby (1990), and the settling velocity module was corrected to consider the density and vertical velocity parameters in the input file, WQSTOKES.INP. In addition, the artificial weir module for EFDC–NIER was developed to simulate various weir structure operations, such as a fixed or movable weir, fishway, and small hydraulic power plant. The model can reflect the variation of flow rate and water elevation controlled by the HC. Details of the EFDC–NIER weir module are described in Shin et al. (2017).

2.2.1.1. Model setup and boundary condition. In the current study, the EFDC–NIER model included a section of the river that extended beyond the HC reach to obtain more accurate calibration results. The model domain includes most of the primary Nakdong River stream extending from the river mouth to the Andong Dam, which is located approximately 340 km upstream from the river mouth. Fig. 1 presents the sub-domain for the HC reach, where curvilinear orthogonal grids were adopted to represent the geometry of the HC reach, thereby generating a total of 480 grid cells (Figs. 1 and 2a). The grid cell width and lengths varied, ranging from 90 m to 100 m and from 190 m to 200 m, respectively. The water body was divided into 11 layers of the same thickness, with an average depth of 10 m.

After the grid generation, the initial and boundary conditions were set using flow and water quality data acquired from the Water Environmental Information System of the Republic of Korea, observed at the outlets of the two tributaries, and using the simulation results from the DS, the upstream boundary (Fig. 2b). The initial conditions were set using the observation data along the river reach, but their effects disappear soon after the spin-up period of the model (Fig. 2b). In the mainstream and tributaries of the HC, the inflow, outflow, water quality constituent loading, and weather data were used to set the initial concentrations and boundary conditions.

The model was calibrated for the entire year of 2015 and then validated for the years from 2016 to 2018. The significant issues for the water quality module calibration were the algal primary production, oxygen consumption after the decomposition of organic matter, and dissolved oxygen concentration (Wu and Xu, 2011). The optimum model parameters were varied based on the typical ranges found in literature sources (Fig. 2c) and were fine-tuned until the root mean square difference between the observed and simulated water quality variables was minimized (Chen et al., 2016).

2.3. Convolutional neural network (CNN)

A CNN is a deep learning model that exploits feature hierarchies of multi-dimensional image data, from low-level to high-level features, and serves as a powerful image processing tool. Compared to standard feedforward neural network, a CNN requires fewer parameters and connections between elements and is, therefore, easier to train (Krizhevsky et al., 2012). EFDC–NIER outputs can be organized into images where grid cells serve as pixels. Thus, this study utilized a CNN model with structural units shown in Fig. 3 to extract the water quality features from the EFDC–NIER output to predict Microcystis biomass. The typical architecture of a CNN is composed of convolutional and pooling layers (LeCun et al., 2015). In the convolutional layers, convolutional filters slide over the input image, and the feature extraction consists of finding weights for each filter element and the bias value.

The output of the convolutional layer is the feature map that serves as the input for the next convolutional layer, which can then learn more complex features. A set of weights and biases is passed through non-linear activation functions such as the sigmoid function and rectified linear unit (ReLU). Each group of the extracted features in a layer shares the same filters, allowing the detection of the same feature pattern in different parts of the image. Conversely, different features result from the various filters that lead to easy feature detection from the input data (LeCun et al., 2015).

The convolution layers may be separated by the pooling layers that merge the extracted features before passing them on to the next convolution layer. The merge consists of averaging or taking maximum value for the highlighted features and increasing the computation efficiency by reducing the size of the features. Max-pooling was applied in this work, during which the dimension of the extracted features is reduced by taking the maximum value from the features. In other words, if a feature with a 2×2 dimension was passed through the max-pooling layer with a 2×2 dimension, the maximum value from the feature values of 2×2 is output as one value.

A CNN model can be configured to have multiple stacked convolutional and pooling layers, followed by additional convolutional layers, and then fully connected layers. The weights in all the filters can be found during the CNN training using the backpropagation algorithm (LeCun et al., 2015). We applied the AlexNet as the standard CNN architecture for the configuration based on the Tensorflow library (Abadi et al., 2016). For training the CNN model, the Adam optimizer (Santoro et al., 2017) was used to optimize the weights by minimizing the difference between the Microcystis biomass predicted by the CNN model and the EFDC–NIER model. The mean square error (MSE) are adopted as the cost function:

$$y = \frac{\sum_{i=1}^{n} \left(C_i^l - E_i^l\right)^2}{n} \tag{3}$$

where n is the number of training data; C_i^l is the predicted Microcystis biomass by CNN in carbon concentration ($gC \cdot m^{-3}$); and E_i^l is the predicted Microcystis from EFDC–NIER.

Although multiple non-linear weights in the stacked CNN layers can be trained from the complicated relationships between inputs and outputs, a significant portion of such relationships may be the result of sampling noise by the training data. The dropout method was adopted in this study to prevent possible overfitting. The dropout layer temporarily removes one node that is chosen randomly from each layer in the network. Previous studies have suggested that the optimal probability of dropping nodes is 0.5 (Krizhevsky et al., 2012;). The weights in the dropped nodes are not renewed in the training network, preventing the convergence of particular weights during the training (Srivastava et al., 2014).

The statistical distribution of input data is changed in the intermediate convolutional layers because the distribution of the weights in the activation function is constantly changed during training. Accordingly, the training speed of CNN model may become slow because each layer has to learn new data distribution in every training iteration (Ioffe and Szegedy, 2015). To address this problem, we placed a batch normalization layer after each convolutional layer. This layer normalizes the output from the previous convolutional layer and feeds the normalized output into the next layer as the input data (DeVries and Taylor, 2017). The batch normalization also prevents the early convergence of weights to a narrow range of values.

2.3.1. CNN setup

The EFDC–NIER model outputs of each grid with a 12-hour interval for the calibration and validation periods were used as inputs for the CNN. The grid cell data consists of Microcystis biomass, water quality variables, and environmental variables, such as flow velocity, temperature, and elevation (Fig. 2d and 4a). The atmospheric data, such as atmospheric pressure, dry and wet temperatures, rainfall, evapotranspiration, solar radiation, and cloud

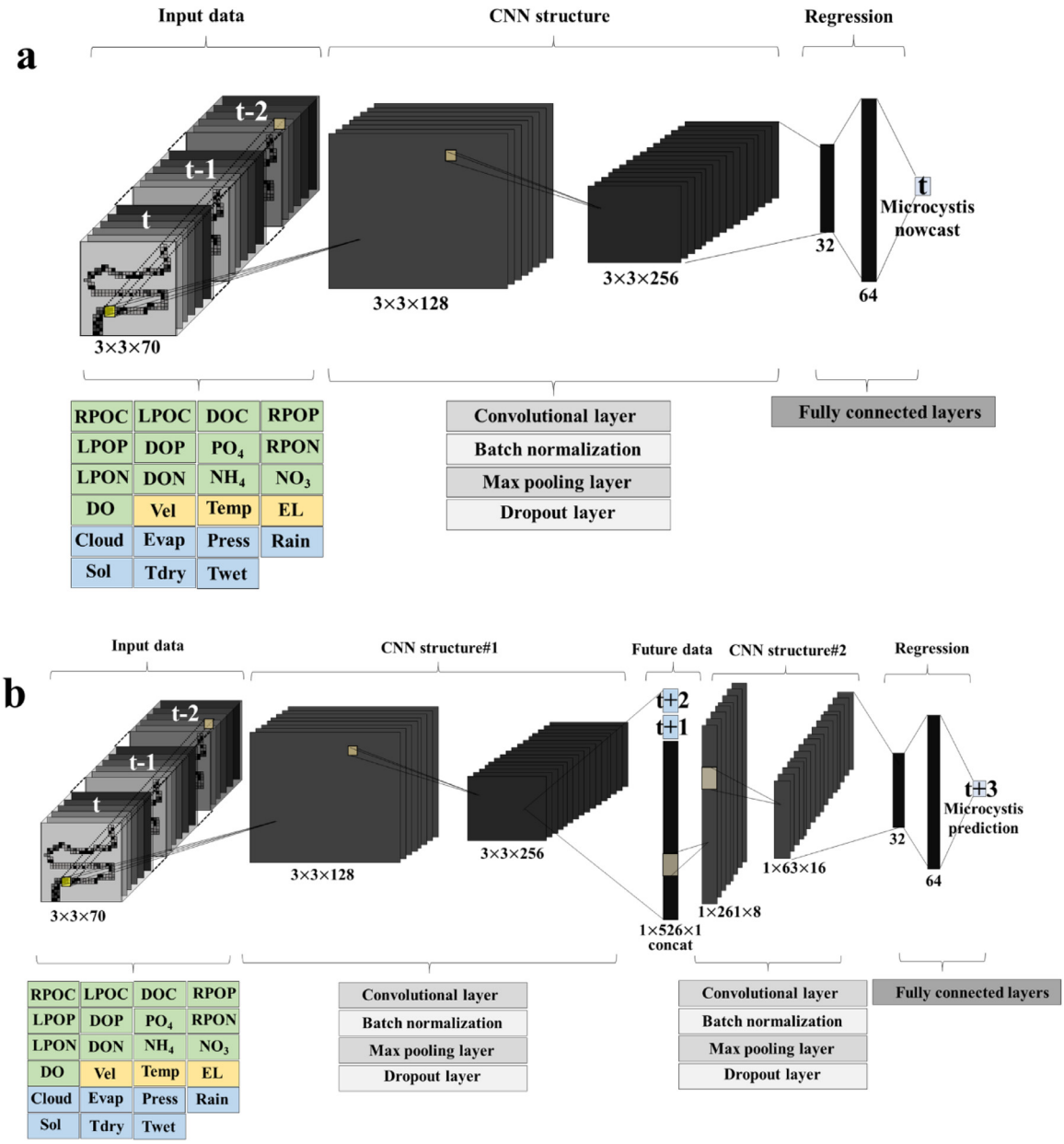

Fig. 3. CNN structure for Microcystis simulation using nutrients, environmental, and weather input, a) Microcystis nowcasting and b) Microcystis prediction with different prediction times (3-day lead time is $t + 3$ and 7-day lead time is $t + 7$).

cover, used for the EFDC–NIER model simulation were assigned to each grid to maintain the same grid data format. All the input data were stacked for the configuration of the CNN model (Fig. 2e and 4b) so that the model can reflect the features of the water quality. The dimensions of the input image data for the CNN were composed of the number of grid cells (N), size of the grid cells including width (W) and height (H), and the number of input variables, including water quality, environmental, and atmospheric variables (C). The amount of data is dependent on the assumption of the number of observations. The width and height of the input image were the size of the input window segmented as 3 × 3 pixel sizes (i.e., $N \times 3 \times 3 \times C$). This study assumed that the observation point was located in the center of the input windows. The model included 23 channel dimensions, including the stacked water quality, environmental, and atmospheric states (Fig. 4c). The internal layers and parameters of the CNN model were manually varied to provide an accurate Microcystis biomass prediction (Fig. 2g). Two convolutional layers were utilized with batch normalization, max pooling, and dropout layers to extract the input features. The size and number of filters in the first convolutional layer were adopted as 2 × 2 pixels and 128, respectively, with the second layer containing filters of 2 × 2 pixel size and 256, respectively. After finishing the feature extraction of the convolutional layers, two sequential, fully connected layers with 32 and 64 nodes were adopted to nowcast the Microcystis concentrations (Fig 3a). Moreover, we added new layers between the second convolutional layer and the fully connected layer.

After the feature extraction of the second convolutional layer, the output was vectorized, and future weather data were then added. The future weather data used were for two days before the target prediction day. Then, two 1-dimensional convolutional layers were used with 8 and 16 filters with 1 × 5 pixel sizes (Fig. 3b). Finally, the same number of fully connected layers with the now-

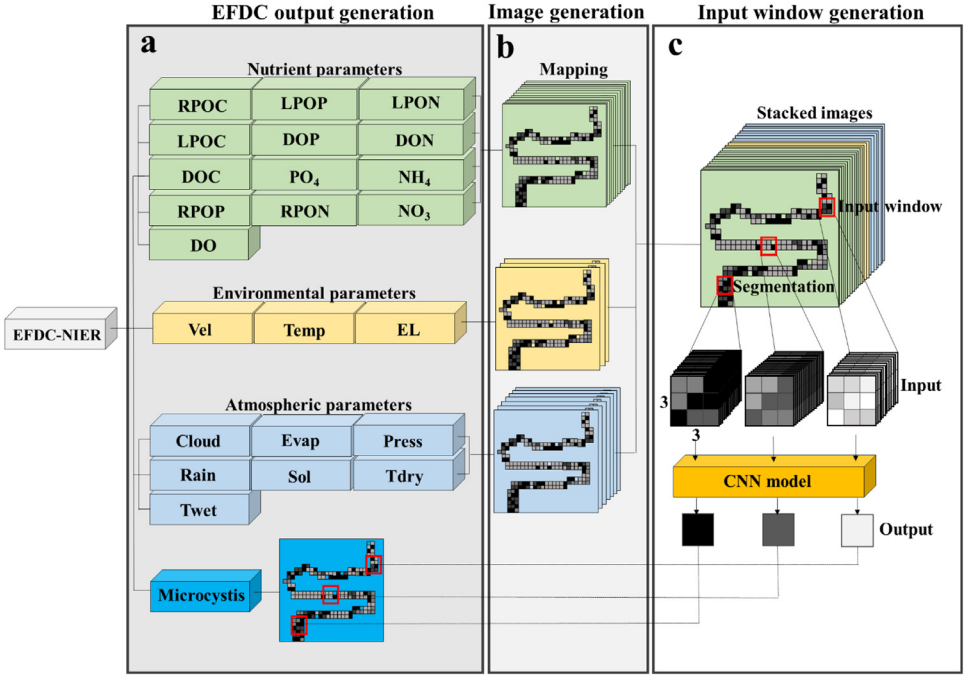

Fig. 4. CNN input generation using EFDC–NIER model simulation. a indicates outputs of EFDC–NIER including algal nutrients, environmental and atmospheric parameters; b presents stacking output grids of EFDC–NIER for generating CNN input images; c indicates input window segmentation from the stacked input imageries for predicting Microcystis concentration.

casting structure was used to predict the Microcystis concentration.

2.4. Scenario analysis

Scenarios regarding Microcystis biomass prediction under various assumptions were analyzed using the CNN model in which the EFDC outputs were assumed to be the true data in an image data format with high temporal frequency (Fig. 2h–j). The scenario analysis included Microcystis predictions with different forecast lead times, observation grid cell densities, and observation errors. Using the EFDC–NIER output allowed us to evaluate the CNN performance in predicting Microcystis concentrations for different prediction times from nowcasting to 7-day lead times (Fig. 2h). The scenarios with various spatial densities of observation grid cells evaluated the effect of the sparsity of observation points on the prediction accuracy. Because the grid cells of the EFDC–NIER model could be treated as real observation points, synthetic data sets with different observation points could be generated by a random selection of grid cells. The scenarios included data sets using 100%, 50%, and 20% of the grid cells, respectively (Fig. 2i). Scenarios with different observation errors were developed by introducing random errors that followed uniform, normal, or gamma distributions with varying magnitudes (i.e., 10%, 50%, and 100%) and were added to the CNN inputs (Fig. 2j)

3. Results and discussion

3.1. EFDC–NIER model for Microcystis simulation

For an accurate simulation of the hydrodynamics of the river reaches confined by the large weirs, the measured flow rate data were used at each weir point as boundary conditions, which forced the simulated water levels to be comparable to those observed. For water quality, Table 1 presents the calibration and validation results of the water quality variables. The unit of Microcystis from the EFDC–NIER output was transformed from carbon unit to cells using the carbon contents of Microcystis cells in HC (MOE, 2016). The results for 2015 showed the lowest calibration errors compared with the validation errors from the other years. The Microcystis validation results produced relatively large errors compared with the other variables. These results were attributed to the inherent characteristics of microorganism measurements in the environment, which normally show significant spatial and temporal heterogeneity (Wu and Xu, 2011). Although EFDC–NIER always has internal simulation errors, the reason for the use of the synthetic data as ground-truth is that this study requires sufficient image data resources with high temporal frequency to evaluate the Microcystis prediction performance of the CNN model.

Table 1
Water quality calibration results of EFDC–NIER.

	RMSE			
	2015	2016	2017	2018
[1]Temp	0.92	1.15	0.769	1.06
[2]DO	1.35	2.33	2.20	3.27
[3]BOD	0.83	1.04	0.84	1.04
[4]TN	0.32	0.51	0.67	0.54
[5]NO_3	0.32	0.54	0.66	0.48
[6]NH_4	0.03	0.21	0.075	0.073
[7]TP	0.012	0.044	0.033	0.034
[8]PO_4	0.006	0.018	0.011	0.017
[9]Chl-a	9.52	16.37	11.80	19.93
[10]MS	6864.84	48,320.78	26,675.62	48,146.47

[1]water temperature (°C), [2]dissolved oxygen(mg L^{-1}), [3]biological oxygen demand (mg L^{-1}), [4]total nitrogen (mg L^{-1}), [5]nitrate (mg L^{-1}), [6]ammonium (mg L^{-1}), [7]total phosphorus (mg L^{-1}), [8]phosphate (mg L^{-1}), [9]chlorophyll-a (mg L^{-1}), [10]Microcystis (cells mL^{-1}).

3.2. CNN for *Microcystis* biomass nowcasting

The nowcasting of Microcystis biomass by the CNN model was conducted for the data generated by the EFDC–NIER model. Only

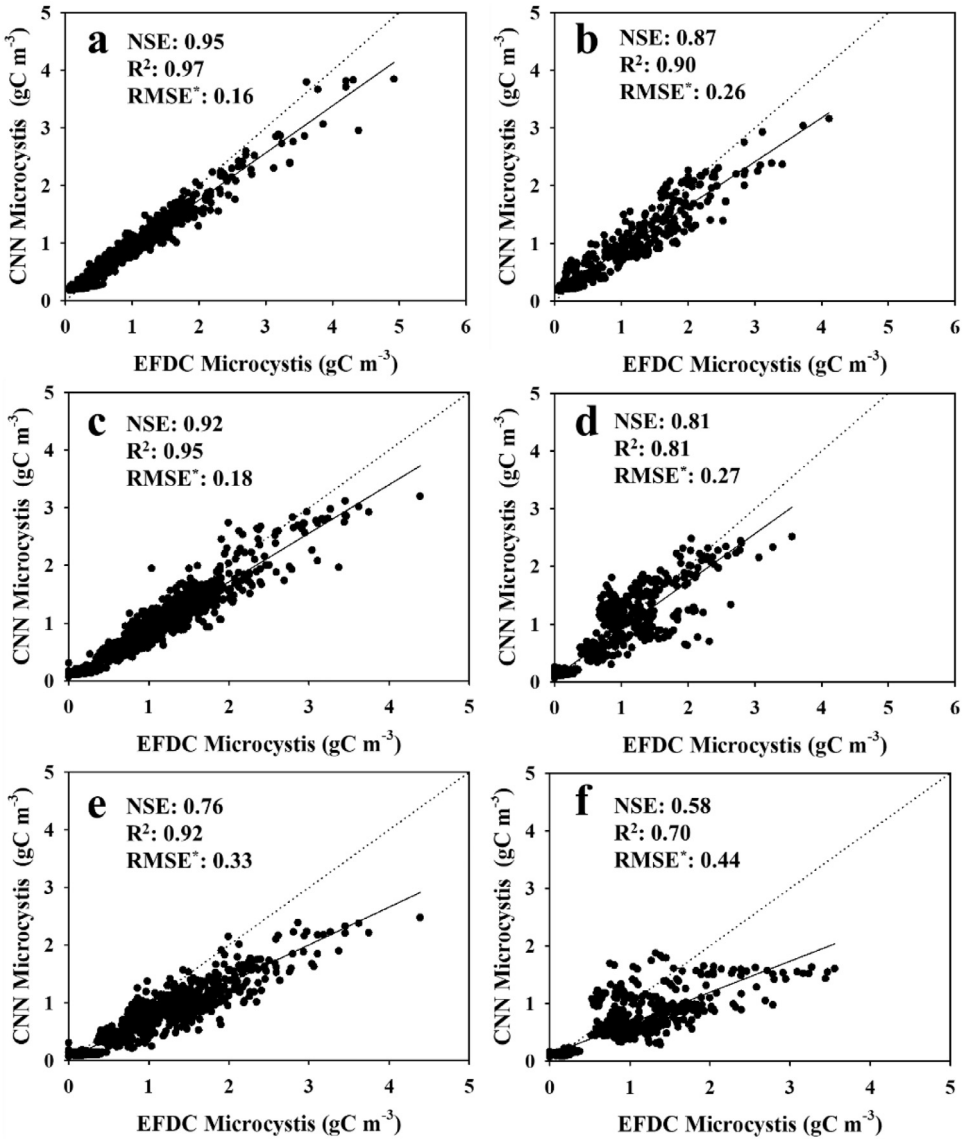

Fig. 5. Microcystis prediction of CNN varied to prediction time difference, a-b) training and validation results of nowcasting, respectively; c-d) training and validation results with 3-day lead time; and e-f) training and validation with 7-day lead time. (unit of RMSE* is gC m^{-3}).

data from August were used to investigate the most vulnerable period of the year to the Microcystis-dominated bloom. To consider the possible influence of previous conditions on the current Microcystis biomass, the input data for three days, from current day (denoted as t) to two days ago (t - 2), were entered in the CNN model to predict the status of Microcystis at the current date t. A total of 30 random grid cells were assumed available for measurements. The earliest 70% of the input data were used for model training, and the remaining 30% were used for validation. The two figures in the first row of Fig. 5 show the CNN performance on the Microcystis nowcasting for the training (Fig. 5a) and validation periods (Fig. 5b). High accuracy, with an NSE of 0.95 and R^2 of 0.97, was achieved for the training (Table 2). The outcomes of the validation showed reasonable results with an NSE of 0.87 and R^2 of 0.90. These results proved that the CNN model could provide reliable nowcasting outcomes for Microcystis by reflecting input variations with time. The combination of the input variables for the nutrients, environment, and atmospheric conditions and the selection of three days (i.e., t - 2, t - 1, and t) for the input time window were sufficient to reflect the conditions of the Microcystis-dominated blooms at day t. Several studies have demonstrated the reasonable accuracy of CNN-based metamodels in various applications. Wang and Xu (2019) built a CNN model with four one-dimensional convolutional layers to estimate the walking and running speeds of a person using acceleration data.

Table 2
CNN performance of difference prediction times.

Prediction time	Training			Validation		
	NSE	R^2	RMSE*	NSE	R^2	RMSE*
Nowcasting	0.95	0.97	0.16	0.87	0.90	0.26
3 days	0.92	0.95	0.18	0.81	0.81	0.27
7 days	0.76	0.92	0.33	0.58	0.70	0.44

* Unit of RMSE is gC m^{-3}.

This study found that the CNN structure provided the estimation errors within 18% compared to the observations. Pyo et al. (2019) designed a CNN structure with two convolutional layers to estimate cyanobacteria concentrations from hyperspectral imagery showing R^2 values of 0.86 and 0.73, with respect to the observed phycocyanin and chlorophyll-a concentrations, re-

spectively. Hong et al. (2019) estimated concentrations of ultra-fine particles using a pretrained ImageNet structure with satellite images, and their results demonstrated a regression accuracy of over 86%. Despite of the different input composition and CNN structures, the results of previous studies indicate that the deep feature extraction allowed for finding robust relationships, thereby providing a reliable performance of CNN metamodels. By extension, our study provided an example of time series data as efficient input for nowcast using a CNN model.

3.3. Scenario analysis

3.3.1. Effect of forecast lead time

New weather data layers for 2 days (t + 1 and t + 2) were added to the nowcasting CNN configuration to configure a new CNN model for forecasting Microcystis for 3 days in the future (t + 3) (Fig. 4). Similarly, weather data layers for 6 days, from t + 1 to t + 6, were added for forecasting 7 days in the future (t + 7). The two figures in the second row of Fig. 5 show the results of Microcystis forecasting with the 3-day lead time for the training (Fig. 5c) and validation periods (Fig. 5d). Fig. 5e and f present the results of the 7-day lead time. As expected, the forecast accuracies of the 3-day lead time simulation were lower than for nowcasting in both the training and validation, but they were still reasonably high. However, the accuracy of the 7-day lead time simulation decreased significantly. In particular, the accuracy difference between the two cases was the largest for the 3-day forecasting, with NSE values of 0.92 (training) and 0.81 (validation) versus 0.76 and 0.58 for the 7-day simulation, respectively (Table 2). These results indicate that the performance of the CNN model can be significantly degraded with a forecast lead time past a certain point. Beyond that point, the nutrient, environment, and weather inputs for t - 2, t - 1, and t could correspond poorly with the future biophysical processes affecting Microcystis concentrations (Fig. 6).

This trend of decreasing CNN performance with increasing lead time can be observed in prior studies; Chattopadhyay et al. (2019) introduced a decrease in accuracy of cold spell class prediction from 73% to 47% when the lead time changed from 1-day to 5-day. The time scale of the prediction also affected the CNN performance in the work of Ghimire et al. (2019) who found that the relative mean absolute error of the global solar radiation estimation with CNN was 9.91% to 19.57% for daily and monthly values, respectively. When the forecast lead time increases, the internal uncertainty of the CNN model and imperfect descriptions of the extracted input features influence on model training, resulting in a decrease in the prediction accuracy (Miao et al., 2019). Thus, the current study demonstrated the robust short-term Microcystis forecasting ability of the CNN model.

3.3.2. Spatial density of observations

The prediction performance of the CNN model was tested in terms of varied observation density. The spatial observation points were randomly assigned to 50% and 20% of the total model grid cells (denoted as OD50 and OD20, respectively, and as OD100 when all the grid cells were utilized). The CNN model configured for Microcystis nowcasting was adopted for these scenarios. The NSE values of the validation results decreased from 0.84 to 0.70 as the observation density decreased from 100% to 20%. The trained models for OD100, OD50, and OD20 were each applied to the entire grid to generate a Microcystis map, and the accuracy of the three models was compared against the EFDC–NIER output grid cells. Fig. 7 presents the RMSE variation of each case, OD100, OD50, and OD20, for August. OD100 showed the lowest RMSEs, while OD20 produced the highest Table 3.

Table 3
CNN performance dependence on observation density variation.

Observation Density	Training			Validation		
	NSE	R^2	RMSE*	NSE	R^2	RMSE*
OD 100	0.93	0.93	0.17	0.84	0.88	0.29
OD 50	0.81	0.97	0.28	0.76	0.91	0.35
OD 20	0.87	0.95	0.22	0.70	0.84	0.37

* Unit of RMSE is gC m^{-3}.

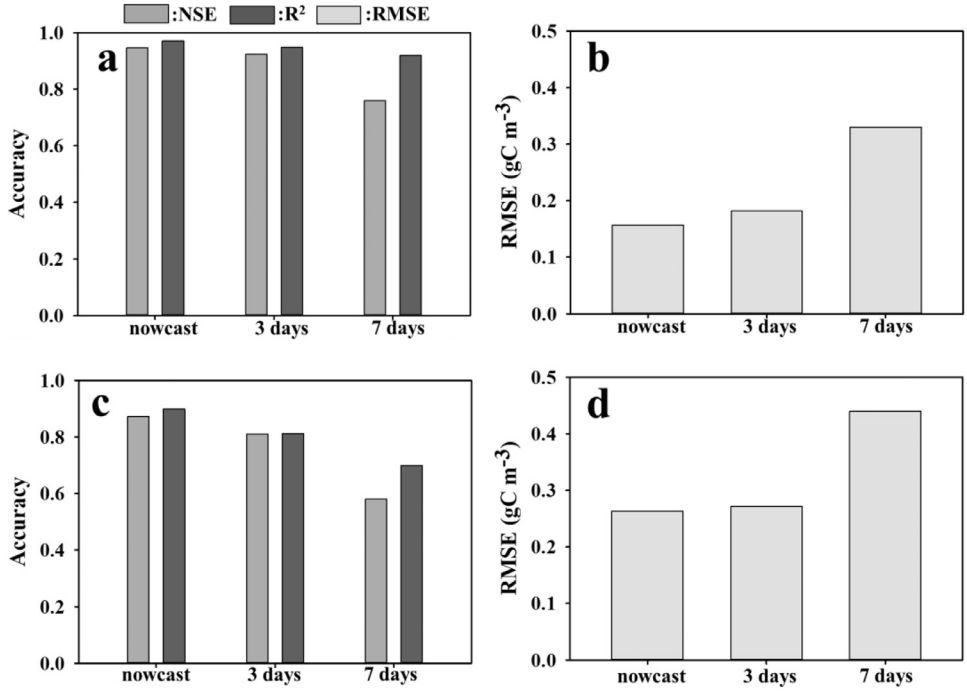

Fig. 6. CNN performance for Microcystis, a-b) accuracy and RMSE of training results; and c-d) accuracy and RMSE of validation results.

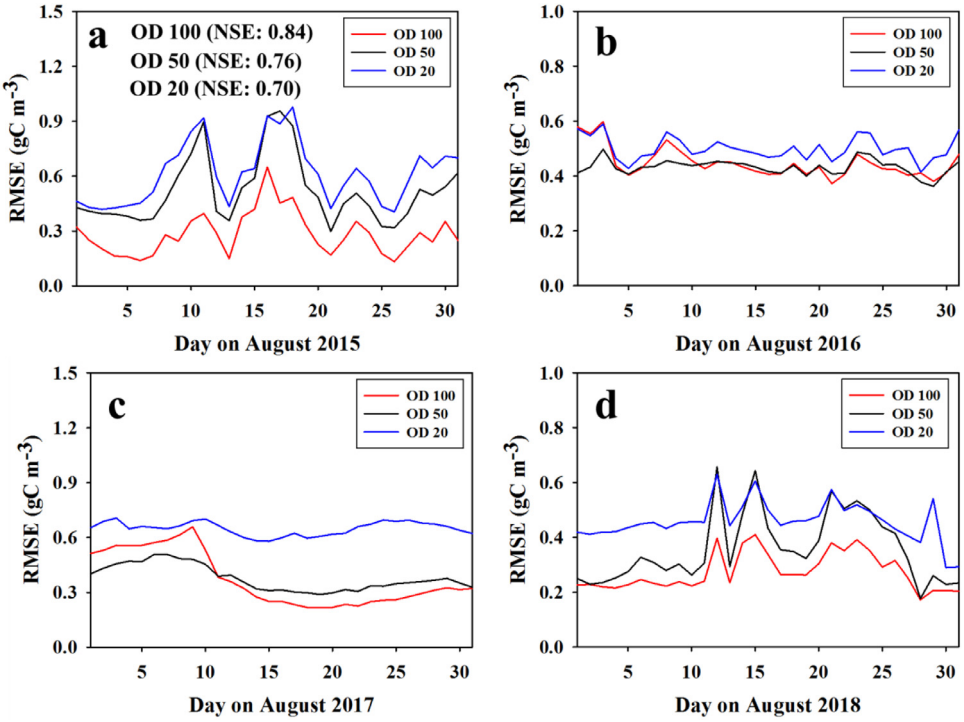

Fig. 7. CNN performance dependent on the observation point density, a) RMSE in August 2015; b) RMSE in August 2016; c) RMSE in August 2017; and d) RMSE in August 2018. OD indicates observation density, OD 100 presents total pixels as the observation points, OD 50 is total pixels of 50% as the points, and OD 20 is 20% as the points.

These results indicate that the CNN regression model with higher observation densities can describe the spatial and serial features of Microcystis relatively well compared with the model with a lower density. Therefore, the deep learning performance is directly influenced by the amount of the available data. Sarikaya et al. (2014) identified training accuracy variations of a deep learning model considering different data sizes, showing the highest classification performance for the case with the largest data size. An insufficient amount of data creates a generalization problem for a deep learning model because the trained model may not be capable of representing the features of new data that have not been or have rarely been present in the training data (Huang and Kingsbury, 2013). Data augmentation techniques, such as deformation and rotation of inputs to increase the size of the input data, have been used in other studies to address this issue (Salamon and Bello, 2017). Furthermore, the CNN performance is also dominated by the complexity of the input data. In particular, water quality predictions with CNNs using multi-dimensional imagery require spatially distributed nutrients, environmental variables, and Microcystis concentrations, which can be acquired only from intensive coupled field monitoring and laboratory experiments. Otherwise, one may be left with the limited available data. However, this study identified the randomly selected 30 grid cells provided reasonable prediction accuracy of CNN model with respect to different forecast lead time. Although the deep learning performance typically benefits from larger datasets, it is difficult to define the minimum or maximum size of the dataset for deep learning models because the complexity of the observation data is an important factor (Du et al., 2018). Pyo et al. (2019) obtained reasonable cyanobacteria concentration estimation using CNN model with complex image inputs despite the limited number of observation data points. Preliminary research is needed to establish an acceptable amount of input data for CNN regression modeling.

3.3.3. Adding noise to CNN inputs

In this study, three different types of distributions—uniform, normal, and gamma distribution were assumed for the random errors that were added to the training and validation data to build three data cases with different synthetic observational errors. The CNN model with the 3-day forecast lead time was utilized for these scenarios. Four cases for each error distribution model were introduced based on different error amplitudes of the input data: no error, 10%, 50%, and 100% errors.

Table 4 shows that the training and validation accuracy decreased as the observational error rate increased for all error distribution functions. As the error rate increased for the gamma distri-

Table 4
CNN performance dependent on adding error function to input.

Uniform distribution	Training			Validation		
	NSE	R^2	RMSE*	NSE	R^2	RMSE*
No error	0.92	0.95	0.18	0.81	0.81	0.27
10%	0.76	0.92	0.33	0.52	0.68	0.43
50%	0.70	0.90	0.36	0.51	0.61	0.44
100%	0.68	0.91	0.37	0.45	0.61	0.46
Normal distribution	Training			Validation		
	NSE	R^2	RMSE*	NSE	R^2	RMSE*
10%	0.85	0.94	0.25	0.68	0.70	0.35
50%	0.82	0.93	0.28	0.67	0.70	0.36
100%	0.76	0.95	0.33	0.60	0.68	0.40
Gamma distribution	Training			Validation		
	NSE	R^2	RMSE*	NSE	R^2	RMSE*
10%	0.91	0.93	0.20	0.76	0.75	0.31
50%	0.85	0.95	0.25	0.65	0.67	0.37
100%	0.71	0.89	0.36	0.60	0.68	0.39

* Unit of RMSE is gC m^{-3}.

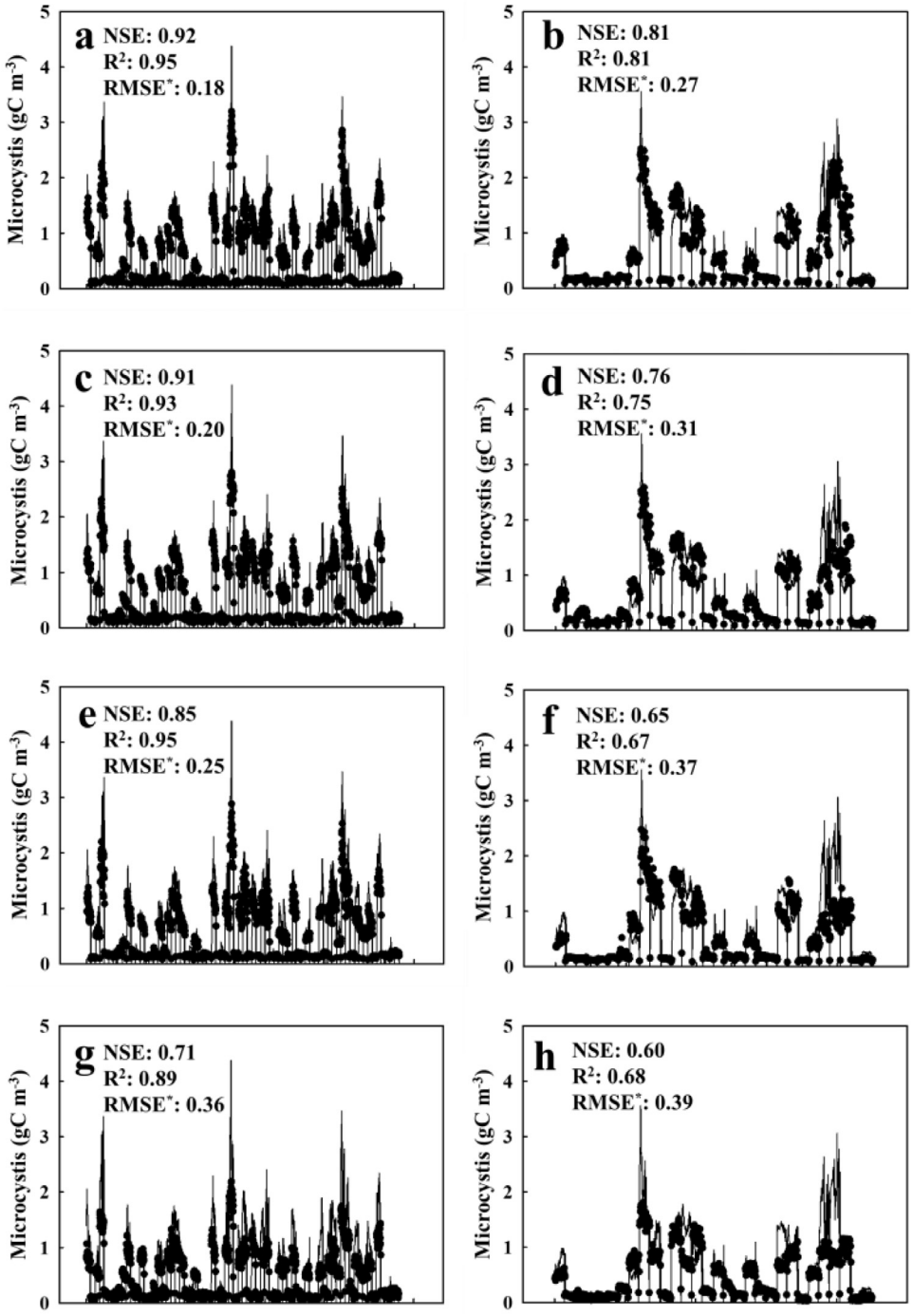

Fig. 8. Microcystis prediction with 3-day lead and gamma distribution errors added to the inputs, a-b) training and validation results with no error; c-d) training and validation with10% error; e-f) training and validation with 50% error; and g-h) training and validation with 100% error. x-axis indicates the data points.

bution, the forecast accuracy decreased significantly: the NSE was reduced from 0.81 to 0.60, and the RMSE increased from 0.27 to 0.39 gC·m^{-3}, respectively, for the validation results (Fig. 8b, d, f, and h). Overall, more substantial errors in the input data led to a decrease in the forecast accuracy of the CNN model because the nutrient and environmental inputs with significant errors could worsen the relationship with the Microcystis data (Fig. 9). For example, nutrient concentrations higher than an optimum level for Microcystis growth (nutrient saturation) cannot reflect the variation in Microcystis biomass in August because the algal growth rate reaches its maximum (Baldia et al., 2007).

Similarly, high solar intensity could reduce the biomass of *Microcystis* because algal growth is photo-inhibited over a certain solar threshold (Zevenboom and Mur, 1984). The optimal water temperature for Microcystis growth falls in the range between 28.8 °C and 30.5 °C, with temperatures over 30.5 °C decreasing the growth rate (Robarts and Zohary, 1987). Furthermore, erroneously high flow rates would inhibit Microcystis bloom formation. Lin et al. (2012) found that the threshold flow rate for Microcystis growth was 25 cm·s^{-1}. Substantial observational errors affect the input data, resulting in incorrect nutrient and environmental feature extraction during the CNN model training. The accuracy

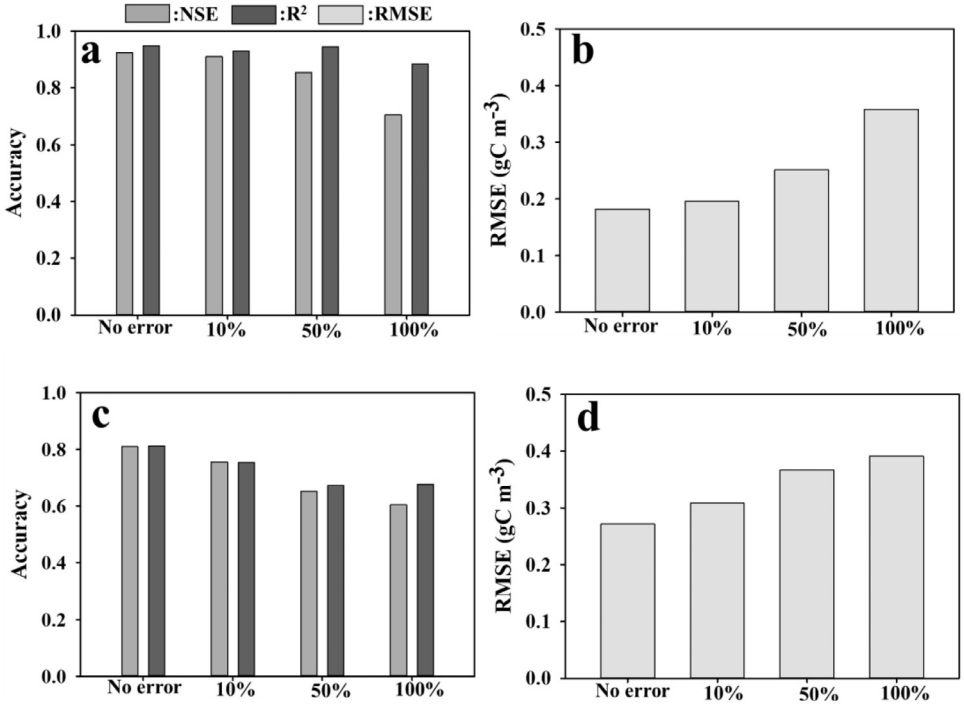

Fig. 9. The 3-day lead Microcysis prediction performance of CNN with gamma distribution error in input, a-b) accuracy and RMSE of training results; and c-d) accuracy and RMSE of validation.

results revealed the significant effect of input error on the CNN prediction performance, implying that reducing data acquisition errors is required to improve the prediction performance. In addition, CNN prediction capability was not compromised by input errors of less than 10%. Elmes et al. (2020) showed that the classification usage of the deep learning model could tolerate the input errors less than 20%. If the error does not approach the critical level, the CNN model may provide reliable prediction results. Furthermore, a large enough dataset can mitigate the effect of randomly distributed input errors during training because the CNN model can learn the actual relationship between input and output (Kang et al., 2019).

3.4. Deep learning application with high temporal frequency images

We used the output of EFDC–NIER as the synthetic dataset to see what degree of detail may be acceptable to train the CNN and obtain reasonable predictions with this CNN. The advantages of using the realistic synthetic dataset are that it can be decimated to different degrees, and the noise in perturbed data can be accurately defined. However, neither the NSE nor RMSE within the scope of the study accounts for spatial patterns in the concentrations of Microcystis. Synthetic data and subsequent visualization could help in defining the metrics for the CNN reproduction of specific magnitude ranges or areas of Microcystis distributions. A video clip was generated to compare the Microcystis prediction map created by the CNN and EFDC–NIER models. In the clip, the Microcystis spatial distribution from the CNN model mostly agreed well with the distribution of the EFDC–NIER model (Fig. 10). For 2015, the CNN demonstrated the best performance regarding spatial distribution along the HC reach (Fig. 10a). The prediction results for 2016 and 2017 showed overestimations of the Microcystis concentration. The southwest and the northeast sections notably overestimated the Microcystis compared with EFDC–NIER prediction (Fig. 10b). For 2017, the CNN model overestimated Microcystis along the HC reach (Fig. 10c), but the prediction for 2018 slightly underestimated the concentrations in the northwest part of the reach (Fig. 10d). These prediction uncertainties may be caused by the internal errors in the synthetic data from the EFDC–NIER model because the results of this work were limited by the list of the water quality variables (Table 1). The available synthetic data may approach the critical level of error. Accordingly, the weak relationship of input variables with Microcystis concentrations resulted in the mediocre validation performance of the CNN model. Overall, a high density of data appears to be critical for the CNN Microcystis prediction performance. Acquiring high temporal frequency images can lead to the robust accuracy of the CNN deep learning applications, which could be a promising tool for the short-term prediction of Microcystis biomass.

This work contributes to the ongoing discussion on the role of deep learning models and mechanistic models (Baker et al., 2018). Whereas mechanistic models, such as EFDC–NIER, have difficulties in incorporating data from different space and time scales and can work with relatively small datasets, deep learning models, such as CNN, easily incorporate data from multiple scales but require large datasets for training ad validation. Mechanistic models once validated, can be used as a predictive tool where experiments are difficult or costly to perform whereas deep learning models can only make predictions that relate to patterns within the data. This work did not contrapose the models of the two types but rather illustrated two aspects of synergetic use of deep learning and mechanistic models. First, a large realistic dataset was required to research the ability of the deep learning model CNN to learn the spatio-temporal patterns of Microcystis biomass concentrations with different spatio-temporal data density. Such a dataset could not be collected in the field at the current level of equipment capabilities, and the dataset was produced as the output of the mechanistic model EFDC–NIER. Second, simulations with complex mechanistic models are often computationally expensive, and EFC–NIER provides an example of that is substantially time-consuming processes spending 20 h compared to spending 4.97 min for CNN simulation (Guo et al., 2016). For such cases, the deep learning model can create the surrogate model (Hong et al., 2017) that can be trained at a limited number of detailed simulations without the configuration of initial and boundary conditions,

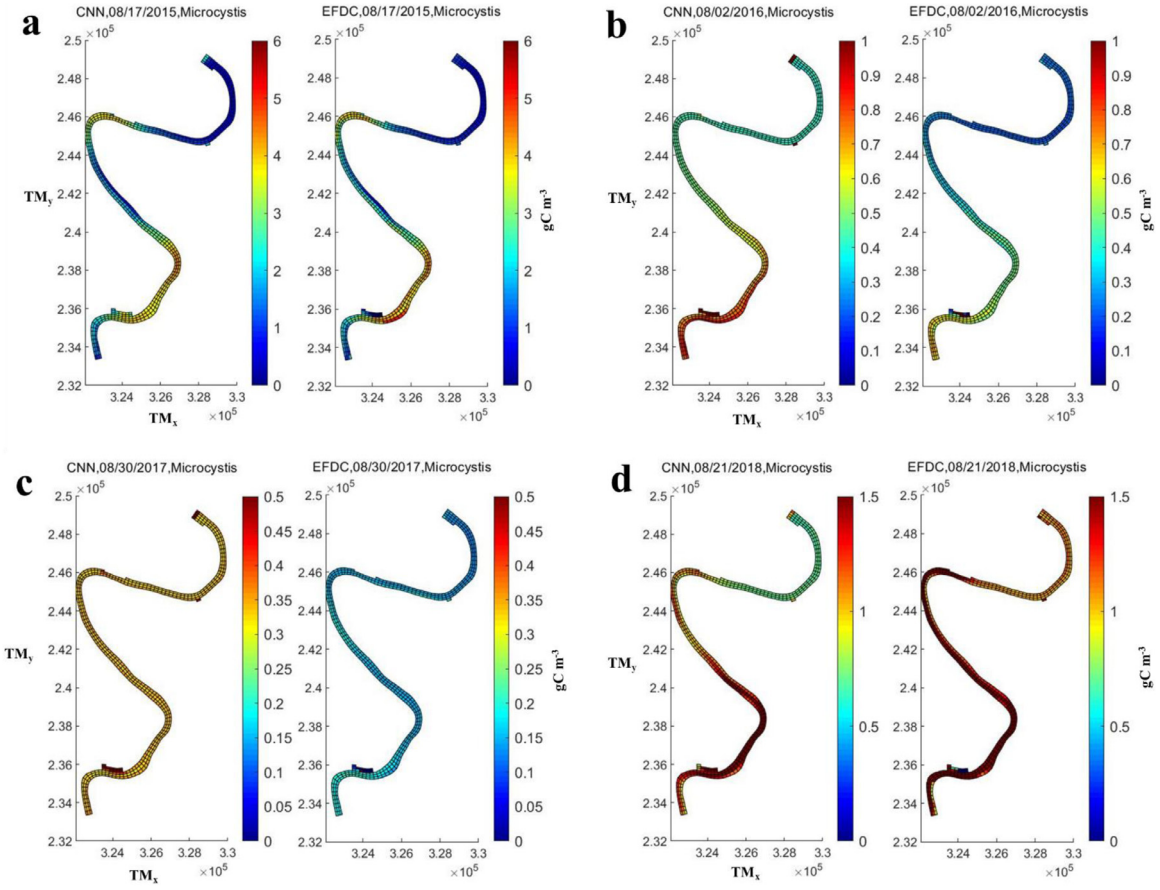

Fig. 10. Microcystis concentration map from the CNN predictions.

the precise characterization of more than 130 number of model parameters (Djurovic et al., 2015; Ji, 2017). The approximate surrogate model, in our case CNN is used for future predictions. Integration of mechanistic and deep learning models is a fast-developing field (Schuwirth et al., 2019; Zaherpour et al., 2019) presenting promising and important research avenues.

Remote sensing, particularly using small unmanned aerial vehicles (drones) and multiple satellites, can provide high-frequency of multi- or hyper-spectral image data. These types of data can be converted to water quality variable maps, including water temperature (Alcântara et al., 2010), dissolved oxygen (Kim et al., 2020), total suspended solid (Wang et al., 2018), and total phosphorus (Xiong et al., 2019), which were used as CNN inputs in this study (Choi et al., 2019; Ta and Wei, 2018). We expect groups of remote sensing images to be replaced as CNN input for water quality prediction study in the near future because the image sets can include high temporal frequencies. Hence, the prediction performance of a CNN model was tested preliminarily in this study, assuming that high-frequency image data are available.

4. Conclusion

This study investigated the CNN performance for Microcystis prediction based on synthetic high temporal frequency images. The EFDC–NIER model was developed to generate the synthetic nutrient, environmental, and atmospheric grid cell data that were utilized as inputs for the CNN model. The CNN model was then configured for Microcystis nowcasting and forecasting. Modeling scenarios were designed and implemented to investigate how the CNN model performance is affected by the forecast lead time, spatial observation density, and noise in the input data. Temporal degradation in the input datasets led to a decrease in CNN accuracy. However, the CNN performance remained acceptable even after a significant decrease in the volume of spatial data and a substantial addition of noise. Thus, this work highlights the significance of spatial monitoring with high temporal frequency and contributes to the development of insights and guidance regarding data acquisition for future deep learning research for water resources. For related future work, a sensitivity analysis would be of interest to determine the relative importance of accurate variable inputs, such as water quality, environmental, and weather conditions, as potential cyanobacterial bloom predictors in the CNN model.

Declaration of Competing Interest

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

Acknowledgement

This research was supported by the Basic Core Technology Development Program for the Oceans and the Polar Regions of the National Research Foundation (NRF) funded by the Ministry of Science, ICT & Future Planning (NRF-2016M1A5A1027457) and was also supported by Water Environment and Infrastructure Research Program (NIER-2018-01-01-036) funded by the National Institute of Environmental Research.

Supplementary materials

Supplementary material associated with this article can be found in the online version at doi:10.1016/j.watres.2020.116349.

References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., Kudlur, M., Levenberg, J., Monga, R., Moore, S., Murray, D.G., Steiner, B., Tucker, P., Vasudevan, V., Warden, P., Wicke, M., Yu, Y., Zheng, X., 2016. TensorFlow: A system For Large-Scale Machine learning, 12th USENIX Symposium on Operating Systems Design and Implementation Savannah, GA, USA.

Alcântara, E.H., Stech, J.L., Lorenzzetti, J.A., Bonnet, M.P., Casamitjana, X., Assireu, A.T., de Moraes Novo, E.M.L., 2010. Remote sensing of water surface temperature and heat flux over a tropical hydroelectric reservoir. Remote Sens. Environ. 114 (11), 2651–2665.

Baker, R.E., Peña, J.M., Jayamohan, J., Jérusalem, A., 2018. Mechanistic models versus machine learning, a fight worth fighting for the biological community? Biol. Lett. 14 (5), 20170660.

Baldia, S.F., Evangelista, A.D., Aralar, E.V., Santiago, A.E., 2007. Nitrogen and phosphorus utilization in the cyanobacterium Microcystis aeruginosa isolated from Laguna de Bay, Philippines. J. Appl. Phycol. 19 (6), 607–613.

Beck, M.B., van Straten, G. (Eds.), 2012. Uncertainty and Forecasting of Water Quality. Springer Science & Business Media.

Chattopadhyay, A., Nabizadeh, E., Hassanzadeh, P., 2019. Analog forecasting of extreme-causing weather patterns using deep learning. arXiv preprint .

Chen, Y., Jiang, H., Li, C., Jia, X., Ghamisi, P., 2016. Deep feature extraction and classification of hyperspectral images based on convolutional neural networks. IEEE Trans. Geosci. Remote Sens. 54 (10), 6232–6251.

Cho, K.H., Sthiannopkao, S., Pachepsky, Y.A., Kim, K.W., Kim, J.H., 2011. Prediction of contamination potential of groundwater arsenic in Cambodia, Laos, and Thailand using artificial neural network. Water Res. 45 (17), 5535–5544.

Choi, J.H., Kim, J., Won, J., Min, O., 2019. Modelling chlorophyll-a concentration using deep neural networks considering extreme data imbalance and skewness. In: 2019 21st International Conference on Advanced Communication Technology (ICACT). IEEE, pp. 631–634.

Clark, L.J., Jaworski, N.A. (1972). Nutrient transport and dissolved oxygen budget studies in the Potomac estuary.

DeVries, T., Taylor, G.W., 2017. Improved regularization of convolutional neural networks with cutout. arXiv preprint .

Djurovic, N., Domazet, M., Stricevic, R., Pocuca, V., Spalevic, V., Pivic, R., Gregoric, E., Domazet, U., 2015. Comparison of groundwater level models based on artificial neural networks and ANFIS. Scientif. World J. 2015.

Du, S.S., Wang, Y., Zhai, X., Balakrishnan, S., Salakhutdinov, R.R., Singh, A., 2018. How many samples are needed to estimate a convolutional neural network? In: Advances in Neural Information Processing Systems, pp. 373–383.

Elmes, A., Alemohammad, H., Avery, R., Caylor, K., Eastman, J.R., Fishgold, L., …, Lunga, D., 2020. Accounting for training data error in machine learning applied to Earth observations. Remote Sens. 12 (6), 1034.

Ghimire, S., Deo, R.C., Raj, N., Mi, J., 2019. Deep solar radiation forecasting with convolutional neural network and long short-term memory network algorithms. Appl. Energy 253, 113541.

Guo, X., Li, W., Iorio, F., 2016. Convolutional neural networks for steady flow approximation. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, pp. 481–490.

Hamrick, J.M. (1992). A three-dimensional environmental fluid dynamics computer code: theoretical and computational aspects.

Hamrick, J.M., Mills, W.B., 2000. Analysis of water temperatures in Conowingo Pond as influenced by the Peach Bottom atomic power plant thermal discharge. Environ. Sci. Policy 3, 197–209.

Hong, E., Pachepsky, Y., Whelan, G., Nicholson, T., 2017. Simpler models in environmental studies and predictions. Crit. Rev. Environ. Sci. Technol. 47 (18), 1669–1712.

Hong, K.Y., Pinheiro, P.O., Minet, L., Hatzopoulou, M., Weichenthal, S., 2019. Extending the spatial scale of land use regression models for ambient ultrafine particles using satellite images and deep convolutional neural networks. Environ. Res. 176, 108513.

Huang, J., Kingsbury, B., 2013. Audio-visual deep learning for noise robust speech recognition. In: *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, pp. 7596–7599.

Ioffe, S., Szegedy, C., 2015. Batch normalization: accelerating deep network training by reducing internal covariate shift. arXiv preprint .

Ji, Z.G., 2017. Hydrodynamics and Water quality: Modeling rivers, lakes, and Estuaries. John Wiley & Sons.

Kim, J., Lee, T., Seo, D., 2017. Algal bloom prediction of the lower Han River, Korea using the EFDC hydrodynamic and water quality model. Ecol. Modell. 366, 27–36.

Kim, Y.H., Son, S., Kim, H.C., Kim, B., Park, Y.G., Nam, J., Ryu, J., 2020. Application of satellite remote sensing in monitoring dissolved oxygen variabilities: a case study for coastal waters in Korea. Environ. Int. 134, 105301.

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 1097–1105.

Kromkamp, J., Walsby, A.E., 1990. A computer model of buoyancy and vertical migration in cyanobacteria. J. Plankton Res. 12 (1), 161–183.

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521 (7553), 436–444.

Lee, G.H., Kim, K.H., Park, Y.G., Lee, S.J., 2014. A study on development of a GIS based post-processing system of the EFDC model for supporting water quality management. Spatial Inf. Res. 22 (4), 39–47.

Lee, S.M., Kim, I.K., 2017. Analysis of correlation between cyanobacterial population and water quality factors in the middle and down stream region of nakdong river. J. Korean Soc. Water Wastewater 31 (1), 93–101.

Li, Z., Chen, Q., Xu, Q., Blanckaert, K., 2013. Generalized likelihood uncertainty estimation method in uncertainty analysis of numerical eutrophication models: take bloom as an example. Math. Probl. Eng. 2013.

Lin, L.I., Weia, Z.H., Feng-lan, C.H.E.N.G., Ting-ting, W.A.N.G., Xiao, T.A.N., 2012. Effects of continuous water flow on growth of the Microcystis Aeruginosa under high nutrient levels. Energy Procedia 17, 1793–1797.

Martin, J.L., 1988. Application of two-dimensional water quality model. J. Environ. Eng. 114 (2), 317–336.

Miao, Q., Pan, B., Wang, H., Hsu, K., Sorooshian, S., 2019. Improving monsoon precipitation prediction using combined convolutional and long short term memory neural network. Water (Basel) 11 (5), 977.

Ministry of Environment (MOE), 2016. Advancing Monitoring and Prediction Model in Order to Improving Algal Prediction Accuracy Incheon, Korea.

Park, Y., Cho, K.H., Park, J., Cha, S.M., Kim, J.H., 2015. Development of early-warning protocol for predicting chlorophyll-a concentration using machine learning models in freshwater and estuarine reservoirs, Korea. Sci. Total Environ. 502, 31–41.

Poliyapram, V., Imamoglu, N., Nakamura, R., 2019. Recurrent feedback CNN for water region estimation from multitemporal satellite images. Image and Signal Processing For Remote Sensing XXV 11155, 111550.

Pyo, J., Duan, H., Baek, S., Kim, M.S., Jeon, T., Kwon, Y.S., Lee, H., Cho, K.H., 2019. A convolutional neural network regression for quantifying cyanobacteria using hyperspectral imagery. Remote Sens. Environ. 233, 111350.

Robarts, R.D., Zohary, T., 1987. Temperature effects on photosynthetic capacity, respiration, and growth rates of bloom-forming cyanobacteria. N. Z. J. Mar. Freshwater Res. 21 (3), 391–399.

Salamon, J., Bello, J.P., 2017. Deep convolutional neural networks and data augmentation for environmental sound classification. IEEE Signal Process. Lett. 24 (3), 279–283.

Santoro, A., Raposo, D., Barrett, D.G., Malinowski, M., Pascanu, R., Battaglia, P., Lillicrap, T., 2017. A simple neural network module for relational reasoning. In: Advances in Neural Information Processing Systems, pp. 4967–4976.

Sarikaya, R., Hinton, G.E., Deoras, A., 2014. Application of deep belief networks for natural language understanding. IEEE/ACM Trans. Audio Speech Lang. Process. (TASLP) 22 (4), 778–784.

Schuwirth, N., Borgwardt, F., Domisch, S., Friedrichs, M., Kattwinkel, M., Kneis, D., Vermeiren, P., 2019. How to make ecological models useful for environmental management. Ecol. Modell. 411, 108784.

Shin, C.M., Min, J.H., Park, S.Y., Choi, J., Park, J.H., Song, Y.S., Kim, K., 2017. Operational water quality forecast for the Yeongsan River using EFDC model. J. Korean Soc. Water Environ. 33 (2), 219–229.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R., 2014. Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. 15 (1), 1929–1958.

Summers, J.K., Wilson, H.T., Kou, J., 1993. A method for quantifying the prediction uncertainties associated with water quality models. Ecol. Modell. 65 (3–4), 161–176.

Ta, X., Wei, Y., 2018. Research on a dissolved oxygen prediction method for recirculating aquaculture systems based on a convolution neural network. Comput. Electron. Agric. 145, 302–310.

Teles, L.O., Vasconcelos, V., Pereira, E., Saker, M., 2006. Time series forecasting of cyanobacteria blooms in the Crestuma Reservoir (Douro River, Portugal) using artificial neural networks. Environ. Manage. 38 (2), 227–237.

Ulanowicz, R.E., 1976. Modeling the Chesapeake Bay and tributaries: a synopsis. Chesapeake Sci. 17 (2), 114–122.

Wang, C., Li, W., Chen, S., Li, D., Wang, D., Liu, J., 2018. The spatial and temporal variation of total suspended solid concentration in Pearl River Estuary during 1987–2015 based on remote sensing. Sci. Total Environ. 618, 1125–1138.

Wang, K.L., Xu, J., 2019. A speed regression using acceleration data in a deep convolutional neural network. IEEE Access 7, 9351–9356.

Wang, P., Lai, G., Li, L., 2015. Predicting the hydrological impacts of the Poyang Lake Project using an EFDC model. J. Hydrologic Eng. 20 (12), 05015009.

Wójcik, P.I., Kurdziel, M., 2019. Training neural networks on high-dimensional data using random projection. Pattern Anal. Appl. 22 (3), 1221–1231.

Wu, G., Xu, Z., 2011. Prediction of algal blooming using EFDC model: case study in the Daoxiang Lake. Ecol. Modell. 222 (6), 1245–1252.

Xie, Z., Lou, I., Ung, W.K., Mok, K.M., 2012. Freshwater algal bloom prediction by support vector machine in macau storage reservoirs. Math. Probl. Eng. 2012.

Xiong, J., Lin, C., Ma, R., Cao, Z., 2019. Remote sensing estimation of lake total phosphorus concentration based on MODIS: a case study of Lake Hongze. Remote Sens. (Basel) 11 (17), 2068.

Yang, C., Everitt, J.H., Fernandez, C.J., 2010. Comparison of airborne multispectral and hyperspectral imagery for mapping cotton root rot. Biosyst. Eng. 107 (2), 131–139.

Yu, S., Jia, S., Xu, C., 2017. Convolutional neural networks for hyperspectral image classification. Neurocomputing 219, 88–98.

Zaherpour, J., Mount, N., Gosling, S.N., Dankers, R., Eisner, S., Gerten, D., Wada, Y., 2019. Exploring the value of machine learning for weighted multi-model combination of an ensemble of global hydrological models. Environ. Modell. Softw. 114, 112–128.

Zevenboom, W., Mur, L.R., 1984. Growth and photosynthetic response of the cyanobacterium Microcystis aeruginosa in relation to photoperiodicity and irradiance. Arch. Microbiol. 139 (2–3), 232–239.