Merged-LSTM and multistep prediction of daily chlorophyll-a concentration for algal bloom forecast

Merged-LSTM and multistep prediction of daily chlorophyll-a concentration for algal bloom forecast

H Cho[1] and H Park[1,2]
[1]Department of Civil and Environmental Engineering, KAIST, 34141 Daejeon, Republic of Korea

E-mail: hkpark@kaist.ac.kr

Abstract. Algal blooms are significant environmental problems which threaten the water supply system and ecosystem. To manage the problem, the effective forecasting model is necessary, but it is still challenging to predict the algal bloom due to its uncertainty and complexity. To improve the prediction performance, this study proposed the advanced model based on LSTM networks. Merged-LSTM model contains the three parallel LSTM layers and merged layers which is available to use the additional data from the diverse sources without problem in the training process. To predict the chlorophyll-a of target area, data from an additional monitoring station in upstream and auxiliary environmental data were put into parallel layers as well as data from the target area. The prediction result of the proposed model outperforms the existing models, and also shows a better training process with larger data dimensions. The proposed model and its result also suggest that the possibility of prediction of algal bloom with more advanced models and corresponding data sources.

1. Introduction

An algal bloom is one of the environmental hazards, which causes the degradation of water quality and ecological impact. Thus, this harmfulness becomes a significant threat to the safety of water resource management [1]. While the phenomenon has been widely noticed in many regions, the awareness has increased in South Korea recently. Social concerns increased after the construction of weirs in major rivers in Korea, and thus public systems to observe and predict the algal bloom has been built up additionally.

Essential ways to respond to algal bloom are prediction and early warning, but the process of occurrence is known to be complicated and uncertain [2]. Thus, many researchers have tried to predict the occurrence and intensity of algal bloom more accurately. Both numerical and statistical methods are used for prediction. Hydrodynamic modeling can be used as a numerical method to analyze and predict the algal bloom [3]. However, the machine learning approaches, as statistical methods, are widely applied for analyzing the observed data from the nonlinear and complex ecological process [4].

Machine learning methods to predict the algal bloom include Artificial Neural Network (ANN) [5,6], Random Forest model [7,8] and autoregressive model such as Autoregressive integrated moving average (ARIMA) [4,9]. Among them, ANN has been developed rapidly in recent days as computing power and algorithms were improved a lot [6]. Especially, Long Short-Term Memory (LSTM) network which is one of the types of recurrent ANN shows good forecasting power in time-series data [10] and also applied well to predict the algal bloom. [11,12].

However, there is still a limitation to applying the model in practice although many previous studies have been attempted to predict the algal bloom and achieved expected results. One of the challenges is that it is still hard to predict the values in multistep ahead compared to one-step prediction [12,13]. As the error accumulates and uncertainty increases, the multistep prediction is more difficult prediction tasks, and more improved model is needed for practical use.

The objective of this study is to suggest an improved model for predicting the algal bloom, especially in multistep forecasting task. As current LSTM model works well enough for a simple time-series data and delicate fine-tuning is not a promising way to practical usage, we tried to expand the model with more diverse data sources. Therefore, the model consists of three parallel layers from different data sources and merged sequence layers to get the multistep output values. It seems to expanded neural network model needs much larger computing power compared to previous models from the perspective of the conventional research environment. However, thanks to the breakthrough development of hardware and compatible programming packages, we could easily add the layers and other technical elements. The concentration of Chlorophyll-a (Chl-a) was used as indicator of the density of harmful algae and other water quality indices were put into model. The improved model of this study is expected to increase the possibility of practical application of machine learning-based model to algal bloom forecasting.

2. Material and Method

2.1 Study site: Geum-river

Geum-river is located in the mid-western part of Korea and a major freshwater resource in adjacent regions (Figure 1). However, Geum-river suffers from the algal bloom frequently, and safe water supply and the ecosystem of the river are threatened [14]. In the downstream below the Daecheong-dam, three weirs were installed through the restoration project to control the water level. Consequently, there are monitoring stations near the Daecheong-dam and three weirs to check the water quality. The regular measurements are conducted weekly including the environmental factors related to the algal bloom.

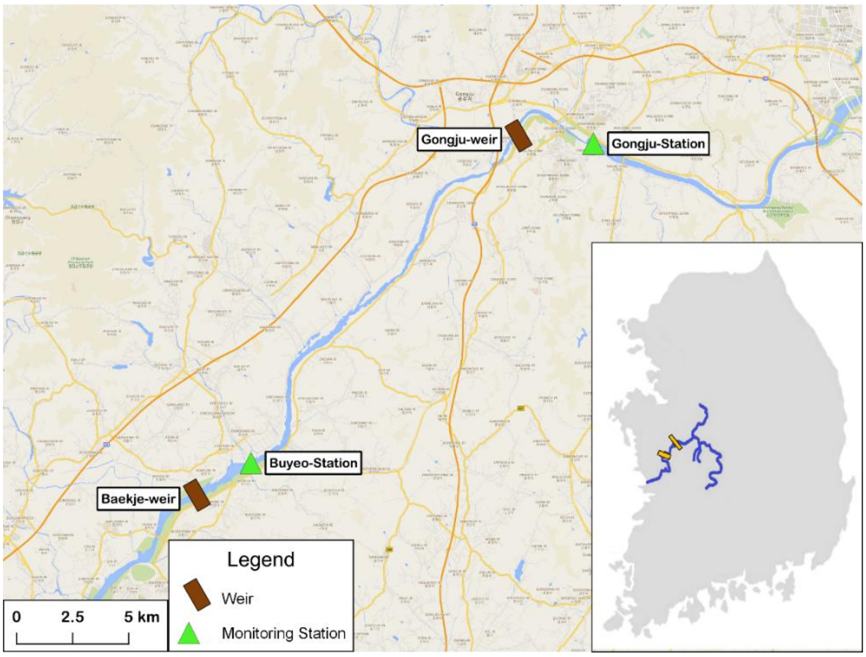

Figure 1. Map of Geum-river and study site.

On the other hand, we used a daily real-time monitoring data provided by the Ministry of Environment [15] in few reasons. First, daily data source provides a larger amount of data, and the sufficient amount of data is an essential condition for the machine learning model. Even the real-time monitoring data contain many missing values compared to regular measurement, the advantage of larger data is enough to compensate for shortcomings of instability. Secondly, the early warning and treatment for algal bloom should be more frequently rather than multi-weeks. When the forecast is conducted to a one step ahead prediction, the weekly data can also yield the practically meaningful result. In multistep tasks, however, it would be too broad and not precise in practical early warning.

Along the Geum-river, real-time monitoring stations are located near the weirs as seen in Figure 1. We selected the Baekje weir, located most downstream as a target area but got the data from both Gongju weir and Baekje weir. Then data as indicators for our model were selected based on previous studies among the 19 values measured from the monitoring station. The selected variables are concentration of chlorophyll-a, total nitrogen, total phosphorus, total organic carbon, dissolved oxygen, electrical conductivity, water temperature and pH [4-7,11,16]. Additionally, auxiliary inputs, known to affect the density of algal bloom, were used which are solar radiation, rainfall, flow rate, and air temperature from the data sources near the Buyeo station [6,17,18].

2.2 Data processing

Acquired data from the monitoring stations and auxiliary data sources were then processed to be used as input in the neural network model. One of the annoying problems in the practical application of the neural network is the presence of missing values. Basic neural network is not robust to missing values, and missing value inside of time-series data breaks the continuity of time-series. As seen in Figure 2, the original values of the Chl-a concentration contain a number of missing values. Moreover, these missing of variables did not occur at the same time. Consequently, some of the missing values should be filled with adequate methods.

First, the auxiliary data were simply interpolated because they rarely contain the missing value due to the stability of data sources. Next, the measured values from the monitoring stations except for the Chl-a in Buyeo station were replaced to -1. It is known as a robust method that put the 0 or -1 to missing value unless there were no 0 or -1 in original data [19]. Next, the whole data were reshaped to sequence data with a length of 7 days, the length of multistep in this study. Then, the sequences which contain the missing values of Chl-a of Buyeo station were deleted. It was available because the daily-based dataset was enough in volume to remove those missing value sequences.

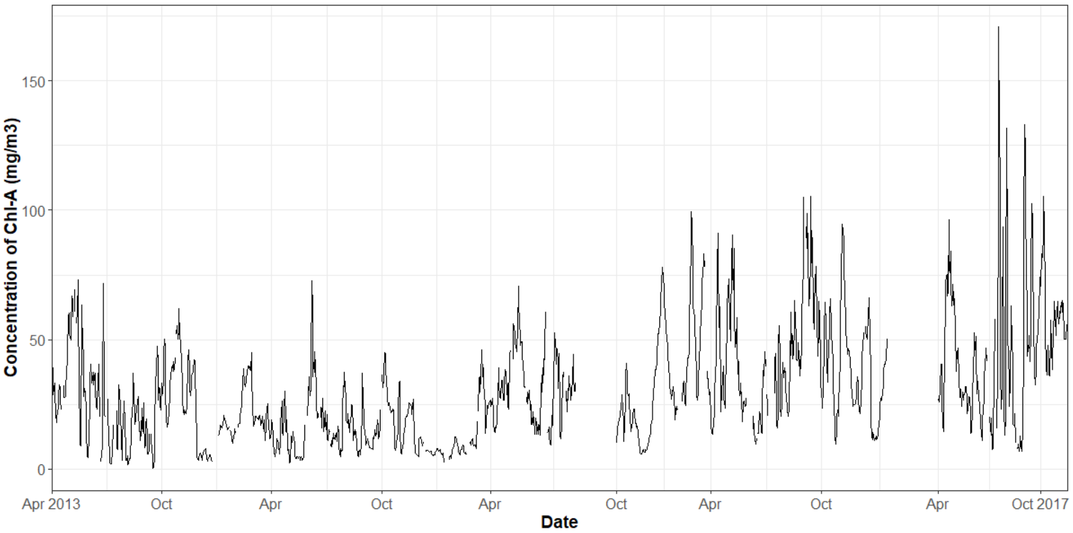

Figure 2. Concentration of Chlorophyll-a at Buyeo station.

The available data period is from April 2013, when real-time monitoring began, to now. However, the weir operation was stopped since the 2018, and hydrological condition was significantly changed compared to before. In this reason, the data used in this study was defined to before 2018. Then the 981 of data sequences with 7 days length were put into a model for train and test. In addition, the Chl-a concentration showed a different trend in 2017, and thus we did not divide the train and test set by the time order but using random selection among the whole dataset.

2.3 Merged-LSTM model

An improved LSTM model for multistep prediction of Chl-a concentration was proposed in this study (Figure 3). The essential point is using the parallel layers for each data source and then merging them later. This structure can be justified because it is clear that upstream data is closely related to downstream data as well as auxiliary data. Then parallel LSTM layers correspond to three data sources; Buyeo station, Gongju station, auxiliary data. Two LSTM layers were used for monitoring data while the one LSTM layer was used for auxiliary data that consists of only four variables. After parallel LSTMs, the merge layer adds the outputs of LSTMs and then, additional LSTM and Fully Connected (FC) layer are placed like common LSTM network models for the prediction task. For better performance of the neural network model, advanced techniques were applied to prevent the overfitting. Dropout [20] and Batch normalization [21] which are known to general improvement of neural network were added between the LSTM or FC layers.

While there is a spatial distance between two monitoring station, we ignored the time lagging between two data sequences. This assumption is because our model does not model the numerical hydrology but for nonlinear process in the overall environmental nears the target area. Furthermore, the development of a model which is robust to distance or time lag is more needed because it is too hard to do the fine tuning for every model along the river. Our proposed model was compiled using the Keras library and every LSTM layer was designated as CuDNNLSTM in Keras library which accelerates the training with using the Graphics Process Unit.

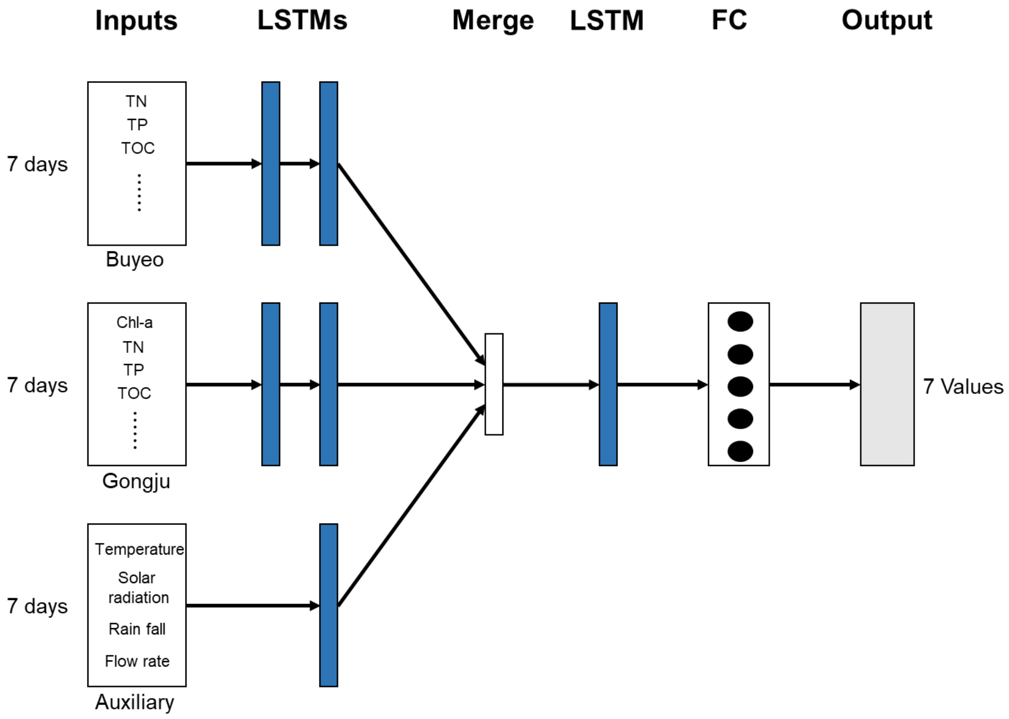

Figure 3. Structure of Merged-LSTM.

3. Results and discussion

The 7 days multistep prediction of Chl-a in Geum-river was conducted with our proposed model with a comparison of existing models. Total 981 of sequence data were divided into train set, validation set, and test set which are 60%, 20% and 20% of the total, respectively. The number of epochs was determined by observing the validation score during the training and early stop was done for preventing overfitting when the validation curve rises. The performance results were measured using the Root Mean Square Error (RMSE) of the test set.

Selected existing models for comparison are conventional LSTM and Multi-layer Perceptron (MLP). For multistep prediction task, MLP model showed the performance that is better than other linear models [13], and the model of multiple LSTM layers was also claimed to work well [12]. These comparable models have one sequential layers and then they use only the data from the Buyeo station. Conventional LSTM model consists of two LSTM layers with a dropout layer between them, and the one FC layer to make outputs. MLP model consists of three FC layers and two dropout layers between them.

The comparison result (Table 1) shows that the proposed Merged-LSTM model performed better than other existing models. This result can be explained because the Merged-LSTM could use more data sources and trained them in each parallel LSTM layers. Meanwhile, the proper techniques such as dropout and batch normalization prevented the overfitting or underfitting even we used a complex neural network model. Also, our results verified again that powerful performance of LSTM in time-series prediction compared to MLP. MLP is too simple structure to train the complex time-series task and reached a point of early stopping earlier than others.

On the other hand, the expanded parallel model with more LSTM and other neural network models increased the necessary computation power and time to train the model. Results also show that prediction performance is correlated with the training time per step. However, it is still less than a second per epoch and the whole training process is done around the ten minutes with GPU acceleration. Thus, the longer training time is not a significant obstacle to use the proposed model given the importance of this prediction task in water management.

Table 1. Prediction results of each model.

Model Description	Number of epochs	Training time (Sec per step)	Test Error (RMSE)
Merged-LSTM (proposed)	1500	0.577	0.0459
LSTM	1500	0.271	0.0500
LSTM without Dropout	1500	0.241	0.0661
MLP	500	0.134	0.0832

4. Conclusion

The forecast for algal bloom is a crucial task to manage the water resource in many regions. However, the precise prediction is still a challenging task, and existing models have limitation in applying for multistep prediction. Therefore the improvement of the prediction model is required for practical application for early warning of algal bloom. Since a monitoring point in the river correlates with upstream, we tried to add the data from the upstream as well as the auxiliary data from the nearby environment. Then, to make the neural network model combining these data from the diverse sources, this study proposed the merged-LSTM model which have the three parallel LSTM sequential layers for each data source and merge layer to yield the outputs. The performance of the proposed Merged-LSTM model outperforms the existing neural network models. This result showed that improvement could be achieved through more data source and extended neural network structure. This improvement suggests further improvement with using more available data sources and potential for practical application in forecast task.

Acknowledgments

This work has supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT)(No. 2018R1D1A1B07050208).

References

[1] Carmichael W W and Boyer G L 2016 Health impacts from cyanobacteria harmful algae blooms: Implications for the North American Great Lakes *Harmful algae* **54** 194-212

[2] Brooks B W, Lazorchak J M, Howard M D, Johnson M V V, Morton S L, Perkins D A, et al. 2016 Are harmful algal blooms becoming the greatest inland water quality threat to public health and aquatic ecosystems? *Environ. Toxicol. Chem.* **35** 6-13

[3] Bae S and Seo D 2018 Analysis and modeling of algal blooms in the Nakdong River, Korea. *Ecol. Model.* **372** 53-63

[4] Kim S 2016 A multiple process univariate model for the prediction of chlorophyll-a concentration in river systems *Int. J. Limnol.* **52** 137-50

[5] Lee G, Bae J, Lee S, Jang M and Park H 2016 Monthly chlorophyll-a prediction using neuro-genetic algorithm for water quality management in Lakes *Desalin. Water Treat.* **57** 26783-91

[6] Zhang F, Wang Y, Cao M, Sun X, Du Z, Liu R, *et al* 2016 Deep-Learning-Based Approach for Prediction of Algal Blooms *Sustainability* **8** 1060

[7] Yajima H and Derot J 2018 Application of the Random Forest model for chlorophyll-a forecasts in fresh and brackish water bodies in Japan, using multivariate long-term databases *J. Hydroinform.* **20** 206-20

[8] Shin J, Yoon S and Cha Y 2017 Prediction of cyanobacteria blooms in the lower Han River (South Korea) using ensemble learning algorithms *Desalin. Water Treat.* **84** 31-9

[9] Chen Q, Guan T, Yun L, Li R and Recknagel F 2015 Online forecasting chlorophyll a concentrations by an auto-regressive integrated moving average model: Feasibilities and potentials *Harmful Algae* **43** 58-65

[10] Hochreiter S and Schmidhuber J 1997 Long short-term memory *Neural computation* **9** 1735-80

[11] Lee S and Lee D 2018 Improved prediction of harmful algal blooms in four Major South Korea’s Rivers using deep learning models *Int. J. Env. Res. Public Hea.* **15** 1322

[12] Cho H, Choi U and Park H 2018 Deep learning application to time-series prediction of daily chlorophyll-a concentration *WIT Trans. Ecol. Environ.* **215** 157-63

[13] Du Z, Qin M, Zhang F and Liu R 2018 Multistep-ahead forecasting of chlorophyll a using a wavelet nonlinear autoregressive network *Knowl-Based Syst.* **160** 61-70

[14] Srivastava A, Ahn C Y, Asthana R K, Lee H G and Oh H M 2015 Status, alert system, and prediction of cyanobacterial bloom in South Korea *BioMed Res. Int.* **2015** 584696

[15] Ministry of Environment Real-Time Water Quality Information System, Available from: http://www.koreawqi.go.kr

[16] Cho K H, Kang J H, Ki S J, Park Y, Cha S M and Kim J H 2009 Determination of the optimal parameters in regression models for the prediction of chlorophyll-a: a case study of the Yeongsan Reservoir, Korea *Sci. Total Environ.* **407** 2536-45

[17] Cha Y, Cho K H, Lee H, Kang T and Kim J H 2017 The relative importance of water temperature and residence time in predicting cyanobacteria abundance in regulated rivers *Water Res.* **124** 11-9

[18] Yi H S, Park S, An K G and Kwak K C 2018 Algal Bloom Prediction Using Extreme Learning Machine Models at Artificial Weirs in the Nakdong River, Korea *Int. J. Env. Res. Public Hea.* **15** 2078

[19] Francois C 2017 *Deep learning with Python* (New York: Manning Publications Company)

[20] Srivastava N, Hinton G, Krizhevsky A, Sutskever I and Salakhutdinov R 2014 Dropout: A simple way to prevent neural networks from overfitting *J. Mach. Learn. Res.* **15** 1929-58

[21] Ioffe S and Szegedy C 2015 Batch normalization: Accelerating deep network training by reducing internal covariate shift *arXiv Preprint* (arXiv:150203167)