DEEP LEARNING APPLICATION TO TIME-SERIES PREDICTION OF DAILY CHLOROPHYLL-A CONCENTRATION

HYUNGMIN CHO[1], U-JIN CHOI[2] & HEEKYUNG PARK[1]
[1]Department of Civil and Environmental Engineering, KAIST, Republic of Korea
[2]Department of Mathematical Sciences, KAIST, Republic of Korea

ABSTRACT

Algal bloom in rivers is a major environmental concern which threatens the stable water supply and river ecosystem. Due to its complexity and nonlinearity, previous studies have tried various machine learning techniques to predict algal bloom. However, conventional approaches have limitations on predicting unobserved near future, and thus it is hard to apply to actual preparation policy. In this study, long short-term memory (LSTM), as a deep learning approach, is applied to predict the concentration of chlorophyll-a. Daily measured water quality information is used as input data and chlorophyll-a is used to output value for representing algal bloom. In addition to 1-day prediction, 4-days prediction task is attempted as sequence data prediction. As a result, LSTM network shows better performance, compared to the previous approaches, in predicting chlorophyll-a in 4-days prediction as well as 1-day prediction. In addition, the regularization methods are applied to model and batch normalization is proved to be a suitable way to improve accuracy. This result can lead to improvement in preventing algal bloom and also suggest various applications of deep learning methods in chlorophyll-a prediction task.

Keywords: algal bloom, chlorophyll-a, long short-term memory, LSTM.

1 INTRODUCTION

Algal blooms are one of the main concerns in water quality management since algae cause the problems of toxicity, taste, and odour in water resources. As preventive actions to reduce the damage from algal blooms, prediction of algal bloom and early warning are conducted in many cases [1]. Many previous studies applied data based machine learning methods to predict the algal bloom represented by the concentration of chlorophyll-a. For example, there have been applications of the artificial neural network (ANN) [1]–[3], support vector machine (SVM) [4] and Random Forest [4], [5]. However, complexity and non-linearity among the factors associated with algal blooms make it difficult to identify the process of algal bloom occurrence. Therefore more advanced prediction model is still required to develop.

Among them, the ANN has recently undergone rapid improvement with the appearance of deep learning which has deeper and advanced network layers [6]. Deeper layers and newly found regularization techniques such as dropout [7] showed significantly improved predictive accuracy for many fields. Although most of the deep learning approaches are highlighted in the field of image analysis or natural language process [6], recent studies showed that the deep learning can also improve the accuracy of predicting environmental problems including algal blooms [2] and air pollution [8]. Especially for sequence data, Long short-term memory (LSTM) network, a kind of recurrent neural net (RNN), is known to work well [8], [9]. By adding special units to avoid the long-term dependencies problems what the conventional RNN had, LSTM networks become the general algorithm used for sequential data problems in recent. Since most of the environmental monitoring data have sequential structures in times, LSTM is a suitable method to capture the temporal dynamicity [8].

WIT Transactions on Ecology and the Environment, Vol 215, © 2018 WIT Press
www.witpress.com, ISSN 1743-3541 (on-line)
doi:10.2495/EID180141

The objective of this study is to apply LSTM neural network to predict the chlorophyll-a concentration in the river and to suggest the guide to construct the model by comparing several multilayer models. Consequently, long-term prediction is also tried with LSTM networks, as well as conventional prediction task, to confirm the improvement by using deep learning approach. To obtain the results with better accuracy with deep learning based models for one point prediction and time-series prediction are the goal of this study.

2 MATERIAL AND METHOD

2.1 Study site and data

Geum River is one of the major rivers in South Korea which supplies water resource to Middle Western part of Korea. In Geum River, there are three weirs to control water flow in flood or drought seasons. On the other hand, in summer of Korea, algal blooms frequently occur in rivers and especially near weirs. Therefore, they are suitable locations for the study on the prediction of the concentration of chlorophyll-a. The selected site for this study is the Gongju observation station near the Gongju weirs which is in the middle of three weirs in the river (Fig. 1). A data source is Real-Time Water Quality Information System operated by the Ministry of Environment [10]. Although other regular observations in monthly or bi-weekly also provide water quality information, daily water quality data in Real-Time Water Quality Information System are more suitable for machine learning applications which need sufficient volume of data. Total available period of data is from April 2013 to October 2017. Fig. 2 shows that the concentration of chlorophyll-a is highly fluctuating value except for the winter period. Even the seasonal factors such as temperature are known to significant factor in the growth of the microorganism, the concentration of chlorophyll-a in study site had more complex behaviors.

The 19 environmental items are measured in the station, and input variables for models were selected based on previous studies. The nine selected indicators; water temperature, pH, dissolved oxygen (DO), electrical conductivity (EC), turbidity, total organic carbon (TOC), total nitrogen (TN), total phosphorus (TP), and chlorophyll-a concentration were found as effective predictive parameters in previous studies [1], [3]–[5], [11]–[13]. Excluded items from observation data are the concentration of volatile organic compounds (VOCs) such as trichloroethane and benzene which are measured because of their strong toxicity. Next, whole

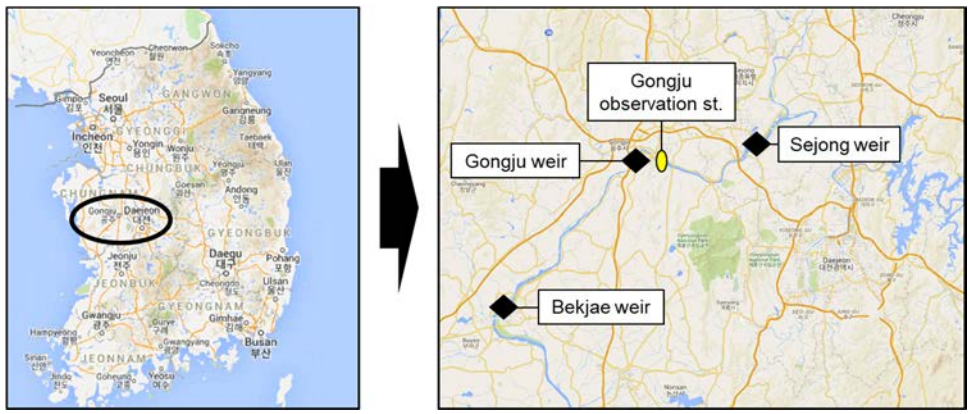

Figure 1: Study site: Geum River and Gongju observation station.

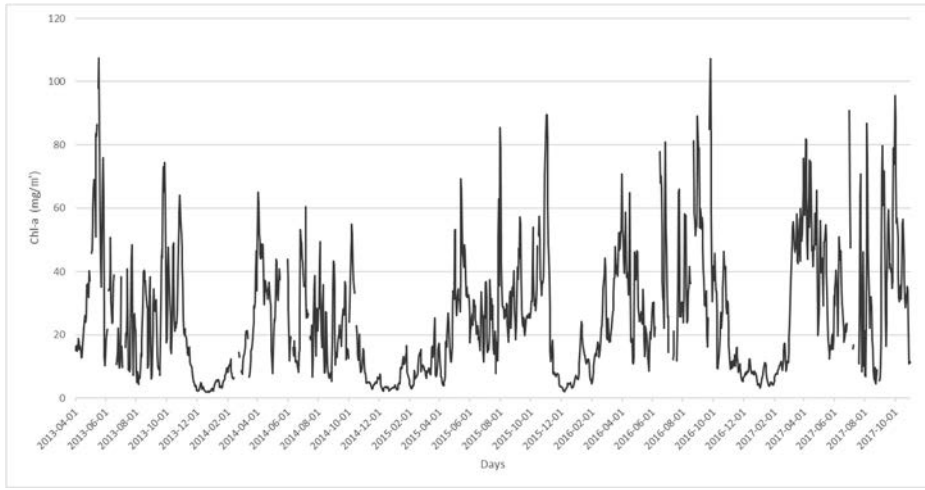

Figure 2: Concentration of chlorophyll-a in Gonju observation station.

data were preprocessed with a certain length of time sequence. One obstacle to process the data was the existence of missing values due to failure or inspection of measuring equipment. When creating the too long sequence, most of the sequence will have missing values and be unusable for the prediction model. Therefore, 688 sequence data, which are consist of 8-days observations were created for prediction by using the total observation period of 1675 days data.

2.2 Deep learning model and prediction task

Deep learning models in this study are made up of the combination of different layers and functions and then constructed into sequential layer model. Densely-connected (Dense) layer and Long Short-Term Memory (LSTM) layer were used as neural net layers which assigned as an input layer, output layer, and hidden layers. As activation function, Relu was used for dense layers and hyperbolic tangent was used for LSTM layers. Models used mean square error for loss function and selected Adam [14] as the optimizer. Additionally, dropout layer and batch normalization layer were applied to deep neural nets as regularizing layers since they are known to help network to avoid overfitting [7], [15]. All the component were based on the functions supported by Keras library [16] to make all neural network models and train them.

Fig. 3 shows a simple schematic of the deep neural network model to predict the time-series task. Models in this study are consist of three neural network layers; three Dense layers or two LSTM layers and one Dense layer. The number of layers was chosen by considering previous studies [1], [8] and computation cost. The number of training epochs is also an important parameter to performance and computation cost but it is also hard to define before training. Therefore, we divided the validation set from data to set the appropriate number of epochs which avoids too little or too much training.

Prediction models performed two tasks which are the 1-day prediction and 4-days prediction. The 1-day prediction was tested for not only comparing the different models but also checking the validity of data and neural networks to use. That is because tasks similar to 1-day prediction have been already tried by previous studies using conventional ANNs. After

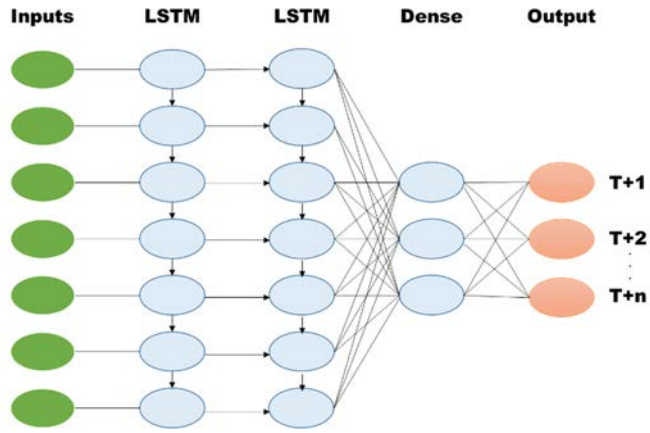

Figure 3: Simple schematic of deep learning abased predictive model.

verification of data and models, the performance of Dense networks and LSTM networks were compared. The 688 of 8-days sequence data with nine measured items were divided into 7-days as inputs and last 1-day as output. Thus, total 63 input variables were used to predict the 1 output value.

The 4-days prediction is more challenging tasks for neural networks because models need to predict more values and farther future with past data. Furthermore, same sequence data were used for both prediction tasks in this study because of many missing values in raw data and the smaller number of variables was available as input values. Therefore, more training epoch and training technique were tested to 4-days prediction task to achieve better performance. As first, we tried the advanced LSTM model, sequence-to-sequence LSTM, to compare the different structures of LSTM for difficult tasks. Sequence-to-sequence LSTM is a model known to be more effective in language translation, another sequence data processing task [17]. Dropout and batch normalization were also applied to networks as regularization methods and they enabled the model to train more epochs without overfitting problem. The 688 of 8-days data were divided into 4-days inputs and 4-days outputs for this task. Therefore, 36 variables were set as input values to predict four output values.

3 RESULT AND DISCUSSION

3.1 1-day prediction

The results of 1-day prediction are shown in Table 1. At first, the comparison between data arranged in chronological order and randomly shuffled data is for understanding the temporal distribution of data and the validity of data. A notable result is that models using randomly shuffled data show consistent higher accuracy than models using arranged in chronological order. Moreover, the large differences between the training error and the test error in arranged data models indicate that the models learned with the older data have low accuracy in predicting the latest data. In other words, there is a temporal bias in the data we used. It can be interpreted that external variables such as precipitation, the operation of dams and weirs affect the chlorophyll-a concentration in addition to the data we put as input variables. Models using shuffled data show fewer of these problems, all subsequent models use the shuffled data without notation of whether they are shuffled.

Table 1: Results of 1-day prediction models.

Model description	No of epochs	Training error (RMSE)	Test error (RMSE)
Arranged – 3 Dense	100	0.04875	0.08233
Arranged – 5 Dense	100	0.03969	0.10228
Arranged – LSTM (2 LSTM, 1 Dense)	100	0.04631	0.07928
Shuffled – 3 Dense	200	0.05364	0.06112
Shuffled – 5 Dense	200	0.03233	0.05870
Shuffled – LSTM (2 LSTM, 1 Dense)	**200**	**0.04968**	**0.04868**

The second thing to look at in the results is whether the number of layers for Dense networks is sufficient. The three Dense layers, which is suggested in the previous study is compared to five layers model. Increasing the number of layers reduce the test error, but there is a risk of overfitting as the difference from training error increases. Considering the complexity of the model and the increasing number of parameters, it can be seen that the 3-Dense layer model used in the previous study is a sufficiently applicable model. The third is a comparison of Dense networks and LSTM networks. Since Dense networks are known to be effective in previous studies, the LSTM networks showing better predictive power is also considered to be an effective model for predicting the chlorophyll-a concentration.

3.2 4-days prediction

The results of 4-days prediction in Table 2 show that the 4-days prediction is relatively difficult task than 1-day prediction in Table 1. When comparing prediction results by type of layer, simple LSTM models show the best performance. It is understood that LSTM networks perform better than Dense networks because the given tasks require handling sequence data in both input and output. The Sequence-to-sequence models, that showed better performance on the language translation tasks, show no special performance improvement for the 4-days chlorophyll-a prediction. However, when the length of the data sequence is extended or the data dimension is expanded, consideration of the advanced model will be necessary if the simple LSTM networks show limitations.

Table 2: Results of 4-days prediction models.

Model description	No of epochs	Test error (RMSE)
3 Dense	200	0.10009
3 Dense	600	0.10479
3 Dense + Dropout	600	0.12135
3 Dense + Batch Norm	600	0.11509
LSTM	200	0.09573
LSTM	600	0.09029
LSTM + Dropout	600	0.08911
LSTM + Batch Norm	**600**	**0.08015**
Seq2Seq	200	0.09954
Seq2Seq	600	0.09718
Seq2Seq + Dropout	600	0.09394
Seq2Seq + Batch Norm	600	0.08871

With increased training epoch, the performance of regularization is also compared. The results show that batch normalization is more effective than dropout to be applied inside the network models. This reconfirms that batch normalization improves the learning speed and accuracy of the neural network. Batch normalization will be an effective way for the field of the environment in which overfitting should be prevented with a limited number of data.

The predicted output of best performance model, LSTM with batch normalization, is presented in Fig. 4. The output shows acceptable prediction performance, but it seems that there is a limit to accurately predict high peaks. However, for preventive response to the algal bloom, it is important to predict the increasing trend rather than predicting the highest value accurately. Furthermore, considering the prediction task that using data of past 4-days to predict next 4-days is with considerable uncertainty, we expect further improvements in accuracy when we expand the amount of data and number of variables.

4 CONCLUSION

This study applied deep learning neural network models to predict the concentration of chlorophyll-a in the study site. The obtained data of variables associated with the chlorophyll-a concentration were sorted into sequential data and divided into 7-days input with 1-day output or 4-days input with 4-days output according to the prediction task. The 1-day prediction is a task to verify the applicability of models and the results of prediction also show acceptable accuracy. These results imply that neural networks with three Dense layers and three LSTM layers, which were suggested to use in this study are sufficiently applicable for prediction task. In addition, obtained data for this study were found to be required random shuffling due to the temporal bias among the data set. The objective of this study is to show that deep learning LSTM model has better performance in 4-days prediction than conventional neural network methods. The results show that LSTM network model achieves higher accuracy than Dense network model and batch normalization help the learning process as regularization method. Considering that 4-days prediction with data of past 4 days is a task with difficult conditions, the improved accuracy with using LSTM layers is a noticeable result. The deep learning neural network model of this study that is composed of LSTM layers is a more advanced method to predict the concentration of chlorophyll-a and the results of improved accuracy suggest that possibility of further advancement of the prediction model with various deep learning methods.

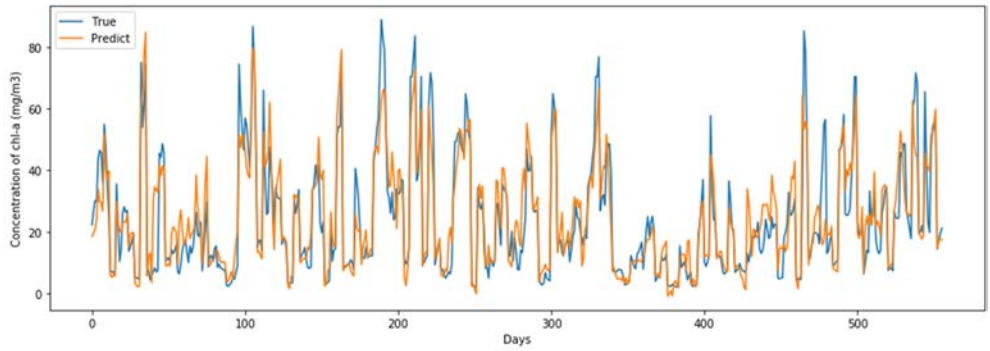

Figure 4: Predicted output of LSTM with batch normalization and true value.

ACKNOWLEDGEMENT

This work is financially supported by Korea Ministry of Land, Infrastructure and Transport (MOLIT) as U-City Master and Doctor Course Grant Program.

REFERENCES

[1] Lee, G., Bae, J., Lee, S., Jang, M. & Park, H., Monthly chlorophyll-a prediction using neuro-genetic algorithm for water quality management in Lakes. *Desalination and Water Treatment*, **57**, pp. 26783–26791, 2016.

[2] Zhang, F. et al., Deep-learning-based approach for prediction of algal blooms. *Sustainability*, **8**(10), p. 1060, 2016.

[3] Park, Y., Cho, K.H., Park, J., Cha, S.M. & Kim, J.H., Development of early-warning protocol for predicting chlorophyll-a concentration using machine learning models in freshwater and estuarine reservoirs, Korea. *Science of the Total Environment*, **502**, pp. 31–41, 2015.

[4] Li, X., Sha, J. & Wang, Z.-L., Application of feature selection and regression models for chlorophyll-a prediction in a shallow lake. *Environmental Science and Pollution Research*, pp. 1–11, 2018.

[5] Yajima H. & Derot, J., Application of the Random Forest model for chlorophyll-a forecasts in fresh and brackish water bodies in Japan, using multivariate long-term databases. *Journal of Hydroinformatics*, **20**, pp. 206–220, 2018.

[6] LeCun, Y., Bengio, Y. & Hinton, G., Deep learning. *Nature*, **521**, p. 436, 2015.

[7] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R., Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research,* **15**, pp. 1929–1958, 2014.

[8] Li, X. et al., Long short-term memory neural network for air pollutant concentration predictions: Method development and evaluation. *Environmental Pollution*, **231**, pp. 997–1004, 2017.

[9] Hochreiter, S. & Schmidhuber, J., Long short-term memory. *Neural Computation*, **9**, pp. 1735–1780, 1997.

[10] Ministry of Environment, Real-Time Water Quality Information System, Online. www.koreawqi.go.kr. Accessed on: 30 Jan. 2018.

[11] Lee, G., Othman, F., Ibrahim, S. & Jang, M., Determination of the forecasting-model parameters by statistical analysis for development of algae warning system. *Desalination and Water Treatment*, **57**, pp. 26773–26782, 2016.

[12] Lee, S., Lee, S., Kim, S.H., Park, H., Park, S. & Yum, K., Examination of critical factors related to summer chlorophyll a concentration in the Sueo Dam Reservoir, Republic of Korea. *Environmental Engineering Science*, **29**, pp. 502–510, 2012.

[13] Cho, K.H., Kang, J.-H., Ki, S.J., Park, Y., Cha, S.M. & Kim, J.H., Determination of the optimal parameters in regression models for the prediction of chlorophyll-a: A case study of the Yeongsan Reservoir, Korea. *Science of the Total Environment*, **407**, pp. 2536–2545, 2009.

[14] Kingma, D.P. & Ba, J., Adam: A method for stochastic optimization. Preprint, arXiv:1412.6980, 2014.

[15] Ioffe, S. & Szegedy, C., Batch normalization: Accelerating deep network training by reducing internal covariate shift. Preprint, arXiv:1502.03167, 2015.

[16] Keras, Online. https://keras.io/.

[17] Sutskever, I., Vinyals, O. & Le, Q.V., Sequence to sequence learning with neural networks. *Conference on Neural Information Processing Systems*, pp. 3104–3112, 2017.