J. Korean Soc. Environ. Eng., 45(4), 201-209, 2023
https://doi.org/10.4491/KSEE.2023.45.4.201
ISSN 1225-5025, e-ISSN 2383-7810

Research Paper

Comparison of Automated Machine Learning Model Performance for Predicting Chlorophyll-*a* Concentration according to Measurement Frequency of Input Data

Jungsu Park[1†]

Department of Civil and Environmental Engineering, Hanbat National University, Republic of Korea

(Received March 21, 2023; Revised April 7, 2023; Accepted April 7, 2023)

Objectives : Automated machine learning is a recent field of study that automates the process of machine learning model development including proper model selection and optimization. In this study, auto H2O, a novel automated machine learning algorithm, was used to develop a model to predict chlorophyll-*a* (chl-*a*).

Methods : This study used datasets with different observation frequencies of 1h, 2h, 8h, 24h and 1 week for the development of a machine learning model using an auto H2O algorithm to analyze the effects of measurement frequency of input data on model performance. The effect of the concentration of the input datasets on the performance of the model was also compared by building a model using datasets with observed values of chl-*a* exceeding 30 mg/m^3. The model performance was evaluated using three indices mean absolute error (MAE), Nash-Sutcliffe coefficient of efficiency (NSE) and root mean squared error-observation standard deviation ratio (RSR).

Results and Discussion : The MAE, NSE, and RSR of the model using the input data with a measurement frequency of 1h were analyzed as 0.8977, 0.9710, and 0.1704, respectively. The higher the measurement frequency of the input data, the better the performance of the model as the NSE of the model using full data was 0.9710, 0.9552, 0.8856, 0.8396, and 0.7509 for the input datasets with 1h, 2h, 8h, 24h and 1 week observation frequencies, respectively. The difference in model performance according to the difference in measurement frequency was larger for the model using data with the measured value of chl-*a* exceeding 30 mg/m^3, as the NSE was analyzed to be 0.8971, 0.8164, 0.5704, 0.5141, and 0.2052, respectively.

Conclusion : The auto H2O model for predicting chl-a showed better model performance as the measurement frequency of the input data increased, and the difference in performance according to the measurement frequency was larger in the range of observed chl-*a* concentrations that exceeded 30 mg/m^3.

Keywords : Algal bloom management, Automated machine learning, Machine learning, Model optimization, Water quality management

† Corresponding author
E-mail: parkjs@hanbat.ac.kr
Tel: 042-821-1265

J. Korean Soc. Environ. Eng., 45(4), 201-209, 2023
https://doi.org/10.4491/KSEE.2023.45.4.201
ISSN 1225-5025, **e-ISSN** 2383-7810

연구논문

입력 자료 측정빈도에 따른 클로로필-*a*농도 예측 자동 머신러닝 모형 성능 비교

박정수[1†]

국립한밭대학교 건설환경공학과

목적 : 자동 머신러닝은 모형의 선정부터 최적화까지 머신러닝 모형의 구축을 자동으로 수행해주는 최신 알고리즘으로, 본 연구에서는 자동 머신러닝(automated machine learning) 알고리즘 중 하나인 auto H2O를 이용하여 하천 클로로필-*a*(chl-*a*) 농도를 예측하는 모형을 구축하였다.

방법 : 본 연구에서는 1h, 2h, 8h, 24h 및 1 week 측정빈도로 구축된 입력 자료를 구축하여 입력 자료의 측정빈도가 auto H2O 알고리즘을 이용하여 구축된 자동 머신러닝 모형의 성능이 미치는 영향에 대한 분석을 수행하였다. 또한, 입력 자료의 농도가 모형 성능에 미치는 영향을 비교하기 위해 chl-*a* 실측값이 30 mg/m^{3}를 초과하는 자료로 구축된 모형과 성능 차이를 함께 비교하였다. 모형 성능은 mean absolute error (MAE), nash-sutcliffe coefficient of efficiency (NSE) 및 root mean squared error-observation standard deviation ration (RSR)의 3가지 지수를 이용하여 평가하였다.

결과 및 토의 : 측정빈도 1h의 입력 자료를 이용한 모형의 MAE, NSE, RSR이 각각 0.8977, 0.9710, 0.1704로 분석되었다. 전체자료를 이용할 경우 1h, 2h, 8h, 24h, 1 week 측정빈도에서의 NSE가 각각 0.9710, 0.9552, 0.8856, 0.8396, 0.7509로 분석되어 입력 자료의 측정빈도가 높을수록 모형의 성능이 좋은 경향을 확인하였다. Chl-*a* 실측값이 30 mg/m^{3}를 초과하는 경우 NSE가 각각 0.8971, 0.8164, 0.5704, 0.5141, 0.2052로 분석되어 전체자료를 이용하는 경우보다 상대적으로 측정빈도 차이에 따른 모형 성능 차이가 큰 것으로 분석되었다.

결론 : 자동 머신러닝 auto H2O 알고리즘을 이용하여 조류예측 모형을 구축하였으며 측정빈도가 높을수록 모형 성능이 좋으며 측정빈도에 따른 성능차이는 chl-*a*의 실측값이 30 mg/m^{3}인 구간에서 더 큰 것으로 분석되었다.

주제어 : 조류 발생 관리, 자동화 머신러닝, 머신러닝, 모형 최적화, 수질관리

1. 서 론

최근 수년간 물환경관리에 머신러닝 모형 등 고도화된 데이터 분석기술의 적용을 위한 노력이 빠르게 증가하고 있다. 인공신경망(artificial neural network), 서포트벡터머신(support vector machine), 의사결정나무(decision tree), 앙상블 머신러닝(ensemble machine learning), 순환신경망 LSTM(long short term memory) 등 다양한 머신러닝 알고리즘이 수질의 예측과 공정의 최적화 등 물환경관리 분야에 지속적으로 활용되고 있다.[1,2,3] 또한 최근에는 머신러닝 모형의 성능향상과 모형의 구현 결과에 대한 해석 등을 통해 머신러닝 모형의 실제 현장 적용성을 높이기 위한 연구도 활발히 이루어지고 있다.[4,5]

머신러닝 모형은 모형의 구축을 위해 활용된 입력 자료의 분석을 통해 미래 변화의 예측, 이상값의 결정, 자료의 분류(classification) 등 개발자가 원하는 결과를 도출하기 위한 다양한 내부 알고리즘으로 구성되어 있으며[6,7] 이러한 내부 알고리즘을 이용하여 입력된 자료간의 관계와 특성에 기반하여 목표로 하는 출력결과를 산정하게 된다. 예를 들어 의사결정나무에 기반한 대표적인 앙상블 머신러닝 알고리즘인 랜덤포레스트(random forest)의 경우 모형의 학습과정에서 활용되는 입력 자료에 따라 생성되는 의사결정나무의 수, 생성되는 나무구조의 깊이 등 최적의 결과를 도출하기 위해 필요한 모형의 내부구조를 결정하게 된다.[8,9]

머신러닝 모형의 구축을 위해서는 모형구축에 충분한 양의 자료확보가 필요하며, 우리나라에서는 최근 수년간 머신러닝 등 고도화된 데이터 분석기술의 적용을 확대하기 위해 필요한 데이터의 취득과 관리를 위한 다양한 노력을 지속하고 있다. 물환경분야에서도 센서 등을 이용한 실시간수질모니터링 등

현장 수질측정 자료의 확보를 위한 국가차원의 노력을 지속적으로 수행하고 있으며, 현재 전국에 70여개소의 수질자동측정망이 운영되고 있다.

입력 자료의 측정빈도는 머신러닝 모형의 구축 및 성능에 영향을 줄 수 있는 중요한 요소이다. 본 연구에서는 환경부 국립환경과학원 물환경정보시스템의 실시간 수질측정자료를 이용하여 취수원 수질의 안전성과 수생태에 다양한 영향을 미치는 하천 녹조발생 현황을 정량적으로 알 수 있는 대표적 수질 지표인 클로로필-*a*(chl-*a*)를 예측하는 머신러닝 모형을 구축하였다. Chl-*a*는 기상, 수질 등 다양한 환경인자가 영향을 미치게 되며, 예측을 위해서는 이러한 환경인자의 복합적인 영향을 고려하는 모형의 선정이 필요하다. 머신러닝 모형은 복잡한 비선형관계의 분석에 우수한 성능을 보여 chl-*a* 예측에 활용하기 위한 노력이 지속되고 있으며, 입력 자료의 특성을 고려하여 다양한 머신러닝 모형 중 주어진 자료에 적합한 최적 모형을 선정하는 것이 필요하다. 자동 머신러닝(automated machine learning)은 심층신경망 등을 포함한 다양한 머신러닝 알고리즘을 내부 모형으로 포함하고, 이러한 내부 모형들의 조합을 통해 입력 자료의 특성에 맞는 최적 모형을 구축하는 알고리즘으로 최근 관련분야의 연구가 활발히 이루어지고 있다. 본 연구에서는 자동 머신러닝 auto H2O 알고리즘을 이용하여 chl-*a*를 예측하는 모형을 구축하였다.[10,11] Auto H2O는 대표적인 자동 머신러닝 알고리즘의 하나로, 본 연구에서는 다양한 측정빈도를 가지는 입력 자료를 활용하여 입력 자료의 측정빈도가 자동 머신러닝 알고리즘을 이용해 구축되는 머신러닝 모형의 성능과 최적화에 미치는 영향을 분석하였다. 또한 각각의 측정빈도에 대하여 auto H2O 모형의 구축시간을 다르게 구성하여 모형 구축시간이 머신러닝 모형의 성능에 미치는 영향을 비교하였다.

2. 실험 방법

2.1. 입력 자료

본 연구에서는 환경부 자동측정망 갑천지점(Site No. S03002)의 수질측정자료를 이용하여 머신러닝 구축에 활용하였다(**Fig. 1**).[12] 갑천의 유역면적은 649 km^2이며 대도시인 대전광역시를 흘러 대청댐 하류와 합류하여 금강 본류로 유입되는 금강 제1지류로 지속적인 수질관리가 중요한 하천이다.[13]

모형의 구축을 위해 2013년 1월1일부터 2021년 12월31일까지 갑천지점에서 측정된 1시간(1h) 측정빈도 수질측정자료를 활용하였으며 수소이온농도(pH), 수온(TEMP), 용존산소농도(DO), 전기전도도(EC), 총유기탄소(TOC), 탁도(TURB)를 독립변수로 하고 조류 발생정도를 정량화하는 대표적인 지표인 chl-*a*를 예측의 대상이 종속변수로 구성하였다. 모형구축에 사용된 자료중 2013년 1월1일 00:00~2018년 12월31일 23:00 까지의 자료를 모형의 학습(training)에 2019년 1월1일 00:00~2021년 12월31일 23:00까지의 자료를 학습된 모형의 평가(testing)에 활용하여 전체자료의 67%를 모형의 training에 전체자료의 33%를 모형의 testing에 사용하였다. 모형의 구축에 사용된 1h 측정빈도의 자료는 측정항목에 따라 13~19%의 결측구간을 포함하고 있으나, 결측이 발생하는 기간이 대부분 수질의 변동이 크지 않은 구간으로 결측이 발생된 측정일에서 가장 가까운 k개의 측정자료를 이용하여 결측값을 보정하는 K-Nearest Neighbor(KNN) 알고리즘을 이용하여 결측값에 대한 보정을 수행하였다. KNN의 수행은 python open source library인 scikit-learn을 이용하였으며 k=5를 적용하였다.[14]

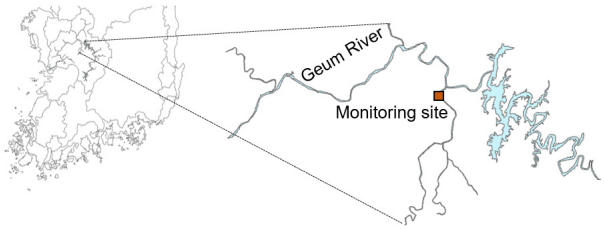

Fig. 1. Research site.

모형구축에 사용된 입력 자료의 측정빈도가 모형이 성능과 최적화에 미치는 영향을 분석하기 위해 1h 측정빈도자료의 평균값을 이용하여 2h, 8h, 24h 및 일주일(1w) 측정빈도의 자료를 구성하고, 모형의 예측성능을 높이기 위해 각각의 측정빈도의 자료에 대해 t-1의 차분을 적용한 값을 독립변수 추가하여 모형의 구축에 활용하였다.

2.2. 모형구축

본 연구에서는 최근에 개발된 open source library인 auto H2O를 이용하여 머신러닝 모형을 구축하였다. Auto H2O는 데이터의 전처리와 모형의 선정 및 최적화를 자동으로 구현해주는 대표적인 자동 머신러닝 알고리즘중 하나로 개별모형으로 좋은 성능을 보여 널리 활용되는 머신러닝 알고리즘인 XGBoost(XGB), gradient boosting machine learning(GBML), random forest(RF), generalized linear models(GLM) 및 deep neural networks를 base learner로 포함하고 있다. Auto H2O는 입력 자료를 이용하여 내부 알고리즘으로 포함하고 있는 다양

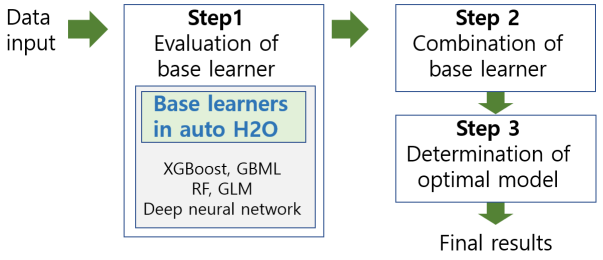

Fig. 2. Schematic of auto H2O model optimization process.

한 base learner의 성능을 평가하고 이러한 base learner의 조합(combination)을 통해 최적의 예측성능을 가지는 모형을 구축하게 된다(**Fig. 2**).[10]

Auto H2O는 모형 성능의 최적화를 위해 소요되는 최대 시간을 설정할 수 있으며, 본 연구에서는 입력 자료의 측정빈도에 따른 모형성능변화를 비교하기 위해 1h, 2h, 8h, 24h 및 1 w(week)의 측정빈도 자료각각에 대하여 1h의 최적화시간을 적용하여 모형을 구축하고 그성능을 비교하였다. 또한 chl-*a*의 실측값 약 상위 10%에 해당되는 30 mg/m^{3}를 초과하는 구간의 자료를 이용하여 종속변수 값이 높은 자료를 이용하는 경우 모형의 성능에 미치는 영향을 분석하였다. Chl-*a* 30 mg/m^{3}를 초과하는 실측값의 비율은 1h, 2h, 8h, 24h, 1w 측정빈도 자료에 대하여 각각 9.7%, 9.7%, 9.7%, 9.8%, 8.3%를 차지하였다.

머신러닝 모형은 입력 자료간의 물리적 혹은 화학적 관계 등을 사전에 분석할 필요가 적어 입력 자료의 구축이 완료되면 상대적으로 빠른 시간에 모형의 구축을 수행할 수 있다. 하지만, 모형의 성능을 최적화하기 위해서는 모형의 구조 혹은 세부 구현 내용을 결정하기 위한 다양한 hyperparameter의 구성에 따른 모형의 성능을 확인해 모형의 성능을 높이기 위한 최적의 hyperparameter 조합을 구하는 과정이 필요하다. Auto H2O는 모형의 선정과 최적화 과정에 분석자의 개입을 최소화하면서 자동으로 모형의 최적화를 수행할 수 있는 내부 알고리즘을 가지고 있으며, 최적화에 소요되는 최대 시간을 설정할 수 있도록 구성되어 있다. 본 연구에서는 모형의 최적화에 소요되는 시간이 모형의 성능에 미치는 영향을 파악하기 위해 동일한 측정빈도별 입력 자료에 대해 60s, 180s, 300s, 900s, 1,800s, 3,600s의 최적화 시간을 적용하여 모형을 구축하여, 입력 자료의 측정빈도 및 모형 최적화 소요시간이 모형 성능에 미치는 영향을 함께 비교하였다.

Table 1. Indices for model evaluation.

Index	Equation	Range
MAE	$MAE = \frac{1}{n}\sum_{i-1}^{n}\left\|y_t - \hat{y_t}\right\|$	0~∞
NSE	$NSE = 1 - \frac{\sum_{t=1}^{n}(\hat{y_t} - y_t)^2}{\sum_{t=1}^{n}(y_t - \overline{y_t})^2}$	-∞~1
RSR	$RSR = \frac{\sqrt{\sum_{t=1}^{n}(y_t - \hat{y_t})^2}}{\sqrt{\sum_{t=1}^{n}(y_t - \overline{y_t})^2}}$	0~1

2.3. 모형 성능 평가

모형의 성능 평가에 널리 활용되는 지수인 mean absolute error (MAE), nash-sutcliffe coefficient of efficiency (NSE) 및

Table 2. Characteristics of input variables with various observation frequencies between January 2013 and December 2021.

	Variables	TEMP (℃)	pH	EC (μS/cm)	DO (mg/L)	TURB (NTU)	TOC (mg/L)	Chl-*a* (mg/㎥)
1h	Average	18.0	7.0	420.3	7.6	5.2	3.9	12.9
	standard deviation	7.5	0.4	85.9	3.0	11.9	1.1	14.0
	Min	1.0	6.2	108.0	0.5	0.0	0.8	0.8
	Max	34.4	9.2	763.0	18.5	388.9	20.7	279.2
2h	Average	18.0	7.0	420.3	7.6	5.2	3.9	12.9
	standard deviation	7.5	0.4	85.8	3.0	11.7	1.1	13.9
	Min	1.1	6.2	109.5	0.5	0.0	0.8	0.9
	Max	34.3	9.2	761.5	18.0	380.0	18.6	191.4
8h	Average	18.0	7.0	420.3	7.6	5.2	3.9	12.9
	standard deviation	7.5	0.4	85.4	2.8	11.3	1.0	13.7
	Min	1.5	6.2	113.5	0.6	0.0	1.1	0.9
	Max	33.9	9.1	745.5	16.8	333.5	15.6	144.7
24h	Average	18.0	7.0	420.3	7.6	5.2	3.9	12.9
	standard deviation	7.5	0.3	83.9	2.4	10.4	1.0	13.2
	Min	2.4	6.3	113.6	0.9	0.1	1.2	1.2
	Max	33.8	9.0	733.3	16.2	182.1	11.9	140.5
1w	Average	18.0	7.0	420.3	7.6	5.2	3.9	12.9
	standard deviation	7.4	0.3	74.5	2.2	7.4	0.9	12.1
	Min	3.9	6.5	209.7	2.0	0.2	1.3	1.8
	Max	31.8	8.1	624.0	13.1	57.8	7.1	91.2

root mean squared error-observation standard deviation ration (RSR)을 이용하여 서로 다른 측정빈도를 가진 입력 자료로 구축된 Auto H2O 모형의 성능 평가를 수행하였다(**Table 1**). **Table 1**에 제시된 공식의 y_t는 시간 t에서의 실측값을, $\overline{y_t}$는 실측값의 평균, $\hat{y}_t$은 시간 t에서의 모형의 예측값을, n은 실측을 수행한 횟수이다. MAE는 실측값과 각 실측값에 대응되는 모형의 예측값의 차이의 총합을 실측된 자료의 수로 나누어준 값으로 모형이 실측값을 잘 예측할수록 그 차이가 작아져서 작은 값을 가지게 된다. RSR은 0에 가까울수록 NSE는 1에 가까울수록 모형이 실측값을 잘 예측한 것으로 평가한다.[15,16]

3. 결과 및 고찰

3.1. 모형구축 자료 현황

모형의 구축에 사용된 1h, 2h, 8h, 24h, 1w의 5가지 측정빈도를 가지는 입력 자료 각각의 평균, 표준편차, 최소값 및 최대값을 **Table 2**에 제시하였다. 모든 측정빈도에서 평균값은 동일하나, 입력 자료의 측정빈도가 낮아질수록 최대값과 최소값은 감소하게 된다. **Fig. 3**은 모형의 training 및 testing에 활용된 종속변수인 Chl-*a*의 측정빈도에 따른 차이를 보여주고 있으며, 측정빈도가 낮아질수록 최대값이 감소하는 경향을 시각적으로 확인할 수 있다.

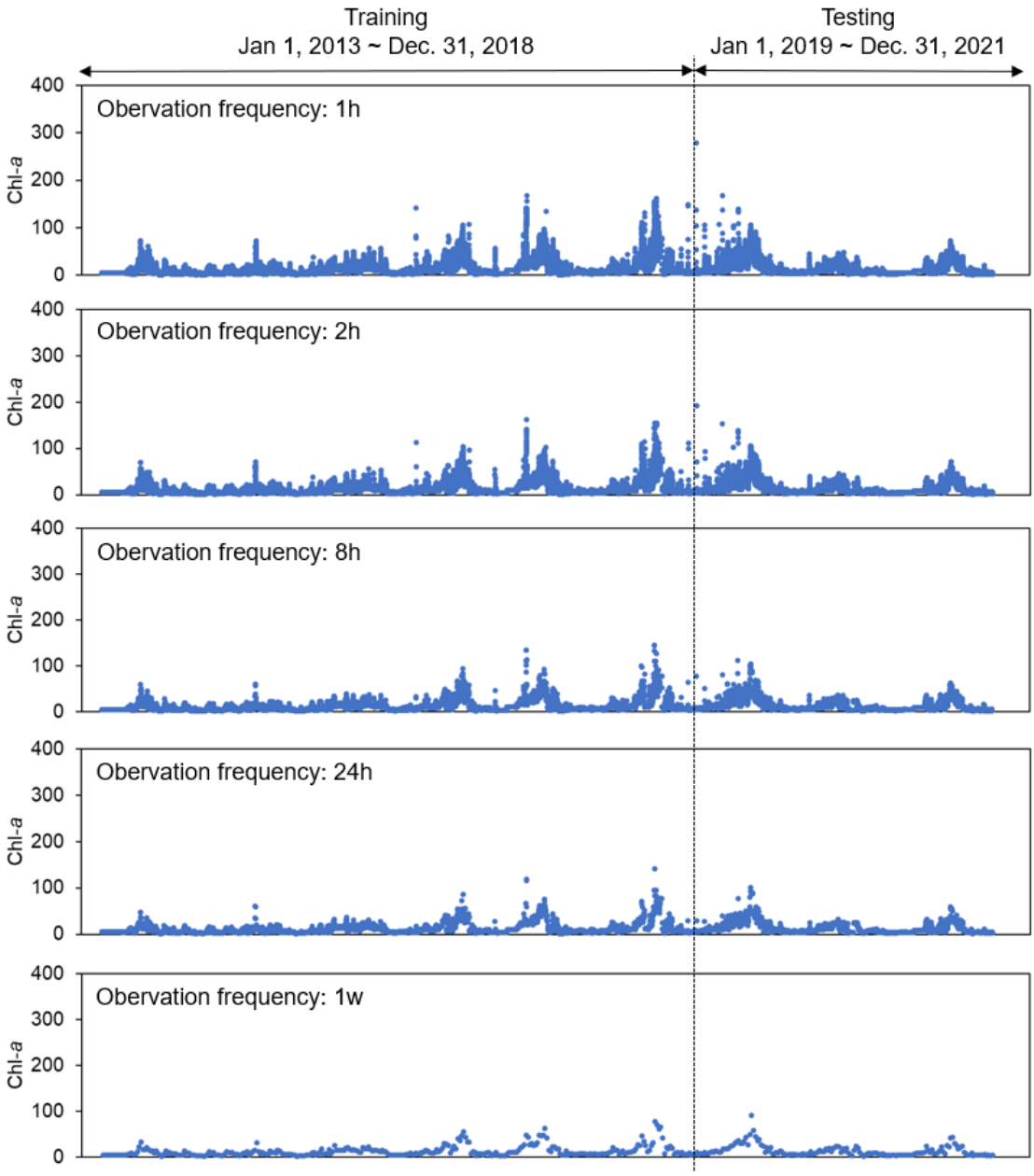

Fig. 3. Dependent variable used for training and testing of the model.

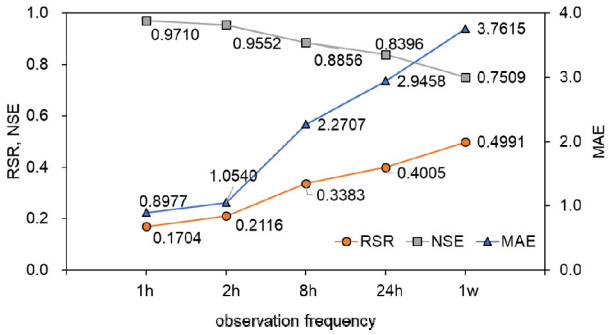

Fig. 4. Comparison of model evaluation results with different observation time frequencies.

3.2. 입력 자료 특성에 따른 모형 성능 분석

Auto H2O를 이용하여 1시간(3,600s)의 최적화 시간을 통해 구축된 모형의 chl-*a* 예측성능평가를 위해 사용한 3가지 지수 MAE, NSE, RSR의 각 측정자료 빈도별 변화를 **Fig. 4**에 제시하였다. 입력 자료의 측정빈도가 1h, 2h, 8h, 24h, 1w인 경우에 대하여 각각 MAE 0.8977, 1.0540, 2.2707, 2.9458, 3.7615, NSE 0.9710, 0.9552, 0.8856, 0.8396 및 0.7509, RSR 0.1704, 0.212, 0.338, 0.401 및 0.4991로 분석되어 3가지 지수 모두 모형의 구축에 사용되는 입력 자료의 측정빈도가 높을수록 모형이 좋은 성능을 보이는 경향을 보여주었다. 현장에서 자료 취득을 위한 측정빈도의 결정은 자료 수집의 목적에 따라 달라지며, 일반적으로 측정빈도가 높을수록 많은 비용과 인력이 소요된다. 머신러닝 모형은 취득된 자료를 기반으로 모형을 구축하게 되며, 모형의 목적에 따라 1w 측정빈도의 성능으로 충분한 경우가 있거나 1h 측정빈도의 성능이 필요한 경우가 있을 수 있는 등 상황에 따라 필요한 적정 성능이 달라질 수 있으며, 필요한 성능을 고려한 적정 빈도 입력 자료의 선택으로 자료 취득에 요구되는 비용과 인력 등의 합리적 운용이 가능할 것으로 판단된다.

Fig. 5는 모형의 구축에 사용된 입력 자료의 측정빈도에 따른 chl-*a*의 실측값과 모형의 예측값의 변화를 비교하여 보여주고 있다. 측정빈도가 낮은 경우 실측값의 최대값과 최소값이 감소하게 되며, 측정빈도가 낮을수록 모형의 성능이 낮아지기는 하나 전체적으로 모형이 실측값을 안정적으로 예측하는 경향을 확인할 수 있었다.

본 연구에서는 또한 모형의 종속변수의 값이 높은 구간의 자료가 모형의 성능에 미치는 영향을 분석하기 위해 chl-*a*의 실측값이 30 mg/m^{3}를 초과하는 구간에 대한 모형의 성능을

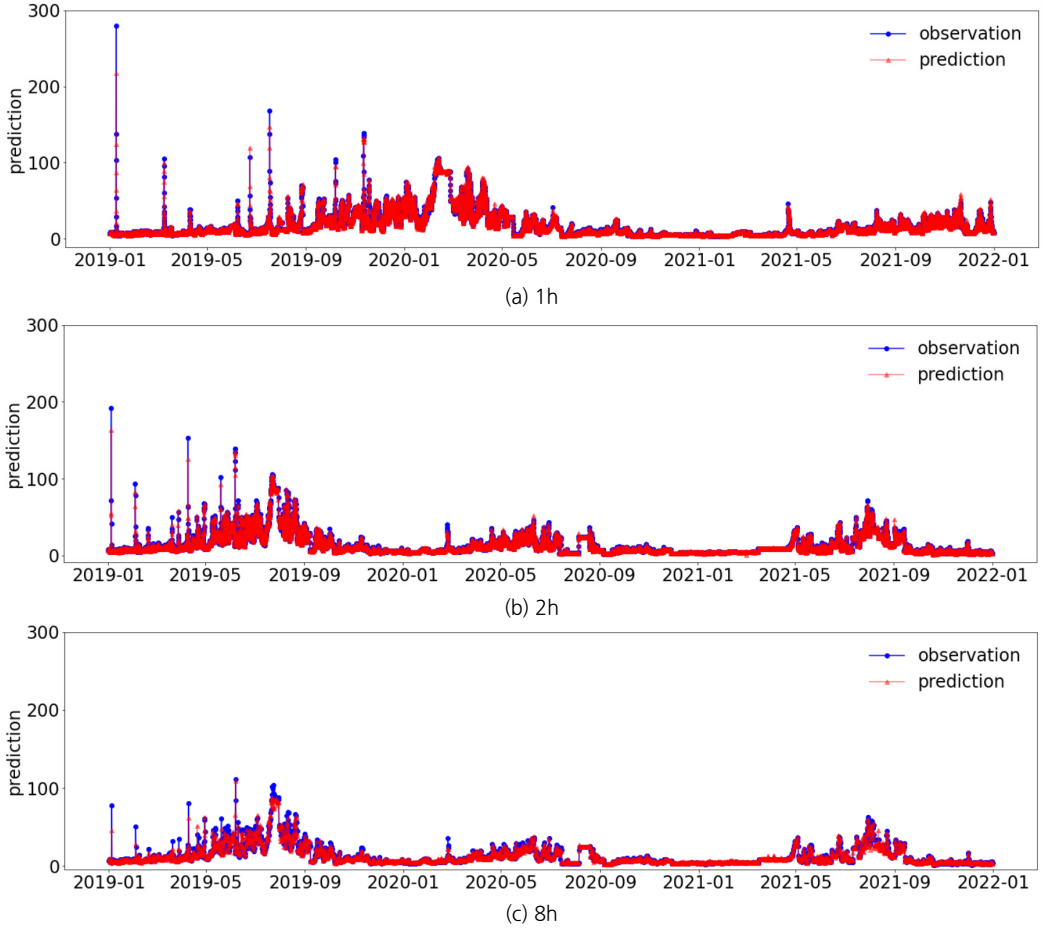

Fig. 5. Comparison of model simulation results with different observation time frequency (Continued).

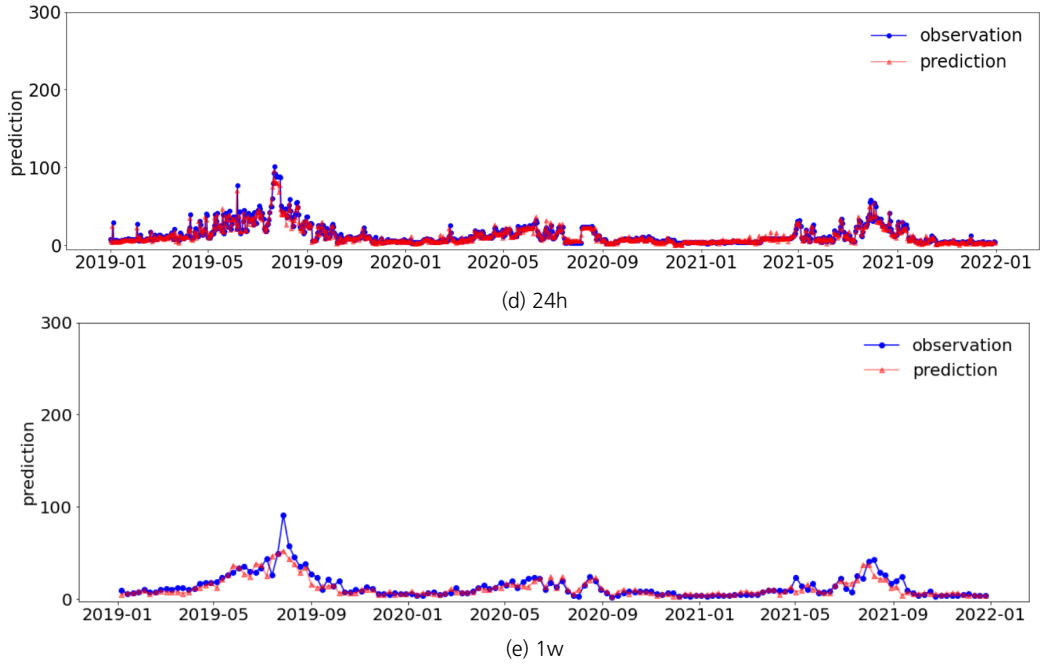

Fig. 5. Comparison of model simulation results with different observation time frequency.

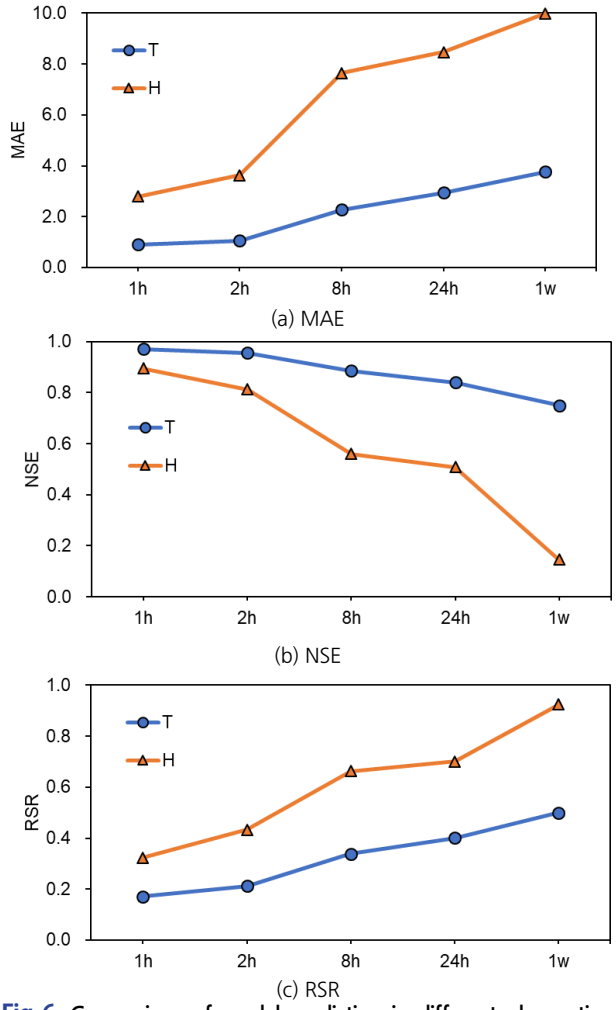

Fig. 6. Comparison of model prediction in different observation range.

별도로 분석하여 전체자료를 이용하여 구축한 모형의 성능과 비교를 수행하였다. **Fig. 6**은 전체자료를 이용한 경우의 모형 성능(T)과 chl-*a*의 실측값이 30 mg/m^{3}를 초과하는 경우의(H) 모형 성능을 비교하여 보여주고 있다. 실측값이 높은 구간(H)의 자료도 전제 자료를 이용한 경우와 마찬가지로 측정빈도가 낮을수록 모형의 성능이 낮아지는 경향을 보여주었다. 하지만 전체자료를 이용할 경우 1h 측정빈도 자료와 1w 측정빈도 자료의 NSE가 각각 0.971과 0.751이었으며, chl-*a*가 높은 구간에 대한 자료에 대해서는 1h 측정빈도 자료와, 1w 측정빈도 자료의 NSE가 각각 0.8955와 0.1460으로 분석되는 등 실측값이 높은 구간의 자료에 대해서 모형의 성능저하가 더 커지는 경향을 확인할 수 있었다.

머신러닝 모형은 입력 자료의 구성과 특성이 모형 성능에 많은 영향을 미치게 되며, 모형의 구축시 입력 자료의 특성을 고려한 모형의 선정 및 전처리 등이 중요하다. 본 연구에서는 입력 자료의 측정빈도와 농도가 chl-*a* 예측을 위한 자동 머신러닝 모형의 성능에 미치는 영향을 분석하여 제시하였다. 이러한 모형의 특징은 예측 대상 항목 등에 따라 다양한 형태로 나타날 수 있으며 향후 다양한 대상 항목 등을 대상으로 하는 지속적인 연구를 통해 수질관리 분야에서 머신러닝 모형의 활용성을 높일 수 있을 것으로 판단된다.

3.3. 모형 최적화 시간의 영향 분석

본 연구에서는 입력 자료의 측정빈도와 함께 모형의 최적화에 소요되는 시간이 모형의 성능에 미치는 영향을 비교하기 위해 1h, 2h, 8h, 24h, 1w 측정빈도의 입력 자료 각각에 대하여 60s, 180s, 300s, 900s, 1,800s, 3,600s의 최적화 시간을 적용하

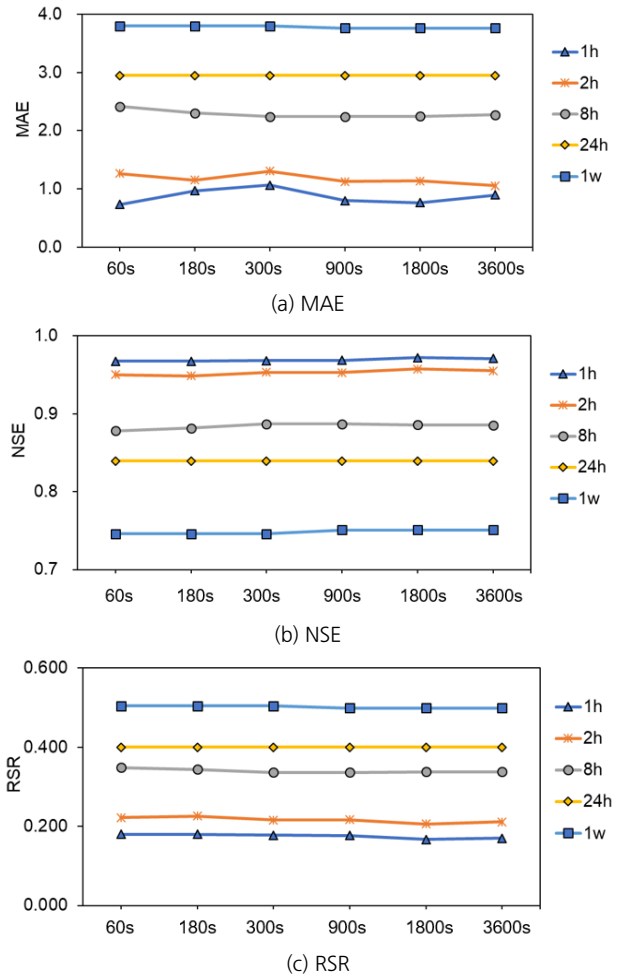

Fig. 7. Model performance in different model optimization time.

여 구축된 모형의 성능에 대한 비교분석을 수행하였다(**Fig. 7**). MAE의 경우 1h, 2h 측정빈도 자료가 모형의 최적화 시간에 따라 성능의 증감이 다소 있는 것으로 분석되었으나 그 변화는 크지 않았으며, 전체적으로 모형의 최적화시간에 따른 성능의 차이는 측정빈도에 따른 차이에 비해 크지 않은 것으로 분석되었다. NSE와 RSR도 MAE와 마찬가지로 모든 최적화 시간대에 대하여 측정빈도가 높을수록 좋은 성능을 보이는 경향을 확인할 수 있었으나, 최적화 시간에 따른 차이는 크지 않은 것으로 분석되었다.

이러한 분석결과는 본 연구에 사용된 입력 자료의 특성이 반영된 사항으로 모든 자료에 일반화하는 것은 한계가 있다. 하지만, 모형의 training에 사용된 자료는 1h, 2h, 8h, 24h 및 1w 측정빈도에 대하여 각각 52,584회 26,292회, 6,573회, 2,191회, 313회의 측정값을 포함하고 있으며 특히 1h, 2h 측정빈도 자료의 경우 물환경분야에서 활용되는 입력 자료로는 작지 않은 관측 횟수를 가지는 자료로, 이러한 다량의 자료에 대해서도 auto H2O를 통해 상대적으로 짧은 시간에도 충분한 성능을 가지는 모형의 최적화가 가능함을 확인할 수 있었다.

4. 결 론

본 연구에서는 모형의 선정부터 최적화까지 머신러닝 구축 과정을 자동으로 수행하는 최신 자동 머신러닝 알고리즘인 auto H2O를 이용하여 chl-*a*를 예측하는 머신러닝 모형을 구축하여 물환경분야에 자동 머신러닝 알고리즘의 적용가능성을 확인하였다. 또한, 하천에서 측정된 1h, 2h, 8h, 24h, 1w의 측정빈도를 가진 입력 자료를 활용하여 입력 자료의 측정빈도가 자동 머신러닝 알고리즘을 이용하여 구축된 머신러닝 모형의 최적화 및 성능에 미치는 영향을 확인하였다. 3,600s의 최적화 시간을 적용하여 구축된 auto H2O 모형의 1h 측정빈도 자료에 대한 MAE, NSE, RSR이 각각 0.8977, 0.9710, 0.1704로 분석되었으며, 측정빈도가 높을수록 모형의 성능이 좋고 측정빈도가 낮을수록 모형의 성능이 낮아지는 경향을 확인하였으며 이러한 측정빈도에 따른 성능의 차이는 chl-*a* 실측값이 높은 자료로 구축된 모형에서 더 큰 것으로 확인되었다. 또한, auto H2O 모형의 최적화에 소요되는 최대 시간이 모형의 성능에 미치는 영향을 분석하였으며 60s, 180s, 300s, 900s, 1,800s, 3,600s의 최적화 시간을 적용하여 모형의 성능을 비교한 결과 최적화 시간과 상관없이 입력 자료의 측정빈도에 따른 성능은 명확한 차이가 있음을 확인할 수 있었으나 모형 최적화에 소요되는 시간에 따른 성능의 차이는 크지 않아, 상대적으로 길지 않은 최적화 시간을 적용하더라도 모형이 충분한 성능을 얻을 수도 있음을 확인할 수 있었다. 본 연구에서는 측정빈도 및 농도와 같은 입력 자료의 특성이 자동 머신러닝 auto H2O 알고리즘의 성능에 미치는 영향을 분석하여 제시하였으며, 향후 다양한 입력 자료를 활용한 지속적인 연구를 통해 자동 머신러닝 모형의 현장 적용성을 높일 수 있을 것으로 판단된다.

Acknowledgement

본 결과물은 환경부의 재원으로 한국환경산업기술원의 수생태계 건강성 확보 기술개발사업의 지원을 받아 연구되었습니다(2020003030006).

References

1. J. Derot, H. Yajima, S. Jacquet, Advances in forecasting harmful algal blooms using machine learning models: A case study with Planktothrix rubescens in Lake Geneva, Harmful Algae, 99, 101906(2020).
2. P. Yu, R. Gao, D. Zhang, Z. P. Liu, Predicting coastal algal blooms with environmental factors by machine learning methods, Ecological Indicators, 123(4), 107334(2021).

3. J. Wen, J. Yang, Y. Li, L. Gao, Harmful algal bloom warning based on machine learning in maritime site monitoring, Knowl.-Based Syst., 245, 108569(2022).
4. Y. Shin, T. Kim, S. Hong, S. Lee, E. Lee, S. Hong, C. Lee, T. Kim, M. S. Park, J. Park, T. Y. Heo, Prediction of chlorophyll-a concentrations in the Nakdong River using machine learning methods, Water, 12(6), 1822(2020).
5. N. Hellen, G. Marvin, Explainable AI for safe water evaluation for public health in urban settings, in 2022 International Conference on Innovations in Science, Engineering and Technology, pp. 1-6(2022).
6. W. Y. Loh, Classification and regression trees, Wiley interdiscip. Rev.: Data Min. and Knowl. Discov., 1(1), 14-23(2011).
7. D. Maulud, A. M. Abdulazeez, A review on linear regression comprehensive in machine learning, J. Appl. Sci. Technol. Trends, 1(4), 140-147(2020).
8. L. Breiman, Random forests, Mach. learn., 45(1), 5-32(2001).
9. G. Biau, E. Scornet, A random forest guided tour, Test, 25(2), 197-227(2016).
10. E. LeDell, S. Poirier, H2o automl: Scalable automatic machine learning, in Proceedings of the 7th ICML Workshop on Automated Machine learing(AutoML), (2020).
11. D. Xin, E. Y. Wu, D. J. L. Lee, N. Salehi, A. Parameswaran, Whither AutoML? Understanding the role of automation in machine learning workflows, 2021, arXiv 2101.04834(2021).
12. NIER(National Institute of Environmental Research) Home Page, Realtime water information system, http://www.koreawqi.go.kr/index _web.jsp, July(2022).
13. ME(Ministry of Environment), Geum river watershed mid area water environment management plan(19-23), Performance evaluation final report, 11-1480355-000108-10(2020).
14. F. Pedregosa, G. Varoquaux, A. Gramfort, Y. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, Scikit-learn: Machine learning in Python, J. Mach. Learn. Res., 12, 2825-2830(2011).
15. D. N. Moriasi, J. G. Arnold, M. W. Van Liew, R. L. Bingner, R. D. Harmel, T. L. Veith, Model evaluation guidelines for systematic quantification of accuracy in watershed simulations, Am. Soc. Agric. Biol. Eng., 50(3), 885-900(2007).
16. N. D. Bennett, B. F. Croke, G. Guariso, J. H. Guillaume, S. H. Hamilton, A. J. Jakeman, S. Marsili-Libelli, L. T. Newham, J. P. Norton, C. Perrin, Characterising performance of environmental models, Environ. Modell. Softw., 40, 1-20(2013).

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.

Authors and Contribution Statement

Jungsu Park

Deptartment of Civil and Environmental Engineering, Hanbat National University, Assistant Professor, ORCID 0000-0002-9780-6988: Conceptualization, Data curation, Data analysis, Methodology, Visualization, Writing – original draft, Writing – review and editing.