KJEE 56(3): 268-279 (2023)
https://doi.org/10.11614/KSL.2023.56.3.268
ISSN: 2288-1115 (Print), 2288-1123 (Online)

Note

담수 유해남조 세포수·대사물질 농도 예측을 위한 머신러닝과 딥러닝 모델링 연구동향: 알고리즘, 입력변수 및 학습 데이터 수 비교

박용은* · 김진휘 · 이한규[1] · 변서현[1] · 황순진[2] · 신재기[3,*]

건국대학교 사회환경공학부, [1]건국대학교 사회환경플랜트공학과, [2]건국대학교 환경보건과학과,
[3]수생태원 한강 (韓江)

Machine- and Deep Learning Modelling Trends for Predicting Harmful Cyanobacterial Cells and Associated Metabolites Concentration in Inland Freshwaters: Comparison of Algorithms, Input Variables, and Learning Data Number.

*Yongeun Park** (0000-0002-1959-0843), *Jin Hwi Kim* (0000-0003-2115-8969), *Hankyu Lee*[1] (0000-0002-6619-6160), *Seohyun Byeon*[1] (0009-0001-6597-7264), *Soon-Jin Hwang*[2] (0000-0001-7083-5036) *and Jae-Ki Shin*[3,*] (0000-0002-5380-5078) (*School of Civil and Environmental Engineering, Konkuk University, Seoul 05029, Republic of Korea; [1]Graduate School of Civil, Environmental and Plant Engineering, Konkuk University, Seoul 05029, Republic of Korea; [2]Department of Environmental Health and Science, Konkuk University, Seoul 05029, Republic of Korea; [3]Limnoecological Science Research Institute Korea (THE HANGANG), Gyeongnam 50440, Republic of Korea*)

Abstract Nowadays, artificial intelligence model approaches such as machine and deep learning have been widely used to predict variations of water quality in various freshwater bodies. In particular, many researchers have tried to predict the occurrence of cyanobacterial blooms in inland water, which pose a threat to human health and aquatic ecosystems. Therefore, the objective of this study were to: 1) review studies on the application of machine learning models for predicting the occurrence of cyanobacterial blooms and its metabolites and 2) prospect for future study on the prediction of cyanobacteria by machine learning models including deep learning. In this study, a systematic literature search and review were conducted using SCOPUS, which is Elsevier's abstract and citation database. The key results showed that deep learning models were usually used to predict cyanobacterial cells, while machine learning models focused on predicting cyanobacterial metabolites such as concentrations of microcystin, geosmin, and 2-methylisoborneol (2-MIB) in reservoirs. There was a distinct difference in the use of input variables to predict cyanobacterial cells and metabolites. The application of deep learning models through the construction of big data may be encouraged to build accurate models to predict cyanobacterial metabolites.

Key words: cyanobacterial cells, deep learning, geosmin, machine learning, 2-MIB, microcystin, metabolites

Manuscript received 13 August 2023, revised 1 October 2023,
revision accepted 1 October 2023
* Co-corresponding author: Tel: +82-55-354-2427,
E-mail: shinjaeki@gmail.com
Tel: +82-2-2049-6106,
E-mail: yepark@konkuk.ac.kr

서 론

담수에서 유해남조의 과도한 대발생(cyanobacterial harmful algal blooms, CHABs)은 유체 역학의 수문학적 구조 변형으로 늘어난 수리학적 체류시간(Mitrovic *et al.*, 2010; Hwang

et al., 2016), 유역 오염원(예, CSOs, 비점오염)의 불완전한 통제(Hwang *et al.*, 2017), 내외적 과잉 영양염 공급과 축적에 의한 부영양화(Schindler, 2012) 및 지구온난화에 의한 기후(예, 온도, 가뭄) 패턴 변화(Paerl and Huisman, 2009; Paerl, 2014; Qin *et al.*, 2021)가 매 순간 단일·복합적 원인으로 작용하여 날로 심화되는 수질환경 문제이다(Anderson *et al.*, 2012; Bruder *et al.*, 2014). CHABs는 그 자체만으로도 전형적인 물 오염 징후 중 하나로 볼 수 있으며(Reynolds and Walsby, 1975; Pearl, 2014), 수색 변화와 투명도 저하(Shin and Park, 2018; Summers and Ryder, 2023), 용존산소 감소와 어류 폐사(Watanabe *et al.*, 1996; OSTP, 2016), 생물 다양성 감소(Dodds *et al.*, 2009) 및 이취미 발생과 식수 공급 장애(Shin *et al.*, 2022) 등과 같은 다양한 부정적 이벤트와도 무관하지 않다(Bruder *et al.*, 2014). 또한, 대발생 수준(10^6 cells mL^{-1})에 다다르면 수체의 유동성에 따라 그 영향권은 공간적으로 급속히 확대되어 가시 규모의 제한적 대응관리 이외에는 속수무책 상태에 이르게 되며(Paerl and Otten, 2013; Summers and Ryder, 2023), 이 무렵에 독소 및 이취미 대사물질의 고농도 이벤트는 무색으로 육안 식별이 쉽지 않아 즉각적인 대응은 더욱 더 어렵고 복잡해진다(Shin *et al.*, 2022).

CHABs는 유해남조의 주요 생리생태학적 특성에 의해 군체(예, *Microcystis*) 또는 사상체(예, *Anabaena*)의 성장과 수체의 안정성이 맞물리면 순식간에 폭발적으로 형성된다(Reynolds and Walsby, 1975; Paerl and Millie, 1996; Watanabe *et al.*, 1996). 외형적 크기 또는 비정형적 구조와 표층에 집적되는 강한 부유력(buoyancy)에 기인한 높은 성장률은 비교적 고온 및 고광 시기의 여름철을 중심으로 더욱 촉발하는 우세한 요인이 되며(Reynolds and Walsby, 1975), 때로는 수표면에 조류막의 스컴(scum)을 이뤄 경관을 해치기도 한다(Raps *et al.*, 1983). 게다가, 일단 대발생 상태로 번성한 후에는 다양한 인위적 환경요인(예, 보 및 저수지 운영패턴, 수위변동, 중·저층수 발전방류 등)의 작용에 의해 시공간적으로 일정 수준 이상의 밀도를 유지하여, 초겨울의 저온기까지도 좀처럼 소멸되지 않는 생태학적 특성을 가진다(Nichols *et al.*, 2006; Shin *et al.*, 2016; Shin and Park, 2018; Shin *et al.*, 2022). 이러한 현상은 자연 상태의 유수하천에 비해 댐 저수지와 보 풀(pool)을 포함한 조절하천의 반 폐쇄성 수역에서 더욱 뚜렷하고 장기화되는 양상을 보이기도 한다(Mitrovic *et al.*, 2010; Shin and Park, 2018). 또한, 수자원 이용 및 수생태계 관리 측면에서 CHABs의 생활사 중 체내·외 대사과정을 통해 분비 후 배출되는 독소(예, microcystin) 및 냄새물질(예, geosmin, 2-MIB)은 세포수와 더불어 중요한 관심의 대상이 되고 있으며, 그 농도의 세기나 노출 또는 잔류 영향은 해를 거듭할수록 더욱 부각되고 있는 실정에 있다(Paerl and Otten, 2013; Kim *et al.*, 2021; Shin *et al.*, 2022).

무엇보다도 CHABs의 독소를 비롯하여 이상과 같은 다양한 영향으로 인해, WHO (World Health Organization)를 중심으로 미국, 캐나다, 유럽연합(EU), 영국, 일본, 호주 및 뉴질랜드 등 세계 주요 선진국가(WHO, 2011)뿐만 아니라 우리나라는 CHABs를 체계적으로 대응 관리하고 상수원의 안전성 확보를 위해 연중 조류경보제를 시행 중에 있으며(Shin *et al.*, 2016), 상수원 및 친수활동 수역에 해당하는 주요 하천·저수지 구간을 대상으로 유해남조 개체군(예, *Anabaena*, *Aphanizomenon*, *Microcystis*, *Oscillatoria*속)의 세포수를 주간 모니터링하고 있다(MOE-NIER, 2020). 또한, 실측 데이터 기반 모델링을 통해 사전에 CHABs의 발생량을 예측하고, 그 결과를 정부 공인 공공포털인 「물환경정보시스템」에 업로드하여 과학적 물 관리에 활용하고 있다(MOE-NIER, 2020).

최근까지, 국내외에서 CHABs 발생으로부터 안전하고 양질의 수자원 관리를 위한 수질예측에는 다양한 물리적 및 데이터 기반 모델들을 목적과 필요성에 따라 꾸준히 사용해 왔다(Peters *et al.*, 2014; LeCun *et al.*, 2015; Baker *et al.*, 2018; Kratzert *et al.*, 2019; Schuwirth *et al.*, 2019; Kim *et al.*, 2021). 우리나라의 경우, 2020년에 조류경보제와 수질예보제를 통합하기 전까지 다기능 보를 대상으로 EFDC 모델을 이용하여 수온과 chlorophyll-*a* (chl-*a*) 항목을 예측하여 사전예방적 수질관리를 실행하였다(Lee *et al.*, 2012). 이후, 머신러닝(machine learning, ML)과 딥러닝(deep learning, DL) 알고리즘에 기반한 모델을 이용하여 CHABs를 예측하였으며, DL 모델링은 아직까지 미진한 상태에 있다(Kim *et al.*, 2021). 또한, CHABs 예측에 가용된 학습 데이터는 chl-*a* 농도 또는 형광 측정값 항목이 더 선호되었다(Harada *et al.*, 2013; Liu *et al.*, 2013; Xiao *et al.*, 2017). CHABs의 세포수, 독소 및 냄새물질 예측에 대한 선행연구는 매우 부족한 실정에 있으나, 점차 다변화되는 수질문제에 능동적으로 대처하기 위해 이에 대한 사전 연구는 필요한 것으로 사료된다. 그 중에서 우선적으로, ML과 DL 모델을 기반으로 한 CHABs의 세포수 및 대사물질 예측에 대한 연구동향 검토를 통해 언급되고 있는 주요 내용 파악과 더불어 미래 방향성도 탐구될 필요가 있다.

본 연구는 국내외에서 ML과 DL 모델 기반으로 CHABs의 세포수와 대사물질을 예측한 선행 연구사례를 과학논문 데이터베이스(예, SCOPUS)로부터 주요 키워드의 검색으로 문헌을 탐색 및 조사하였다. 수집된 연구 결과들은 상세하게 분석하여, 해당 주제에 관한 ML과 DL의 현황 및 경향을 검

토하였다. 이를 근간으로, 모델의 알고리즘 적용성, 입력(매개)변수의 종류 및 가용된 학습 데이터 수의 비교를 통해 주요 특성을 파악하였고, 향후 ML 또는 DL 모델링 기술의 발전에 따른 CHABs 예측 연구의 성능 향상에 필요한 사항을 중점적으로 고찰하였다.

재료 및 방법

1. 국내외 문헌 데이터 수집

CHABs 세포수 및 대사물질(예, microcystin, geosmin 및 2-MIB) 농도에 대한 ML과 DL 모델링 예측과 관련된 선행 연구문헌의 탐색과 수집은 다중 데이터베이스 검색 포털인 SCOPUS (https://www.scopus.com/)를 이용하였다. SCOPUS는 과학 관련 문헌들을 출판하는 대표적 기업인 Elsevier에 의해 만들어진 세계 최대 규모의 학술 데이터베이스 검색 플랫폼으로 과학, 기술, 의학 및 사회과학 분야를 총망라하여 영어로 작성된 다양한 문헌의 검색을 가능하게 해준다. 또한, 본 연구를 위한 문헌 검색도 영문 키워드만 사용하였다. ML과 DL 모델의 알고리즘 관련 주요 검색 키워드는 기계학습(machine learning)과 심층학습(deep learning)으로 하였고, CHABs와 연관된 검색 키워드는 남조 세포수(cyanobacterial cells), 독소(microcystin) 및 냄새물질(geosmin, 2-MIB)로 설정하였다. 예를 들어, 남조 세포수 예측에 ML 알고리즘이 적용된 문헌을 검색하는 경우 “cyanobacterial cells + machine learning”을 사용하였고, DL 알고리즘이 적용된 경우 “cyanobacterial cells + deep learning”을 사용하였다(Table 1). 남조 대사물질의 경우에도 microcystin 예측 연구에서 ML이 적용된 경우 “microcystin + machine learning”, DL이 적용된 경우 “microcystin + deep learning”을 각각 사용하였다(Table 1). Goesmin과 2-MIB도 동일한 방식으로 검색하여 문헌을 탐색하였다(Table 1). 한편으로, 각 예측변수와 ML과 DL 모델 알고리즘의 키워드 조합 시 검색되지 않는 문헌 정보를 최소화하기 위해 유사 키워드도 보조적으로 활용하여 확인하였다(Al-Sulttani *et al.*, 2021; Sibanda *et al.*, 2021). 한 예로써, CHABs 세포수와 관련해서 “cyanobacterial cells”뿐만 아니라 “cyanobacterial cell density”, “cyanobacterial abundance” 등과 같이 남조의 생물량을 의미할 수 있는 유사 키워드들도 활용하여 문헌을 반복적으로 검색하는 부가적 분석과정도 거쳤다. 문헌데이터 검색기간은 1960년 이전부터 2023년 2월까지로 설정하였다.

2. 탐색 문헌 데이터 분석

CHABs의 각 예측변수와 ML과 DL 모델 알고리즘 조합으로 검색된 문헌들을 목록화하였고, 사전에 개별 문헌의 초록과 주요 내용을 검토하여 상세 분석이 필요한 연구들을 심층 분류하였다. 그 결과, 본 연구 수행에서 요구하는 내용을 포함하지 않는 문헌들은 제외하였다. CHABs 세포수, 독소(1종) 및 냄새물질(2종) 예측을 위해 ML과 DL 모델 알고리즘을 적용한 선행 연구사례 또는 심층 분류를 마친 문헌들은 상세 분석을 통해 적용된 알고리즘의 종류별로 각각 구분하였다. 아울러 각 연구에 활용된 입력변수(예, 기상, 수리·수문 및 수질), 학습 데이터 수 및 모델 적용사례를 동시에 조사한 후 상호 비교하였다.

3. 텍스트마이닝 및 동시 출현단어 분석

SCOPUS에서 얻은 키워드 검색 결과는 csv format 파일로 저장하였고, 이 파일은 서지 네트워크의 구축 및 시각화를 하기 위해 VOS (visualization of similarities) viewer (ver. 1.6.16, CWTS, The Netherlands) 소프트웨어를 사용하였다 (van Eck and Waltman, 2009, 2010). VOSviewer는 논제, 초록, 키워드 및 보문 섹션에 사용된 단어의 빈도를 기준으로 검색 결과를 분석하였다(van Eck and Waltman, 2007; van Eck *et al.*, 2010a). 텍스트마이닝 기법을 통한 주제어(또는

Table 1. Key search words used in this study.

Search platform	Search criterion	Total number of articles	Number of articles retained
SCOPUS	(TITLE-ABS-KEY ((“machine learning”) OR (“deep learning”) AND ((“cyanobacterial cell”)))	55	10
	(TITLE-ABS-KEY ((“machine learning”) OR (“deep learning”) AND ((“microcystin”)))	20	5
	(TITLE-ABS-KEY ((“machine learning”) OR (“deep learning”) AND ((“geosmin”) OR (“2-MIB”)))	6	4

중심어, 핵심어) 간의 네트워크 분석과 주제어의 중요도, 밀도 등은 VOSviewer program의 클러스터링 맵(clustering map), 네트워크 시각화(network visualization) 및 단어 밀도 맵(density map)으로 표현하였다(Waltman *et al.*, 2010). VOSviewer 결과는 mapping과 clustering 기법을 이용하여 분석한 후 이를 시각화하였다(van Eck *et al.*, 2010b). 네트워크 시각화는 주제어의 동시 출현 빈도를 기준으로 그 연관성을 Kullback-Leibler distance를 이용하여 공간적으로 도해화하였다(Waaijer *et al.*, 2011).

참고로, 상호 연관성이 있는 주제어들은 동일 색상(color)을 통해 집괴화(clustering)되며, 동일한 색상의 집단 내에서도 동시 출현 빈도가 많을수록 가깝게 위치하고, 적을수록 상호 멀리 떨어져 놓이게 된다(van Eck and Waltman, 2009; Waltman *et al.*, 2010). 이것은 두 개 이상의 주제어가 하나의 문서(document)에서 동시에 활용되고 있다면, 이들 주제어가 상호 연관된다는 가정을 기초하고 있는 것이다(van Eck *et al.*, 2010a). 중요도는 주제어 밀도와 출현 수에 따라 자동적으로 결정되며, 출현 수의 빈도가 많을수록 네트워크의 시각화에서 크게 나타난다(van Eck *et al.*, 2010a). 주제어의 밀도 또한 출현 빈도에 의해 결정되며, 밀도가 높아지면 짙은 색에, 낮아지면 연한 색에 근접하는 강·약의 색상 구배를 가진다. 이렇게 하여 전체 자료 중 각 주제어 밀도와 빈도를 일목요연하게 나타내주게 된다(van Eck and Waltman, 2009, 2010; van Eck *et al.*, 2010a).

결과 및 고찰

1. CHABs 세포수와 대사물질 농도 예측을 위한 ML과 DL 모델링의 선행 연구현황 및 네트워크 분석

ML과 DL 모델링을 통해 담수 CHABs의 세포수 및 대사물질을 예측한 선행 연구는 전체 키워드 검색 결과 총 81개 중 최종적으로 20개가 해당하였다(Table 2). 이때, 하나의 문헌에서 여러 대사물질의 결과를 동시에 소개한 것은 각 개별 건수로 다루었다. 이 중에서, ML을 활용한 예측 연구는 15개(75%)였고, DL은 5개(25%)였다. 모델 별로는 ML에서 세포수가 6개(30%), 독소가 4개(20%)였고, DL에서 세포수가 4개(20%)를 차지하였다. 예측변수별로 보면, 세포수가 10개(50%), 대사물질이 10개(50%)였다. 대사물질 중 독소가 5개(25%), 냄새물질이 5개(25%)였다. ML과 DL 모델링 기법을 활용한 CHABs의 세포수 및 대사물질 예측은 양적으로 많지 않았으나, 냄새물질(geosmin, 2-MIB)까지도 포함한 다양한 연구가 진행된 특성을 나타냈다.

Table 2. Number of previous research literatures predicting the cyanobacterial cells and associated metabolites concentration using machine- and deep learning models.

Models/ Variables	Cyanobacterial cells	Metabolites		
		Microcystin	Geosmin	2-MIB
Machine learning	6	4	4	1
Deep learning	4	1	–	–

최근까지, CHABs를 예측하는 데 ML과 DL 모델의 활용이 증가하고 있으며(Russo *et al.*, 2020), 이는 CHABs에 대한 기본 지식만으로도 최적화된 입력변수 선택 및 학습을 통해 모델의 성능을 향상시킬 수 있는(Chen and Mynett, 2003; Teles *et al.*, 2008; Fornarelli *et al.*, 2013) 반면, 학습에 활용된 데이터 내에서 존재하는 패턴만 학습 가능하여 미지의 데이터에 존재하는 새로운 패턴은 예측하기가 쉽지 않은 면도 있다(Welk *et al.*, 2008). 즉, 예측 규칙 중 일부는 선택된 변수의 값을 예상하는 데 유용할 수도 있지만, 직간접적으로 물리적인 인과관계를 설명하기는 어려울 수도 있다(Recknagel *et al.*, 2014).

CHABs의 세포수, 독소 및 냄새물질 예측에 대한 ML과 DL 모델링의 선행 연구를 중심어(keywords)로써 네트워크를 분석한 결과는 Fig. 1과 같다. 특히, ML과 CHABs의 세포수 관계는 다른 모델-변수에 비해 다소 복잡한 양상을 보였다(Fig. 1A). 이에 반해, 남조 독소와 냄새물질은 단순한 관계성을 나타냈다(Fig. 1B, C). VOSviewer의 집괴된 결과에 기초하여 주제어(적정점수(relevance score) 상위 60% 또는 출현 빈도 최소 2~3회 이상에 해당하는 단어)가 각각 2~3개의 영역(cluster)으로 나뉘었다(Fig. 1). 첫 번째, 「ML과 CHABs 세포수」에서 주제어는 37개였고(Fig. 1A), 영역 1은 조류번성(algal bloom), 영역 2는 수질예측(prediction) 및 영역 3은 남조(cyanobacteria)를 중심으로 그와 관련된 단어가 높은 빈도로 출현하여 네트워크를 이루었다. 두 번째, 「ML과 microcystin」에서 주제어는 40개였고(Fig. 1B), 영역 1은 조류독소(microcystin), 영역 2는 유해조류번성(harmful algal bloom) 및 영역 3은 부영양화(eutrophication)를 중심으로, 세 번째, 「geosmin과 ML」에서 주제어는 10개였고(Fig. 1C), 영역 1은 냄새물질(geosmin), 영역 2는 남조(cyanobacteria)를 중심으로 각각 관련성을 가졌다. 그리고 네 번째, 「DL과 CHABs 세포수」에서 주제어는 26개였고(Fig. 1D), 영역 1은 심층학습(deep learning), 영역 2는 유해조류번성(harmful algal blooms) 및 영역 3은 남조(cyanobacteria)를 중심으로 관계망을 형성하였다. 주요 집괴된 영역들은 유해남조 세포

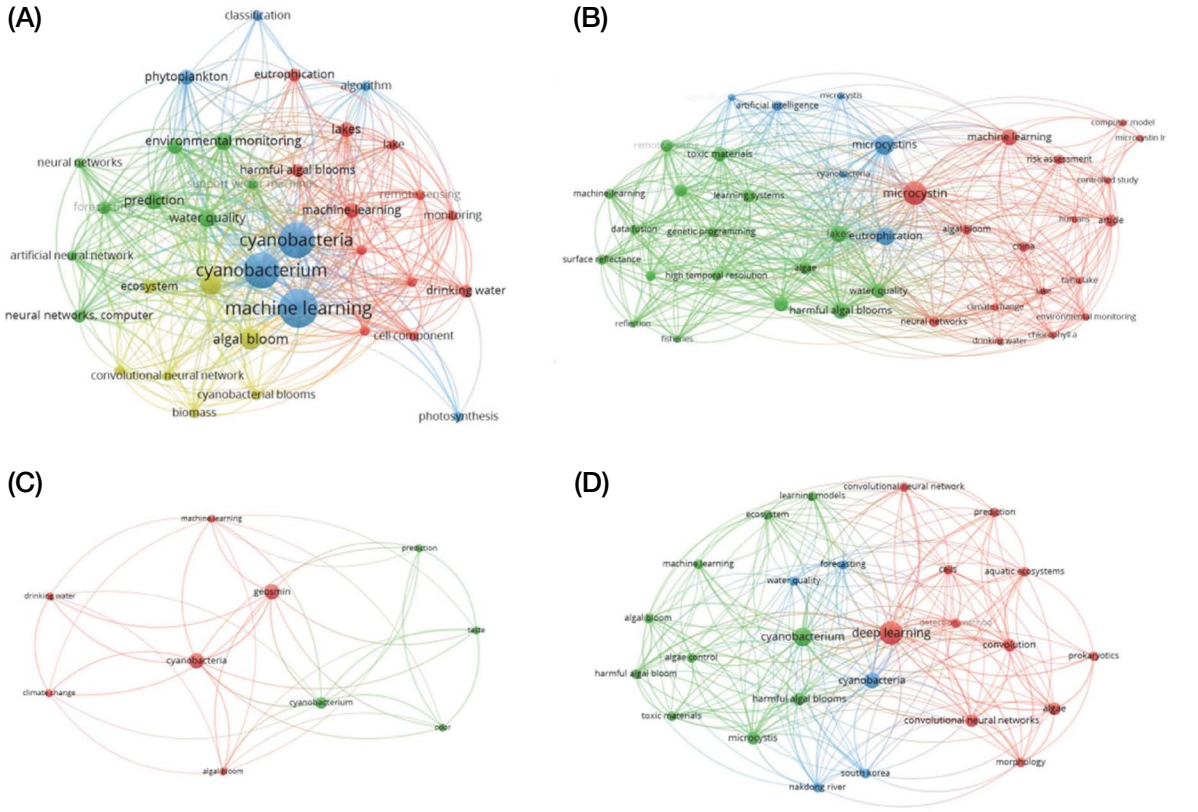

Fig. 1. The VOSviewer results showing the keywords co-occurrence network visualization using data collected in the scopus database. Each cluster is distinguished by different colors. Size of circle represents quantity of network degrees. And gradient of color shows total frequency of the term. (A): cyanobacterial cells and ML, (B): microcystin and ML, (C): geosmin and ML, and (D): cyanobacterial cells and DL.

수 및 대사물질의 발생과 예측 도구(예, ANN 또는 CNN 알고리즘)의 상호 연계성을 가진 주제 분야로 구분되었다. 결과적으로, 문헌 검색에 대한 키워드가 구체적이어서 각 영역의 주제어 범위도 비교적 단순하였으며, ML은 DL에 비해 보다 포괄적인 것으로 사료되었다.

2. CHABs 세포수와 대사물질 농도 예측을 위한 ML과 DL 모델에 적용된 알고리즘 비교

ML은 CHABs의 세포수, 독소 및 냄새물질 예측변수에 대해 다양하게 이용되었고, 특히 CHABs의 세포수보다 대사물질 예측에 사용되는 종류가 월등히 많았다(Table 3). 세포수 예측에 적용된 알고리즘은 ANN과 RF였다. 대사물질 중 독소 예측에는 ANN, BT, Cubist, HMM, LASSO 및 SVM의 6종이었고, geosmin과 2-MIB는 CART, FLM, MARS 및 RF의 4종이었다(Table 3). ANN은 세포수와 독소 예측에, RF는 세포수와 geosmin 예측에 각각 공통적으로 사용되었다(Table 3). 반면에, DL 알고리즘은 CHABs의 세포수 예측에 대부분 이용되었으며, 그 외 독소 대사물질에서 LSTM 1종이 유일하였다(Table 3).

주요 알고리즘의 적용을 한 예로써 살펴보면, ANN은 CHABs의 미래(단기 또는 장기) 예측(Yabunaka *et al.*, 1997; Recknagel *et al.*, 1998; Wei *et al.*, 2001; Xiao *et al.*, 2017; Kim *et al.*, 2021)과 수자원 데이터 분석(Maier and Dandy, 2000) 등에 많이 사용되었다. 단기(5~30일 범위) 예측에서 만족할 만한 정확도를 보였는데, 예측기간이 짧을수록 최소한 2배 이상 높은 정확도(예, 단기: r^2 = 0.74~0.89 (4 hrs~ < 1 month), 장기: 0.32~0.46 (> 1 month~10 yrs))를 가졌다(Harris and Graham, 2017). 남조의 총세포수에 대해 비교적 높은 결정계수(r^2) 값을 가지지만(Luo *et al.*, 2017; Xiao *et al.*, 2017; Kim *et al.*, 2021), 기본 프로세스에 대한 자세한 설명은 제공하지 못하였다(Recknagel, 1997; Wei *et al.*, 2001). 그리고 *Microcystis* 개체군의 모의에서 우수하였으나, 예측변수를 확인하기 위해 민감도 분석이 요구되었다(Harris and Graham, 2017).

CHABs의 세포수, 독소 및 냄새물질을 예측하기 위한 ML 모델의 알고리즘을 비교한 것은 Table 4와 같다. ML에 속하는 주요 알고리즘은 BT, Cubist, RF 및 SVM이었다. CHABs

Table 3. Algorithms of ML and DL used in the study predicting the cyanobacterial cells and associated metabolites concentration using machine- and deep learning models. Numerics indicate the number of literature included in each case.

Features	Algorithm	Cyanobacterial cells	Metabolites		
			Microcystin	Geosmin	2-MIB
Machine learning[a]	ANN	5	1	•	•
	BT	•	1	•	•
	CART	•	•	2	•
	Cubist	•	1	•	•
	FLM	•	•	1	1
	HMM	•	1	•	•
	LASSO	•	1	•	•
	MARS	•	•	1	•
	RF	1	•	3	•
	SVM	•	1	•	•
Deep learning[b]	CNN	2	•	•	•
	GRU	2	•	•	•
	HDL	1	•	•	•
	LSTM	3	1	•	•
	RETAIN	1	•	•	•
	Transformer	1	•	•	•

Abbreviation: [a] Artificial Neural Network (ANN), Boosted Tree (BT), Classification And Regression Trees (CART), Fuzzy Logic Model (FLM), Hidden Markov Model (HMM), Least Absolute Shrinkage and Selection Operator (LASSO), Multivariate Adaptive Regression Splines (MARS), Random Forest (RF), Support Vector Machine (SVM); [b] Convolutional Neural Network (CNN), Gated Recurrent Unit (GRU), Hierarchical Deep Learning (HDL), Long Short-Term Memory (LSTM), Reverse Time Attention (RETAIN).

Table 4. Comparison of machine learning model algorithms applied predicting the cyanobacterial cells and associated metabolites concentration.

Targets\Model algorithms	BT	Cubist	RF	SVM
CHABs abundance (cells mL^{-1})				
– Below 60,000	•	•	+	•
– Above 60,000	•	•	+	•
Microcystin (μg L^{-1})				
– Below 2.5	+	+	•	+
– Above 2.5	•	•	•	•
Geosmin (ng L^{-1})				
– Below 20	•	•	+	•
– Above 20	•	•	+	•

See Table 3 for the definitions of the abbreviations on the model algorithm.

의 세포수가 6.0×10^4 cells mL^{-1} 이하 또는 그 이상에서 RF만 적용되었다(Table 4). Microcystin은 2.5 μg L^{-1} 이하에서 BT, Cubis 및 SVM만 적용되었고, geosmin은 20 ng L^{-1} 이하 또는 그 이상 농도에서 RF가 적용되었다(Table 4). 이때, CHABs 세포수와 대사물질의 발생 또는 촉발 시점은 대체로 비슷하게 맞추었으나, 피크(peak)의 높은 값은 비교적 과소평가되는 경향을 나타냈다. 이러한 결과는 개별 알고리즘 성능의 특성과 관련되는 것으로 사료되었다(Harris and Graham, 2017).

3. CHABs 세포수와 대사물질 농도 예측을 위한 ML과 DL 모델링에 적용된 입력변수 비교

CHABs의 세포수와 대사물질 농도를 예측하기 위해 ML과 DL 모델에 사용된 입력변수는 총 30개로 확인되었고, 이 중에서 기상변수는 10개, 수리·수문변수는 3개 및 수질변수는 17개였다(Table 5). 그중 기상 및 수리·수문변수는 CHABs의 세포수와 독소 예측에만 사용되었고, geosmin과 2-MIB의 냄새물질 예측에는 전혀 이용되지 않았다(Table 5). 예측변수에 따른 입력변수는 CHABs의 세포수에서 22개로 다소 많았고, 독소와 geosmin에는 각각 13개, 15개로 구성되었으나 geosmin은 수질환경 변수만 포함하였다(Table 5). 2-MIB는 수온, 염분도, 총인 및 세포수의 4개 인자만 사용되어 비교적 단순하였다(Table 5).

전술한 바와 같이, 선행 연구에서 입력변수 적용의 유형과 수는 상당히 다양하였다. 대부분의 모델에는 기상 및 수리·수문 데이터와 함께 물리, 화학 및 생물학적 수질 매개변수의

Table 5. Comparison of input variables used predicting the cyanobacterial cells and associated metabolites concentration using machine- and deep learning models.

Features	Input variables	Cyanobacterial cells	Metabolites		
			Microcystin	Geosmin	2-MIB
Meteorological variables	Air temperature	●	●	·	·
	Rainfall	●	·	·	·
	Amount of light	●	·	·	·
	Evaporation	●	·	·	·
	Cloudiness	●	·	·	·
	Atmospheric pressure	●	●	·	·
	Wind speed	●	●	·	·
	Wind direction	·	●	·	·
	Rainfall intensity	●	·	·	·
	Rainfall duration	●	·	·	·
Hydraulics and hydrological variables	Flow discharge	●	●	·	·
	Current	●	●	·	·
	Water level	●	●	·	·
Water quality variables	Water temperature	●	·	●	●
	Dissolved oxygen	●	·	●	·
	pH	●	·	●	·
	Conductivity	●	·	●	·
	Salinity	·	·	●	●
	Total dissolved solids	·	·	●	·
	Turbidity	●	●	●	·
	Suspended solids	·	●	●	·
	Transparency	●	●	·	·
	Biological oxygen demand	●	·	●	·
	Chemical oxygen demand	·	●	●	·
	Organic matter	·	·	●	·
	Total Kjeldahl nitrogen	·	·	●	·
	Total phosphorus	·	·	·	●
	Nutrients (loads)	●	●	●	·
	Chlorophyll-*a*	●	●	●	·
	Algal cells	·	·	●	●

조합으로 구성되었으나, 예측변수에 따라 유연성이 높고 가변성이 컸다(Guven and Howard, 2006; Mooij *et al.*, 2010). CHABs 세포수의 경우는 유해남조의 생활사 전반에 걸친 과정이 관련되므로 그만큼 고려되어야 하는 입력변수가 많아지게 된다(Rousso *et al.*, 2020). 반면에, 대사물질의 경우는 남조의 생활사 중에서 조체의 노쇠 또는 사멸기에 현저하게 배출되므로(Shin *et al.*, 2022), 이와 관련된 환경요인 변수가 다소 제한적일 수 있다. ML과 DL 예측모델은 적절하게 선택된 입력변수를 축소하더라도 만족할 만한 성능을 달성할 수 있다(Liu *et al.*, 2013; Xiao *et al.*, 2017). 입력변수의 선택은 비용적 측면 외에도 가용성, 신뢰성 및 데이터 획득의 용이성을 고려해야 할 필요가 있다(Ostfeld *et al.*, 2015). 이러한 이유 때문에, 많은 연구에서는 입력변수의 가지 수를 줄인 예측모델로 개발하는 것을 목표로 한다(Liu *et al.*, 2013; Xiao *et al.*, 2017). 그 예로써, CHABs에 대한 단기 예측모델 개발에서 chl-*a* (Harada *et al.*, 2013) 또는 형광 데이터(Liu *et al.*, 2013; Xiao *et al.*, 2017)만을 기반으로 하거나, 남조 동태를 예측하기 위해 사전에 여러 변수를 예비평가하기도 한다(Wang *et al.*, 2018).

한편으로, CHABs에 가장 큰 영향을 미치는 입력변수를 찾는 것을 목표로 하는 모델이 많이 사용되었다. CHABs 모델링에 가장 민감한 변수 즉, 주요 입력변수를 확인하고, 여

러 기법들(예, 민감도 분석, 주성분 분석, 데이터 마이닝 등)을 이용하여 입력변수와 출력값 간의 통계적 관계를 평가한 후 예측 또는 성능에 가장 민감한 변수의 우선순위를 매기기도 한다(Gelman and Hill, 2006). 그러나 통계적 상관관계는 인과관계를 의미하는 것은 아니며(Gardner, 2000), 통계적으로 민감한 변수가 반드시 CHABs에 대한 인과관계의 변수요인이 아닐 수도 있다(Recknagel *et al.*, 1997). 따라서, 입력변수의 선택은 CHABs 예측 정도의 변동성을 파악하는 것으로써, 입력변수는 모델 성능을 향상시킬 수도 있지만, 반드시 기본적인 현상들을 이해하도록 하는 것은 아니다.

또한, 남조의 동태를 보여주는 입력변수는 매우 다양하고 복잡한데, 그것과 직접적으로 연관된 입력변수를 포함하게 되면, CHABs 예측 및 모델 성능의 추정력에 긍정적 효과를 제공할 수도 있다(Cawley and Talbot, 2010). 한편, 입력변수를 많이 포함하는 것도 반드시 성능을 향상시키기 위한 전제 조건은 아니다. 그 이유는, 교차 상관 효과로 인해 패턴 인식을 방해하거나 데이터의 과적합(overfitting)을 초래할 수 있기 때문이다(Cawley and Talbot, 2010). 따라서 입력변수 개수의 감소에 의한 차원 축소는 대다수 모델의 목표가 되며(Chen and Mynett, 2003; Teles *et al.*, 2008; Fornarelli *et al.*, 2013), 이와 관련한 주요 방법으로써 주성분분석(principal component analysis, PCA)과 정준상관분석(canonical correspondence analysis, CCA)이 가장 많이 사용되고 있다. 이것은 모델링의 기능적 향상뿐만 아니라 비용 절감, 계산시간 단축 및 효율성 등을 제고할 수 있는 잇점이 있다(Millie *et al.*, 2014; Qin *et al.*, 2015; Xiao *et al.*, 2017).

4. CHABs 세포수와 대사물질 농도 예측을 위한 ML과 DL 모델링에 사용된 학습 데이터 수 비교

CHABs의 세포수, 독소 및 냄새물질을 예측하기 위해 ML과 DL 모델에 각각 사용된 학습 데이터의 수를 요약한 것은 Table 6과 같다. ML과 DL 모델로 CHABs의 세포수 예측에 사용된 평균 데이터의 수(범위)는 각각 109개(39~185개), 2,528개(1,826~2,922개)였고, 독소는 240~277개, 냄새물질은 112~127개 범위였다(Table 6). 대사물질의 경우, ML과 DL에서 데이터의 수적 차이는 거의 없었으나, CHABs의 세포수 예측에서는 달랐다(Table 6). DL 모델은 ML에 비해 23배 이상 많은 데이터가 사용되었다.

일반적으로, ML과 DL 모델링 연구에서 많은 빈도의 학습 데이터를 사용할수록 모델의 예측 성능은 보다 향상되는 것으로 언급하고 있다(Moe *et al.*, 2016; Tromas *et al.*, 2017; Page *et al.*, 2018; Wang *et al.*, 2018; Wilkinson *et al.*, 2018). 그러면서도 데이터 수의 양적 범위를 명확하게 제시하지는 않았다. 현재 사용 가능한 데이터의 양과 질은 모델링 접근 방식을 선택하는 데 중요할 수 있으며, 모델은 개발 및 검증을 위해 최소한의 일상적 모니터링 데이터가 필요하다(Bertone *et al.*, 2018). 또한, ML과 DL 모델은 학습 데이터의 가용성뿐만 아니라 질적 수준에도 크게 의존하기 때문에 데이터가 모델의 목적에 적합하지 않거나, 사전에 적절하게 전처리되지 않은 경우 모델 성능은 예상과 달리 심각하게 저하될 수 있다(O'Hara and Kotze, 2010; Sheng *et al.*, 2012). 더욱이, ML과 DL 모델의 예측 규칙은 오로지 데이터 그 자체에서만 추출되기 때문에 통상적으로 성능은 관찰 또는 측정된 데이터의 범위로 제한될 수밖에 없다. 또한, CHABs를 예측하기 위해 다중 소스(sources)의 데이터를 조합하는 것도 모델 성능을 개선하는 데 기여할 수 있다(Chen *et al.*, 2019).

Table 6. Mean number of data used predicting the cyanobacterial cells and associated metabolites using machine- and deep learning models.

Models/ Variables	Cyanobacterial cells	Metabolites		
		Microcystin	Geosmin	2-MIB
Machine learning	109 (39~185)	277 (17~731)	127 (72~185)	112
Deep learning	2,528 (1,826~2,922)	240	–	–

Parentheses indicate range (minimum to maximum value) that suggested in literatures.

또 다른 중요한 측면은 학습에 가용할 수 있는 데이터 취득의 기간과 빈도(예, 조사 또는 측정주기)를 들 수 있다(Rousso *et al.*, 2020). 본 연구에서 CHABs 예측을 위해 ML 모델에 짧게는 2개월에서 길게는 40년 동안의 입력 데이터와 기간이 이용된 것으로 파악되었다. 대체적으로, 데이터가 많은 게 바람직하지만, CHABs의 고유 변동성을 제대로 반영하지 못하면 모델 성능은 결코 향상되지 않을 수도 있다(Rousso *et al.*, 2020). 또한, 예측모델의 개발 및 검증을 위해서는 CHABs 발생 유·무 기간 동안의 데이터가 요구되는데, 기존 모니터링 방법(예, 고비용적 현장의 샘플링, 시료 운반, 실험실 분석 및 데이터 처리)에 의존하는 경우, 많은 빈도의 데이터 취득은 CHABs 시기에만 제한적으로 가능하다(Millie *et al.*, 2014). 그러나 모델 검증의 기초가 되는 경우, 다중 샘플링 빈도가 있는 데이터를 사용하게 되면 오히려 결과가 편향될 가능성이 있다(Millie *et al.*, 2014). 또한, 데이터 취득일의 일 중 시간대에 대해서도 중요할 수 있다. 그 예로써, 남조는 부유력 조절(Reynolds *et al.*, 1987) 또는 성층화에 대한 주간(diurnal) 변동(Hamilton *et al.*, 2010)의 생리

생태학적 특성으로 인해 하루 중 수층에 따라 세포밀도를 달리하기 때문이다.

결 론

담수 CHABs의 세포수 및 대사물질을 예측한 ML과 DL 모델링의 선행 연구는 총 20개(ML 75%, DL 25%)로써 비교적 양적으로 많지 않았으나 다양하게 적용되었다. ML은 세포수, 독소 및 냄새물질(geosmin, 2-MIB)까지도 포함하였다. 중심어를 이용한 선행 연구의 네트워크를 분석한 결과, ML과 CHABs의 세포수 관계는 다른 모델-변수에 비해 다소 복잡한 양상을 보였으나, 대사물질은 상대적으로 단순한 관계성을 나타냈다. 주요 영역은 유해남조 및 대사물질의 발생과 예측 도구의 상호 연계성을 가진 주제 분야로 구분되었다. ML 알고리즘(10종)은 CHABs 세포수와 대사물질의 예측에 폭넓게 사용된 반면에 DL (6종)은 CHABs의 세포수 예측에만 주로 쓰였다. CHABs의 세포수와 대사물질 농도를 예측하기 위해 ML과 DL 모델의 입력변수는 총 30개였다. 이 중에서 기상 및 수리·수문변수는 CHABs의 세포수와 독소 예측에만 사용되었고, 냄새물질 예측에는 전혀 이용되지 않았다. 예측변수에 따른 입력변수는 CHABs의 세포수에서 22개로 다소 많았고, 독소와 geosmin에는 각각 13개, 15개로 구성되었으나 geosmin은 수질환경 변수만 포함하였다. 2-MIB는 오직 4개 인자만 사용되어 매우 단순하였다. CHABs의 세포수 예측을 위한 학습에 사용된 데이터 양은 ML에 비해 DL 모델이 23배 정도 많았다.

본 연구는 ML과 DL 모델을 이용한 CHABs 세포수 및 대사물질 농도를 예측한 선행 연구 성과를 통해 주요 알고리즘 적용, 입력변수 및 학습데이터 수의 비교를 문헌적 분석에 의한 것으로써 심층적 검토를 하는 데는 한계가 있었다. 그 이유는 선행연구 모델링의 시공간적 차이와 환경요인의 배경적 이질성 등을 감안할 때 자세한 탐구와 이해는 단시간에 쉽지 않는 것으로 사료되었다. 반면에, 예측변수에 대한 주요 속성사항을 추출하여 내용을 비교한 부분은 어느 정도 가치가 있는 것으로 판단되나 지속적인 업데이트 보완이 필요할 것으로 보이며, 향후 이 분야 연구 수행에 기초자료로써 유용하게 활용될 수 있기를 기대한다.

적 요

근래에 들어, 머신러닝과 딥러닝 모델은 다양한 수체 내 수질변화를 예측하기 위해 광범위하게 사용되고 있다. 특히, 담수호의 물 이용과 수생태계 건강성에 위협 요인으로 작용할 수 있는 유해남조의 발생을 예측하기 위해 많은 연구자들이 인공지능 모델을 활용하고 있다. 따라서, 본 연구에서는 최근까지 유해남조의 발생을 예측하기 위해 적용된 인공지능 모델링의 선행 연구들을 검토하였고, 딥러닝을 포함하여 머신러닝 모델을 이용한 이 분야 연구의 발전방향을 모색하고자 하였다. 먼저, Elsevier의 초록 인용 데이터베이스인 Scopus를 활용하여 체계적인 문헌 연구를 수행하였다. 주요 키워드를 이용하여 탐색 및 정리된 문헌들을 리뷰한 결과, 딥러닝 모델은 주로 남조 세포수 예측에만 사용되었고, 머신러닝 모델은 남조 세포수 이외에 microcystin, geosmin, 2-MIB와 같은 대사물질 예측에도 초점을 맞추고 있었다. 또한, 남조 세포수와 대사물질의 예측을 위해 활용된 입력변수들은 현저한 차이가 있었다. 남조의 대사물질을 예측하기 위해 딥러닝 모델이 적용된 바가 없었는데, 향후 빅데이터 구축을 통한 대사물질을 예측하는 연구가 필요할 것으로 사료된다.

저자정보 박용은(건국대학교 사회환경공학부 교수), 김진휘(건국대학교 사회환경공학부 연구교수), 이한규(건국대학교 사회환경플랜트공학과 박사과정 연구원), 변서현(건국대학교 사회환경플랜트공학과 석사과정 연구원), 황순진(건국대학교 환경보건과학과 교수), 신재기(수생태원 한강(韓江) 원장)

저자기여도 개념설정: Y. Park, J.K. Shin, and S.J. Hwang, 연구방법론: Y. Park, J.K. Shin, and S.J. Hwang, 자료제공관리: S.J. Hwang and J.K. Shin, 자료분석: Y. Park, J.H. Kim, H. Lee, S. Byeon, and J.K. Shin, 원고초안작성: J.K. Shin and Y. Park, 원고교정 및 최종검토: J.H. Kim, H. Lee, S. Byeon, and S.J. Hwang, 과제관리: Y. Park, 연구비 수주: Y. Park. 본 논문의 모든 공저자는 내용을 면밀히 검토하였고, 전적으로 이에 동의합니다.

이해관계 본 논문에는 저자간 이해관계 충돌의 여지가 전혀 없습니다.

연구비 본 연구는 한강수계관리위원회 환경기초조사사업의 지원을 받아 수행되었습니다.

사사 본 논문의 심사과정에서 세세한 검토와 코멘트를 해 주신 익명의 심사위원들께 감사드립니다.

REFERENCES

Al-Sulttani, A.O., M. Al-Mukhtar, A.B. Roomi, A.A. Farooque, K.M. Khedher and Z.M. Yaseen. 2021. Proposition of new ensemble data-intelligence models for surface water quality prediction. *IEEE (Institute of Electrical and Electronics*

Engineers) *Access* **9**: 108527-108541.

Anderson, D.M., A.D. Cembella and G.M. Hallegraeff. 2012. Progress in understanding harmful algal blooms: paradigm shifts and new technologies for research, monitoring, and management. *Annual Review of Marine Science* 3: 143-176.

Baker, R.E., J.M. Pena, J. Jayamohan and A. Jerusalem. 2018. Mechanistic models versus machine learning, a fight worth fighting for the biological community? *Biology Letters* **14**: 20170660.

Bertone, E., M.A. Burford and D.P. Hamilton. 2018. Fluorescence probes for real-time remote cyanobacteria monitoring: a review of challenges and opportunities. *Water Research* **141**: 152-162.

Bruder, S., M. Babbar-Sebens, L. Tedesco and E. Soyeux. 2014. Use of fuzzy logic models for prediction of taste and odor compounds in algal bloom-affected inland water bodies. *Environmental Monitoring Assessment* **186**: 1525-1545.

Cawley, G.C. and N.L. Talbot. 2010. On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal Machine Learning Research* **11**: 2079-2107.

Chen, C., J.C. Huang, Q.W. Chen, J.Y. Zhang, Z.J. Li and Y.Q. Lin. 2019. Assimilating multi-source data into a three-dimensional hydro-ecological dynamics model using Ensemble Kalman Filter. *Environmental Modelling and Software* **117**: 188-199.

Chen, Q.W. and A.E. Mynett. 2003. Integration of data mining techniques and heuristic knowledge in fuzzy logic modelling of eutrophication in Taihu Lake. *Ecological Modelling* **162**: 55-67.

Dodds, W.K., W.W. Bouska, J.L. Eitzmann, T.J. Pilger, K.L. Pitts, A.J. Riley, J.T. Schloesser and D.J. Thornbrugh. 2009. Eutrophication of U.S. freshwaters: Analysis of potential economic damages. *Environmental Science and Technology* **43**: 12-19.

Fornarelli, R., S. Galelli, A. Castelletti, J.P. Antenucci and C.L. Marti. 2013. An empirical modeling approach to predict and understand phytoplankton dynamics in a reservoir affected by interbasin water transfers. *Water Resources Research* **49**: 3626-3641.

Gardner, R.C. 2000. Correlation, causation, motivation, and second language acquisition. *Canadian Psychology/Psychologie Canadienne* **41**: 10-24.

Gelman, A. and J. Hill. 2006. Data Analysis Using Regression and Multilevel/Hierarchical Models. Cambridge University Press, Cambridge, England. 648p.

Guven, B. and A. Howard. 2006. A review and classification of the existing models of cyanobacteria. *Progress in Physical Geography: Earth and Environment* **30**: 1-24.

Hamilton, D.P., K.R. O'Brien, M.A. Burford, J.D. Brookes and C.G. McBride. 2010. Vertical distributions of chlorophyll in deep, warm monomictic lakes. *Aquatic Sciences* **72**: 295-307.

Harada, M., T. Tominaga, K. Hiramatsu and A. Marui. 2013. Real-time prediction of chlorophyll-*a* time series in a eutrophic agricultural reservoir in a coastal zone using recurrent neural networks with periodic chaos neurons. *Irrigation and Drainage* **62**: 36-43.

Harris, T.D. and J.L. Graham. 2017. Predicting cyanobacterial abundance, microcystin, and geosmin in a eutrophic drinking-water reservoir using a 14-year dataset. *Lake and Reservoir Management* **33**: 32-48.

Hwang, S.J., K. Kim, C. Park, W. Seo, B.G. Choi, H.S. Eum, M.H. Park, H.R. Noh, Y.B. Sim and J.K. Shin. 2016. Hydrometeorological effects on water quality variability in Paldang Reservoir, confluent area of the South-Han River-North-Han River-Gyeongan Stream, Korea. *Korean Journal of Ecology and Environment* **49**: 354-374.

Hwang, S.J., Y.B. Sim, B.G. Choi, K. Kim, C. Park, W. Seo, M.H. Park, S.W. Lee and J.K. Shin. 2017. Rainfall and hydrological comparative analysis of water quality variability in Euiam Reservoir, the North-Han River, Korea. *Korean Journal of Ecology and Environment* **50**: 29-45.

Kim, S.H., J.H. Park and B. Kim. 2021. Prediction of cyanobacteria harmful algal blooms in reservoir using machine learning and deep learning. *Journal of Korea Water Resources Association* **54**: 1167-1181.

Kratzert, F., D. Klotz, M. Herrnegger, A.K. Sampson, S. Hochreiter and G.S. Nearing. 2019. Toward improved predictions in ungauged basins: Exploiting the power of machine learning. *Water Resources Research* **55**: 11344-11354.

LeCun, Y., Y. Bengio and G. Hinton. 2015. Deep learning. *Nature* **521**: 436-444.

Lee, E., E.H. Na and K. Kim. 2012. The establishment of water quality forecasting system for preemptive water quality management. *Rural Resources* **54**: 50-55.

Liu, Y., Z. Wang, H. Guo, S. Yu and H. Sheng. 2013. Modelling the effect of weather conditions on cyanobacterial bloom outbreaks in Lake Dianchi: a rough decision-adjusted logistic regression model. *Environmental Modeling and Assessment* **18**: 199-207.

Luo, Y., K. Yang, Z.Y. Yu, J.Y. Chen, Y.F. Xu, X.L. Zhou and Y. Yang. 2017. Dynamic monitoring and prediction of Dianchi Lake cyanobacteria outbreaks in the context of rapid urbanization. *Environmental Science and Pollution Research* **24**: 5335-5348.

Maier, H.R. and G.C. Dandy. 2000. Neural networks for the prediction and forecasting of water resources variables: a review of modelling issues and applications. *Environmental Modelling and Software* **15**: 101-124.

Millie, D.F., G.R. Weckman, G.L. Fahnenstiel, H.J. Carrick, E. Ardjmand, W.A. Young II, M.J. Sayers and R.A. Shuchman. 2014. Using artificial intelligence for cyanoHAB niche modeling: discovery and visualization of *Microcystis*-environmental associations within western Lake Erie. *Canadian Journal of Fisheries and Aquatic Sciences* **71**:

1642-1654.

Ministry of Environment-National Institute of Environmental Research (MOE-NIER). 2020. A Manual of Algal Alert System. NIER-GP2020-019. Incheon, Republic of Korea.

Mitrovic, S.M., L. Hardwick and F. Dorani. 2010. Use of flow management to mitigate cyanobacterial blooms in the Lower Darling River, Australia. *Journal of Plankton Research* **33**: 229-241.

Moe, S.J., S. Haande and R.M. Couture. 2016. Climate change, cyanobacteria blooms and ecological status of lakes: a Bayesian network approach. *Ecological Modelling* **337**: 330-347.

Mooij, W.M., D. Trolle, E. Jeppesen, G. Arhonditsis, P.V. Belolipetsky, D.B.R. Chitamwebwa, A.G. Degermendzhy, D,L. DeAngelis, L.N.D. Domis, A.S. Downing, J.A. Elliott, C.R. Fragoso, U. Gaedke, S.N. Genova, R.D. Gulati, L. Hakanson, D.P. Hamilton, M.R. Hipsey, J. `t Hoen, S. Hulsmann, F.H. Los, V. Makler-Pick, T. Petzoldt, I.G. Prokopkin, K. Rinke, S.A. Schep, K. Tominaga, A.A. van Dam, E.H. van Nes, S.A. Wells and J.H. Janse. 2010. Challenges and opportunities for integrating lake ecosystem modelling approaches. *Aquatic Ecology* **44**: 633-667.

Nichols, S., R. Norris, W. Maher and M. Thoms. 2006. Ecological effects of serial impoundment on the Cotter River, Australia. *Hydrobiologia* **572**: 255-273.

O'Hara, R.B. and D.J. Kotze. 2010. Do not log-transform count data. *Methods in Ecology and Evolution* **1**: 118-122.

Office of Science and Technology Policy (OSTP). 2016. Harmful Algal Blooms and Hypoxia Comprehensive Research Plan and Action Strategy: An Interagency Report. National Science and Technology Council Subcommittee on Ocean Science and Technology, USA. 94p.

Ostfeld, A., A. Tubaltzev, M. Rom, L. Kronaveter, T. Zohary and G. Gal. 2015. Coupled data-driven evolutionary algorithm for toxic cyanobacteria (blue-green algae) forecasting in Lake Kinneret. *Journal of Water Resources Planning and Management* **141**: 04014069-13

Paerl, H.W. 2014. Mitigating harmful cyanobacterial blooms in a human- and climatically-impacted World. *Life* **4**: 988-1012.

Paerl, H.W. and D.F. Millie. 1996. Physiological ecology of toxic aquatic cyanobacteria. *Phycologia* **35**: 160-167.

Paerl, H.W. and J. Huisman. 2009. Climate change: a catalyst for global expansion of harmful cyanobacterial blooms. *Environmental Microbiology Reports* **1**: 27-37.

Paerl, H.W. and T.G. Otten. 2013. Harmful cyanobacterial blooms: causes, consequences and controls. *Microbial Ecology* **65**: 995-1010.

Page, T., P.J. Smith, K.J. Beven, I.D. Jones, J.A. Elliott, S.C. Maberly, E.B. Mackay, M. De Ville and H. Feuchtmayr. 2018. Adaptive forecasting of phytoplankton communities. *Water Research* **134**: 74-85.

Peters, D.P., K.M. Havstad, J. Cushing, C. Tweedie, O. Fuenres and N. Villanueva-Rosales. 2014. Harnessing the power of big data: Infusing the scientific method with machine learning to transform ecology. *Ecosphere* **5**: 1-15.

Qin, B., J. Deng, K. Shi, J. Wang, J. Brookes, J. Zhou, Y. Zhang, G. Zhu, H.W. Pearl and L. Wu. 2021. Extreme climate anomalies enhancing cyanobacterial blooms in eutrophic Lake Taihu, China. *Water Resources Research* **57**: e2020WR029371.

Qin, B., W. Li, G. Zhu, Y. Zhang, T. Wu and G. Gao. 2015. Cyanobacterial bloom management through integrated monitoring and forecasting in large shallow eutrophic Lake Taihu (China). *Journal of Hazardous Materials* **287**: 356-363.

Raps, S., K. Wyman, H.W. Siegelman and P.G. Falkowski. 1983. Adaptation of the cyanobacterium *Microcystis aeruginosa* to light intensity. *Plant Physiology* **72**: 829-832.

Recknagel, F., M. French, P. Harkonen and K.I. Yabunaka. 1997. Artificial neural network approach for modelling and prediction of algal blooms. *Ecological Modelling* **96**: 11-28.

Recknagel, F., P.T. Orr and H.Q. Cao. 2014. Inductive reasoning and forecasting of population dynamics of *Cylindrospermopsis raciborskii* in three sub-tropical reservoirs by evolutionary computation. *Harmful Algae* **31**: 26-34.

Recknagel, F., P.T. Orr, M. Bartkow, A. Swanepoel and H. Cao. 2017. Early warning of limit-exceeding concentrations of cyanobacteria and cyanotoxins in drinking water reservoirs by inferential modelling. *Harmful Algae* **69**: 18-27.

Recknagel, F., T. Fukushima, T. Hanazato, N. Takamura and H. Wilson. 1998. Modelling and prediction of phyto- and zooplankton dynamics in Lake Kasumigaura by artificial neural networks. *Lakes and Reservoirs: Research and Management* **3**: 123-133.

Reynolds, C.S. and A.E. Walsby. 1975. Water-blooms. *Biological Reviews* **50**: 437-481.

Reynolds, C.S., R.L. Oliver and A.E. Walsby. 1987. Cyanobacterial dominance: the role of buoyancy regulation in dynamic lake environments. *New Zealand Journal of Marine and Freshwater Research* **21**: 379-390.

Rousso, B.Z., E. Bertone, R. Stewart and D.P. Hamilton. 2020. A systematic literature review of forecasting and predictive models for cyanobacteria blooms in freshwater lakes. *Water Research* **182**: 115959.

Schindler, D.W. 2012. The dilemma of controlling cultural eutrophication of lakes. *Proceedings of The Royal Society B* **279**: 4322-4333.

Schuwirth, N., F. Borgwardt, S. Domisch, M. Friedrichs, M. Kattwinkel, D. Kneis, M. Kuemmerlen, S.D. Langhans, J. Martinez-Lopez and P. Vermeiren. 2019. How to make ecological models useful for environmental management. *Ecological Modelling* **411**: 108784.

Sheng, H., H. Liu, C. Wang, H. Guo, Y. Liu and Y. Yang. 2012. Analysis of cyanobacteria bloom in the Waihai part of Dianchi lake, China. *Ecological Informatics* **10**: 37-48.

Shin, J.K. and Y. Park. 2018. Spatiotemporal and longitudinal variability of hydro-meteorology, basic water quality and

dominant algal assemblages in the eight weir pools of regulated river (Nakdong). *Korean Journal of Ecology and Environment* **51**: 268-286.

Shin, J.K., B.G. Kang and S.J. Hwang. 2016. Water-blooms (green-tide) dynamics of algae alert system and rainfall-hydrological effects in Daecheong Reservoir, Korea. *Korean Journal of Ecology and Environment* **49**: 153-175.

Shin, J.K., Y. Park, N.Y. Kim and S.J. Hwang. 2022. Downstream transport of geosmin based on harmful cyanobacterial outbreak upstream in a reservoir cascade. *International Journal of Environmnetal Research and Public Health* **19**: 9294.

Sibanda, M., O. Mutanga, V.G. Chimonyo, A.D. Clulow, C. Shoko, D. Mazvimavi, T. Dube and T. Mabhaudhi. 2021. Application of drone technologies in surface water resources monitoring and assessment: A systematic review of progress, challenges, and opportunities in the global south. *Drones* **5**: 84.

Summers, E.J. and J.L. Ryder. 2023. A critical review of operational strategies for the management of harmful algal blooms (HABs) in inland reservoirs. *Journal of Environmental Management* **330**: 117141.

Teles, L.O., E. Pereira, M. Saker and V. Vasconcelos. 2008. Virtual experimentation on cyanobacterial bloom dynamics and its application to a temperate reservoir (Torrão, Portugal). *Lakes and Reservoirs: Research and Management* **13**: 135-143.

Tromas, N., N. Fortin, L. Bedrani, Y. Terrat, P. Cardoso, D. Bird, C.W. Greer and B.J. Shapiro. 2017. Characterising and predicting cyanobacterial blooms in an 8-year amplicon sequencing time course. *The ISME* (*International Society for Microbial Ecology*) *Journal* **11**: 1746-1763.

van Eck, N.J. and L. Waltman. 2007. Bibliometric mapping of the computational intelligence field. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* **15**: 625-645.

van Eck, N.J. and L.Waltman. 2009. VOSviewer: A Computer Program for Bibliometric Mapping. Technical Report ERS-2009-005-LIS, Erasmus University Rotterdam, Erasmus Research Institute of Management. Rotterdam, The Netherlands. 19p. http://hdl.handle.net/1765/14841

van Eck, N.J. and L. Waltman. 2010. Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics* **84**: 523-538.

van Eck, N.J., L.R.Waltman, E.C.M. Noyons and R.K. Buter. 2010a. Automatic term identification for bibliometric mapping. *Scientometrics* **82**: 581-596.

van Eck, N.J., L.Waltman, R. Dekker and J. van den Berg. 2010b. A comparison of two techniques for bibliometric mapping: Multidimensional scaling and VOS. *Journal of the American Society for Information Science and Technology* **61**: 2405-2416.

Waaijer, C.J.F., C.A. van Bochove and N.J. van Eck. 2011. On the map: Nature and Science editorials. *Scientometrics* **86**: 99-112.

Waltman, L., N.J. van Eck and E.C.M. Noyons. 2010. A unified approach to mapping and clustering of bibliometric networks. *Journal of Informetrics* **4**: 629-635.

Wang, H., R. Zhu, J. Zhang, L.Y. Ni, H. Shen and P. Xie. 2018. A novel and convenient method for early warning of algal cell density by chlorophyll fluorescence parameters and its application in a highland lake. *Frontiers in Plant Science* **9**: 869.

Watanabe, M.F., K. Harada, W.W. Carmichael and H. Fujiki. 1996. Toxic *Microcystis*. CRC Press, Boca Raton, London, U.K. 262p.

Wei, B., N. Sugiura and T. Maekawa. 2001. Use of artificial neural network in the prediction of algal blooms. *Water Research* **35**: 2022-2028.

Welk, A., F. Recknagel, H. Cao, W.S. Chan and A. Talib. 2008. Rule-based agents for forecasting algal population dynamics in freshwater lakes discovered by hybrid evolutionary algorithms. *Ecological Informatics* **3**: 46-54.

Wilkinson, G.M., S.R. Carpenter, J.J. Cole, M.L. Pace, R.D. Batt, C.D. Buelo and J.T. Kurtzweil. 2018. Early warning signals precede cyanobacterial blooms in multiple whole-lake experiments. *Ecological Monographs* **88**: 188-203.

World Health Organization (WHO). 2011. Management of Cyanobacteria in Drinking-water Supplies: Information for Regulators and Water Suppliers. Technical Brief WHO/FWC/WSH/15.03. 11p.

Xiao, X., J. He, H. Huang, T.R. Miller, G. Christakos, E.S. Reichwaldt, A. Ghadouani, S. Lin, X. Xu and J. Shi. 2017. A novel single-parameter approach for forecasting algal blooms. *Water Research* **108**: 222-231.

Yabunaka, K., M. Hosomi and A. Murakami. 1997. Novel application of a backpropagation artificial neural network model formulated to predict algal bloom. *Water Science and Technology* **36**: 89-97.