Journal of Korean Society on Water Environment, Vol. 36, No. 6, pp. 611-620 (November, 2020)
pISSN 2289-0971 eISSN 2289-098X https://doi.org/10.15681/KSWE.2020.36.6.611

자료기반 물환경 모델의 현황 및 발전 방향

차윤경[a,†] · 신지훈[b] · 김영우[c]

서울시립대학교 환경공학과

Data-Driven Modeling of Freshwater Aquatic Systems: Status and Prospects

YoonKyung Cha[a,†] · Jihoon Shin[b] · YoungWoo Kim[c]

School of Environmental Engineering, University of Seoul

(Received 05 October 2020, Revised 10 November 2020, Accepted 12 November 2020)

Abstract

Although process-based models have been a preferred approach for modeling freshwater aquatic systems over extended time intervals, the increasing utility of data-driven models in a big data environment has made the data-driven models increasingly popular in recent decades. In this study, international peer-reviewed journals for the relevant fields were searched in the Web of Science Core Collection, and an extensive literature review, which included total 2,984 articles published during the last two decades (2000-2020), was performed. The review results indicated that the rate of increase in the number of published studies using data-driven models exceeded those using process-based models since 2010. The increase in the use of data-driven models was partly attributable to the increasing availability of data from new data sources, e.g., remotely sensed hyperspectral or multispectral data. Consistently throughout the past two decades, South Korea has been one of the top ten countries in which the greatest number of studies using the data-driven models were published. Among the major data-driven approaches, i.e., artificial neural network, decision tree, and Bayesian model, were illustrated with case studies. Based on the review, this study aimed to inform the current state of knowledge regarding the biogeochemical water quality and ecological models using data-driven approaches, and provide the remaining challenges and future prospects.

Key words : Data-driven approach, Ecological model, Freshwater aquatic system, Water quality model, Water quality and resources management

[a,†] Corresponding author, 교수(Professor), ykcha@uos.ac.kr, https://orcid.org/0000-0001-9638-9476
[b] 박사과정(Ph.D. Student), sjh3473@uos.ac.kr, https://orcid.org/0000-0002-3763-9938
[c] 박사과정(Ph.D. Student), youngwoo0508@uos.ac.kr, https://orcid.org/0000-0001-7123-6190

1. Introduction

하천과 호소의 이화학·생물학적 수질 및 생물군집의 반응을 모의하는 수질 및 수생태 모델은 물환경 관리를 위한 의사결정의 과학적 근거를 제공하여 왔다(Arhonditsis et al., 2006; Robson et al., 2018). 이러한 모델은 관리가 요구되는 주요 현안인 부영양화에 따른 녹조 발생, 수온 성층 형성으로 인한 심층 혐기화, 기후 변화에 따른 생물 군집 변화 등을 예측함으로써 물환경 관리를 위한 핵심 도구로 활용되어 왔다(Arhonditsis and Brett, 2004; Jeong, 2012). 모델의 유형은 다양한 스펙트럼을 가지고 있으나 크게 기작기반(process-based)과 자료기반(data-driven)으로 분류할 수 있다. 기작기반 모델은 시스템에 대한 개념적 이해를 바탕으로 이를 구동하는 물리학적, 화학적, 생물학적 기작을 일련의 연립 방정식으로 수식화한 체계인 반면, 자료기반 모델은 통계기법이나 기계학습 알고리즘 적용을 통해 자료에서 발견된 패턴 및 변수 간 상관성을 기반으로 시스템에 대한 예측을 수행한다.

전통적으로 물환경 관련 분야에서는 시스템의 구조와 기능에 대한 통찰력을 제공하고 예측 결과에 대한 인과적 추론이 가능한 기작기반 모델을 선호하여 왔다(Baker et al., 2018; Shen, 2018). 그러나 기작기반 모델은 미지의 시스템에 대해 적용이 어렵고, 여러 방정식의 매개변수 추정을 위해 고해상도의 다변량 자료를 요구하며, 충분한 자료를 확보하지 못할 시 예측 성능을 보장할 수 없는 단점이 있다(Dormann et al., 2012). 반면에 자료기반 모델은 시스템에 대한 사전 지식 없이 적용 가능하고, 기작기반 모델에 비해 구조가 간단하며 모델 구축 및 계산에 소요되는 시간이 적으면서도 예측력이 뛰어난 장점을 지니고 있다(Altunkaynak and Wang, 2011; Cloern and Jassby, 2010). 더욱이 최근 다량의 자료가 빠른 속도로 생성되고 이에 발맞추어 자료의 저장 능력 및 계산력이 비약적으로 향상됨에 따라 자료기반 모델을 적용하기에 적합한 환경이 조성되고 있다(Schuwirth et al., 2019). 자료기반 모델 중 특히 인공지능 기반 기계학습 알고리즘은 기존 기법으로는 처리가 어려운 대용량 고차원(high-dimensional) 자료에 대해서도 숨은 패턴 및 관계를 밝혀내 새로운 정보를 발굴하고 정확한 예측 수행하는 데 있어 탁월한 능력을 보여 다양한 학문 분야에 걸쳐 각광받고 있다(Hutchinson et al., 2019; Kratzert et al., 2019; LeCun et al., 2015). 물환경 관련 분야에서도 축적되는 자료의 활용성을 높이고 모델의 예측 성능을 향상시키기 위해 자료기반 모델의 수요가 증가하고 있다(Peters et al., 2014; Schuwirth et al., 2019). 최근 이공학 연구에서부터 일상생활에까지 자료기반 모델이 가져온 영향과 성과를 감안할 때(Kratzert et al., 2019; LeCun et al., 2015), 물환경 분야에서도 국가 경쟁력을 확보하기 위해 관련 전문가들이 자료기반 모델링 기술을 제고하고 이에 수반되는 막대한 잠재력을 선제적으로 실현하기 위한 노력을 기울여야 하는 시점이라고 판단된다. 따라서 본 연구에서는 물환경 분야 국제학술지를 대상으로 수행한 문헌조사를 통해 자료기반 모델링 기술의 세계적 추세 및 현황을 파악하고, 우리나라의 기술 수준을 평가하고자 하였다. 이에 더해 물환경 분야에서 적용성이 높고 전망이 밝은 자료기반 모델링 기법을 선정하여 소개하고, 국내 적용 사례 위주로 설명하였다. 마지막으로 자료기반 모델링 기술의 과제에 대한 고찰을 바탕으로 향후 발전 방향을 제시하였다.

2. Current status

2.1 문헌조사

자료기반 물환경 모델링 연구의 현황 및 추세를 파악하기 위해 국제 학술논문에 대한 문헌조사를 수행하였다. 문헌조사는 국제 학술 데이터베이스인 Web of Science 핵심 컬렉션에서 관련 주제어 검색을 통해 2001년 1월부터 2020년 6월까지 최근 20여 년간 국제 학술지에 게재된 물환경 모델링 연구 성과를 대상으로 이루어졌다. 수체는 담수체인 하천과 호소를 포함하였고, 모델은 생지화학적 수질과 수생태 모델을 포함하였다. 수리·수문 혹은 유역 모델의 경우 수질이나 수생태와 연계 모델은 포함하였고, 단일 모델은 제외하였다. 중복성 검토를 거친 후 최종적으로 선정된 논문은 총 2,944편이었다.

선정된 논문에서 사용된 모델링 기법을 조사하기 위해 먼저 기작기반과 자료기반으로 대분류하였으며, 자료기반 기법을 사용한 경우 회귀 모델(regression model), 인공신경망(artificial neural network), 결정나무(decision tree), 베이지안 모델(Bayesian model), 서포트 벡터 머신(support vector machine), 유전 알고리즘(genetic algorithm), 퍼지 모델(fuzzy model), 클러스터링/차원축소(clustering/dimension reduction), 기타로 소분류하였다. 모델링의 목적 및 대상을 유추하기 위하여 종속변수(혹은 출력변수)를 조사하였으며, 결과를 제시하기 위해 이화학적 수질 항목, 수생태 항목으로 분류하였고 생물 항목 중 클로로필-a와 녹조 관련 항목은 독립적으로 구분하여 분류하였다. 자료기반 물환경 모델링 기술의 국가별 역량을 비교하기 위해 연구 대상 수체의 위치 및 연구자의 국적을 기준으로 국가를 구분하였다. 논문에서 사용된 기법, 종속변수 및 국가가 다수인 경우 중복 계수하였다.

2.2 현황 및 추세

문헌조사 결과 자료기반 모델이 총 1,784편, 기작기반 모델이 총 1,200편의 논문에 게재되어 최근 20년간 물환경 모델링 방법으로써 자료기반 모델의 활용도가 더 높은 것으로 나타났다(Fig. 1a). 논문의 누적 게재 편수를 비교했을 때 2009년까지는 비슷한 증가 추세를 보이다가 2010년을 기점으로 자료기반 모델링 기법 사용의 증가율이 기작기반 모델 사용 증가율보다 확연히 높아지는 추세를 확인할 수 있었으며(Fig. 1b), 이는 빅데이터 시대의 도래와 맞물려 증가해 온 자료기반 모델링 기술 수요를 반영한다고 하겠다.

자료기반 물환경 모델링 연구를 위해 가장 많이 사용되어 온 방법은 전통적 통계기반 기법이자 가장 기본적인 기계학습 기법인 회귀 모델이며, 더 복잡한 기계학습 기반 알고리즘의 강세에도 불구하고 오히려 최근 10년간 사용 증가 추세

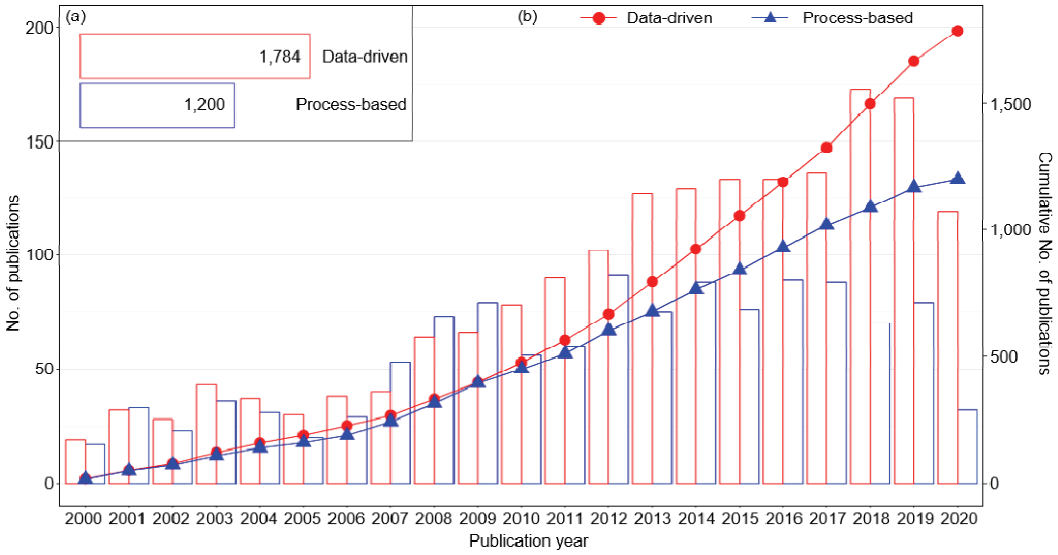

Fig. 1. Number of published studies (a: total and b: annual) during two decades (2000-2020) using data-driven versus process-based approaches for modeling freshwater aquatic systems.

가 강화되는 양상을 띠었다(Fig. 2). 회귀 모델 다음으로는 인공신경망 모델을 사용한 논문 편수가 다른 자료기반 모델보다 눈에 띄게 많았으며, 그 뒤로 결정나무, 클러스터링/차원축소, 베이지안 모델의 순서로 활용도가 높은 것으로 나타났다(Fig. 2). 이 중 클러스터링/차원축소는 다변량 자료의 전처리 및 변수 선택, 특징 추출을 목적으로 다른 모델과 연계 활용도가 높은 것으로 판단된다(Kim and Seo, 2015; Kim et al., 2017).

국제학술지에 게재된 물환경 자료기반 모델링 관련 논문 편수를 기준으로 해당 기술의 국가 역량을 평가하였다(Fig. 3). 논문 편수는 정량적 척도이나 연구의 다양성이나 질적 측면과도 상관성을 보인다고 가정하고 제한적이나마 국가 역량 비교를 위한 지표로 사용하였다. 최근 20년간 게재된 논문 편수를 비교한 결과 지역별로는 북미(총 659편), 유럽연합(총 385편), 동북 아시아(총 323편)의 순서로 많은 연구 성과를 거둔 것으로 나타났다(Fig. 3a). 국가별로는 미국이 최근 20년간 꾸준히 가장 활발하게 관련 연구를 진행하였다(Fig. 3b). 중국은 5위권 이내의 순위를 유지하다가 2010년도부터는 매해 미국 다음으로 많은 연구 성과를 보였으며, 캐나다의 경우 중국과는 반대로 5위권 내 순위를 유지하던 2000년대와 비교해 2010년대에는 10위권으로 국가 순위가 하락하는 경향을 나타냈다(Fig. 3b). 우리나라는 호주와 유사하게 연도별로 큰 등락폭을 보였음에도 불구하고 최근 20년간 꾸준히 10위권 이내의 양적 연구 성과를 거두었다(Fig.

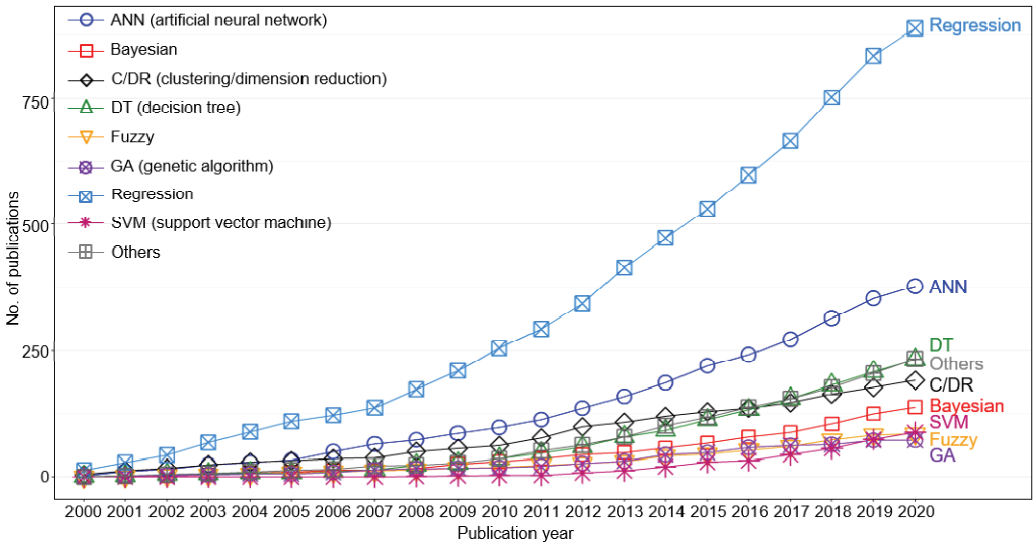

Fig. 2. Annual total number of published studies during two decades (2000-2020) using data-driven approaches, which were classified into nine categories, for modeling freshwater aquatic systems.

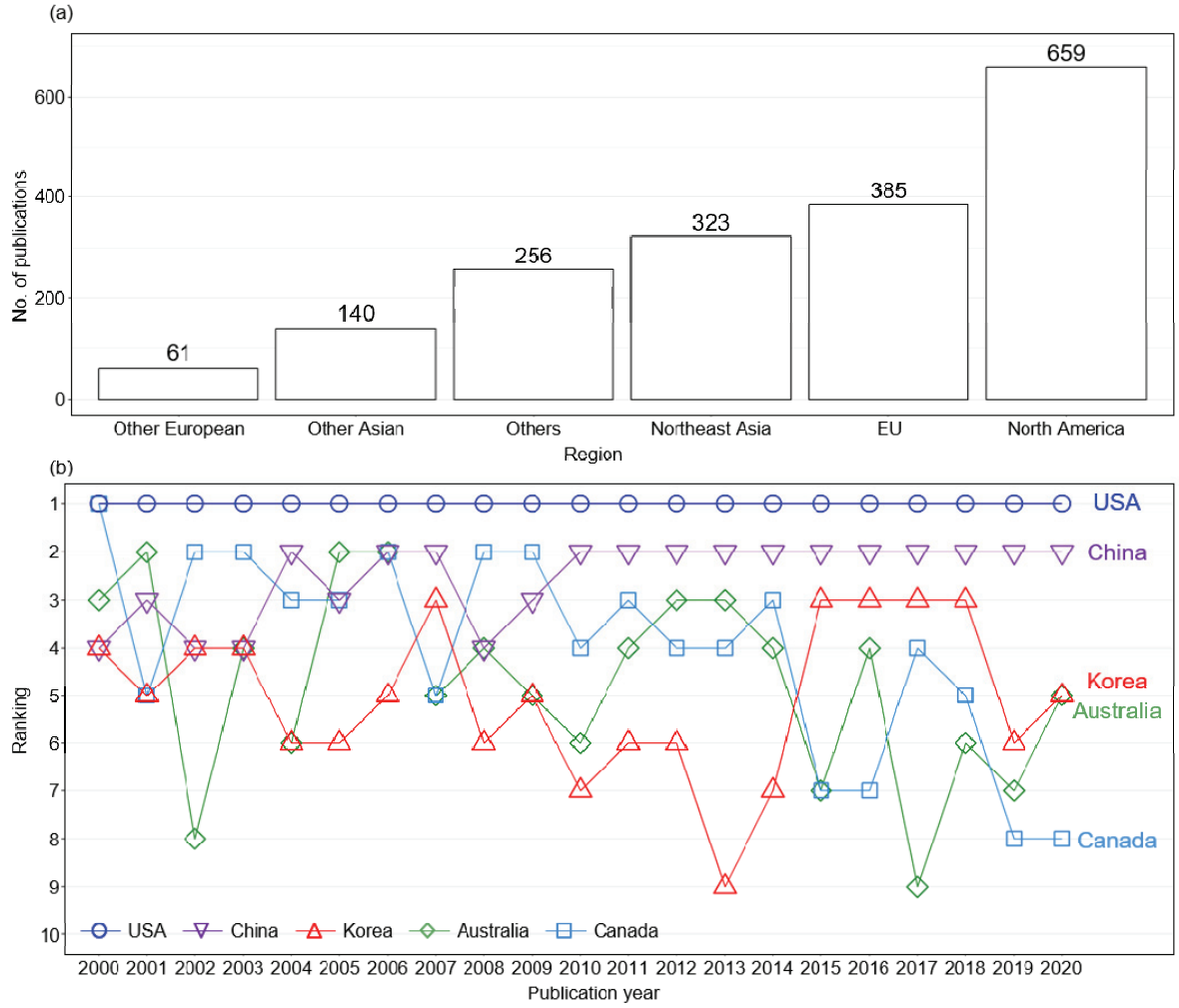

Fig. 3. Performance rankings among a) regions and b) countries evaluated based on the number of published studies (during 2000-2020) that used data-driven approaches for modeling freshwater aquatic systems.

3b). 국가별 누적 논문 편수를 비교하였을 때 우리나라는 5위에 해당하는 연구 성과를 축적하여 물환경 분야에서 세계적으로 우수한 자료기반 모델링 기술을 보유한 국가인 것으로 평가할 수 있다.

모델링의 대상이 되는 종속변수를 이화학적 수질, 수생태, 남조류(세포수, 생체량, 생물부피 등), 클로로필-a의 네 가지 범주로 분류한 후 문헌조사를 진행하였다. 수체 내 수질오염 탐지 및 관리 지표로 사용되는 이화학적 수질 항목은 수온, 용존산소량, 생물화학적 산소요구량, 다양한 영양물질 농도 등을 포함하며, 비교적 계측이 간단하고 자료의 시공간적 해상도가 상대적으로 높은 경향을 보인다. 남조류 관련 항목과 클로로필-a 농도는 정체성 수역인 호소의 부영양화 현상 관리를 위한 주요 지표이나 우리나라의 경우 4대강 사업 이후 증가한 녹조 문제로 인해 보 대표지점을 비롯해 하천에서도 가용한 모니터링 자료가 축적되고 있다. 외부오염원의 수리 및 수질 영향에 대한 생물 반응을 평가하기 위해 지표로 사용되는 수생태 항목은 어류, 저서성대형무척추동물, 동식물플랑크톤 등 생물분류군의 종분포, 종풍부도, 군집지수, 기능군 특성 등 다양한 항목을 포함한다. 생태 항목은 계측 비용 및 시간 소요가 크고 수질 항목에 비해 장기간 누적 영향을 파악하기 위한 지표이므로 자료의 시공간적 해상도가 낮은 특성을 보인다.

문헌조사를 통해 최근 20년간 게재 편수를 계수한 결과 이화학(총 836편), 수생태(총 739편), 클로로필-a(총 256편), 남조류(총 150편)의 순서로 활발한 연구가 진행된 것으로 나타났다(Fig. 4a). 상대적으로 풍부한 자료가 뒷받침되는 이화학적 수질 항목에 비해 자료의 가용성이 낮은 수생태 항목의 모델링에서 자료기반 기법의 높은 활용도는 주목할 만한 점이다. 수생태 모델 중 먹이망 모델은 주로 기작기반인 데 반해, 생물군집 예측 모델은 기후, 수리 · 수문, 토지이용 변화와 생물군집 반응 간 상관성에 기초한 자료기반 기법을 주로 사용한다. 이는 수생태 계측자료의 중요성과 특히 현재 미흡한 호소 수생태자료에 대한 데이터베이스 구축의 시급성을 시사한다고 하겠다. 항목별로 최근 20년간 추세를 살펴본 결과 모든 수질 · 수생태 항목에 대한 자료기반 모델링 연구가 증가하는 경향을 나타냈다(Fig. 4b).

한편 문헌조사를 통해 나타난 주목할 만한 추세는 원격탐사(remote sensing) 영상을 활용한 자료기반 수질 모델의 급

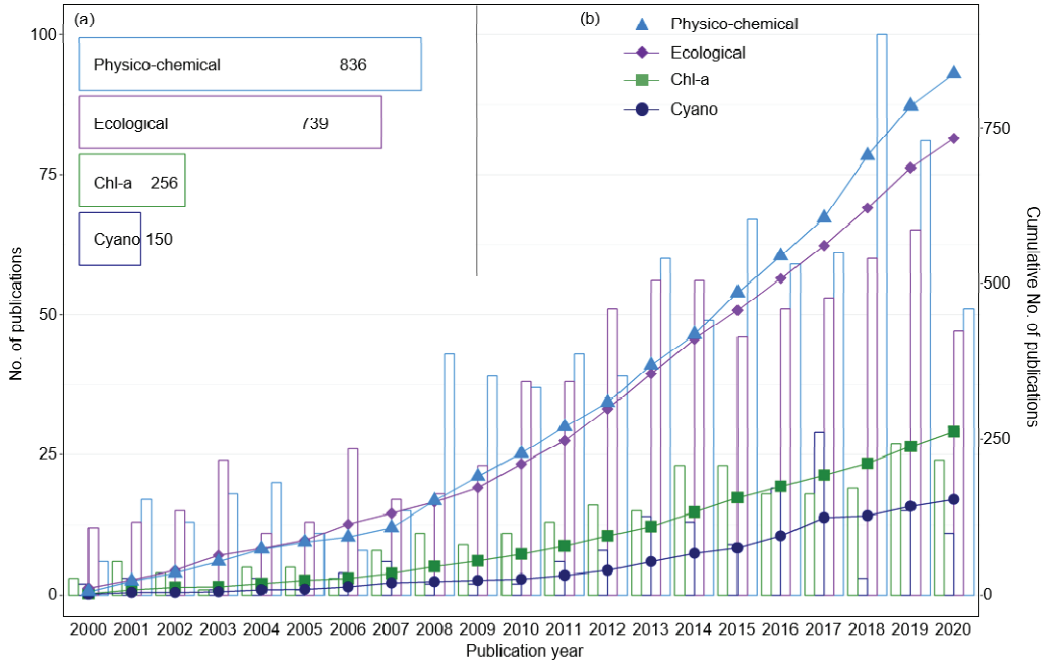

Fig. 4. Number of published studies (a: total and b: annual total and annual cumulative) during two decades (2000-2020) using data-driven approaches for modeling four different categories of response variables (physico-chemical, ecological, chlorophyll-a, and cyanobacterial abundance) in freshwater aquatic systems.

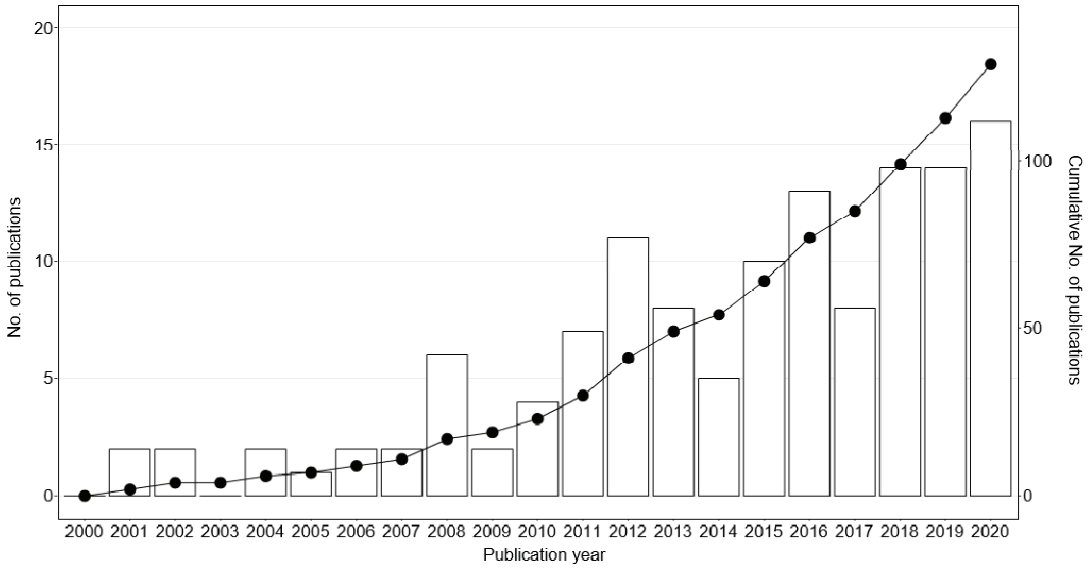

Fig. 5. Annual total number of published studies that used remote sensing data and data-driven approaches for modeling freshwater aquatic systems.

격한 증가이며, 특히 2010년대부터 연구 성과가 크게 증가한 것으로 나타났다(Fig. 5). 원격탐사 자료를 활용한 수질 모델은 원격탐사 영상의 반사율 정보로부터 총 부유물질(total suspended solids)이나 클로로필-a(chlorophyll-a), 피코시아닌(phycocyanin) 등 분광특성을 갖는 수질 항목의 농도를 추정하는 모델로서, 기존의 점단위로 존재하는 수질 자료를 면단위로 확대하는 의미를 지닌다. 과거 원격모니터링은 주로 육상이나 해양에 적용되어 왔으며 담수역의 경우 탁도가 높아 간섭효과가 크고, 수체의 경계에 인접한 육상의 영향으로 인해 적용이 용이하지 않았다. 그러나 공간적 해상도가 높은 위성영상 자료가 생산되고, 또한 초분광(hyperspectral) 센서로부터 높은 분광 해상도를 가진 자료가 생산되면서 담수역에 대한 원격모니터링 연구 사례가 증가하고 있다. 원격모니터링을 위한 모델은 기작기반이라 할 수 있는 고유분광특성(inherent optical properties, IOPs) 변환 모델이 널리 사용되어 있으나(Randolph et al., 2008), Fig. 5에서 나타난 것처럼

자료기반 모델, 예를 들어 반사율 밴드의 비를 사용하는 간단한 모델이나 인공신경망 기반 알고리즘을 사용하는 사례도 늘어나는 추세로(Pyo et al., 2020; Yim et al., 2020), 향후 원격모니터링을 위한 방법론은 더욱 다양화 될 전망이다.

2.3 자료기반 모델링 기법 적용 사례

자료기반 물환경 모델링 기법의 현황 및 추세 분석 결과, 현재까지 성과가 많고 향후 적용성 또한 높을 것으로 판단되는 인공신경망, 결정나무, 베이지안 모델에 대하여 설명하고, 문헌조사 대상 논문 중 국내 수체에 적용된 연구 위주로 모델의 적용 사례를 소개하고자 한다. 국내 수체 적용 사례가 없는 경우 국내 연구자에 의해 수행된 연구 사례를 들기로 한다.

2.3.1 인공신경망

인공신경망은 인간의 뇌에서 뉴런이 정보를 처리하는 기능을 모사하여 계산을 수행하는 모델이다. 뉴런을 구성요소로 하는 층의 다겹 구조로 이루어져 있으며, 입력층으로 투입된 자료는 은닉층을 거치며 변환되어 출력층을 통해 출력값이 생성된다. 인공신경망은 개발 후 적용 단계에서 난관에 봉착할 때마다 그에 대한 해결책을 찾으며 흥망성쇠를 반복해 왔다. 예를 들어 과적합 문제에 대해 드롭아웃(drop-out)의 개념을 도입하여 해결하고, 역전파기법(backpropagation)의 단점인 기울기 소실(vanishing gradient)은 가중치 초기화(weight initialization) 등을 통해 극복하였다. 특히 하드웨어 성능의 발달로 인해 인공신경망의 약점으로 지적되어 온 느린 학습속도 문제가 보완되고, 빅데이터 시대에 접어들며 대용량 자료에 구조 적합성이 뛰어난 딥러닝 기반 인공신경망이 기존의 기계학습 모델을 넘어서는 성능을 보이며 인공신경망은 전성기를 맞이하게 되었다. 인공신경망 알고리즘은 매우 다양하며 자료의 특성이나 모델링의 목적에 따라, 예를 들어 시계열 등 시퀀스(sequence) 자료에는 long short-term memory (LSTM), 영상 자료에는 convolutional neural network (CNN), 특징 추출이나 차원 축소 등 비지도학습(unsupervised learning)을 위한 목적으로는 autoencoder 등을 사용할 수 있다.

국내 물환경 분야에서 인공신경망은 녹조 관리를 위한 연구에 주로 적용되어 왔다. 예를 들어 Kim, Hong et al., (2019)은 인공신공망을 사용하여 낙동강 강정고령보 구간의 클로로필-a를 예측하였으며, Park et al. (2015)은 주암호와 영산호의 클로로필-a를 예측하고, 서포트 벡터 머신과 성능 비교를 수행하였다. Kim and Seo (2015)은 ANN 앙상블 모델을 사용하여 총인, 총질소를 포함한 낙동강의 수질을 예보하였으며, 데이터 불균형으로 인한 예측력 저하의 문제를 클러스터링 기법을 사용하여 해결하였다. 또한 녹조 원격모니터링을 위한 모델링 기법으로 많이 사용되었는데 예를 들어 Yim et al. (2020)은 stacked-autoencoder deep neural network (SAE-DNN)을 사용하여 남조류 고유 색소인 피코시아닌 예측에 있어서 유용성이 높은 초분광 파장(wavelengths)을 선정하였으며, Pyo et al. (2020)은 stacked-autoencoder artificial neural network (SAE-ANN)를 사용하여 대기보정 및 녹조예측을 수행하였다. 또한, Park et al. (2017)은 인공신경망과 기작기반 모델인 Environmental Fluid Dynamics Code (EFDC)을 함께 사용하여 클로로필-a 농도의 공간적 분포를 예측하였다.

2.3.2 결정나무

결정나무는 독립변수에 대해 일련의 기준을 적용하는 가지생성과, 자료를 기준에 부합하도록 가장 동질적인 결과로 분류하는 잎 생성으로 구성된 모델이며, 대체적으로 지도학습(supervised learning)을 사용한다. 결정나무는 인공신경망 계열의 모델에 비해 일반적으로 소요시간이 짧고 결과 해석이 용이한 장점을 가지고 있다. 단일 결정나무 모델은 결과에 편향이 발생하거나 과분산을 나타내는 단점이 있으나, 여러 개의 단일 결정나무 모델을 통합하는 보팅(voting), 배깅(bagging), 부스팅(boosting), 스태킹(stacking), 랜덤 포레스트(random forest) 등 다양한 앙상블(ensemble) 기법을 적용할 경우 이러한 단점을 보완하고 모델의 예측력을 안정적으로 향상시킬 수 있다.

결정나무 모델의 국내 적용 사례를 살펴보면 먼저 다양한 앙상블 기법을 사용하여 성능을 비교하는 연구가 많았으며, 앙상블 기법 중 랜덤 포레스트에 대한 선호도가 가장 높았다. 예를 들어, Choi and Seo (2018)은 대장균군의 수질 기준 달성 여부를 예측하기 위해 배깅과 랜덤 포레스트를 사용하여 단일 모델인 classification and regression tree (CART)와 성능을 비교하였다. Shin et al. (2017)은 자료의 불균형도(data imbalance)가 모델의 분류 성능 저하에 미치는 영향을 분석하기 위해 배깅, 부스팅, 랜덤 포레스트를 함께 사용하였다. 또한 결정나무 모델은 생물종 혹은 분류군의 분포를 예측하기 위한 방법으로 많이 사용되었다. 예를 들어, Kwon et al. (2015)은 CART와 랜덤 포레스트를 사용하여 섬진강을 포함한 5대강 수계에서 22종의 고유 어류종 분포를 예측하였고, Kim, Cho et al. (2019)은 랜덤 포레스트를 사용하여 4대강 수계에서의 규조류 분포를 예측하였다.

2.3.3 베이지안 모델

베이지안 모델은 조건부 확률을 정의하는 베이즈 정리(Bayes' theorem)에 기초하여, 사전 지식을 반영하는 사전확률(prior probability)과 현재 획득한 자료의 우도(likelihood)를 종합한 사후확률(posterior probability) 분포로써 모델의 매개변수(parameter)를 추정한다. 베이지안 모델은 매개변수나 예측 결과의 불확실성 정량화에 최적화된 방법론적 체계를 제공하기 때문에 기후변화에 수반되는 물환경 여건 변화로 인해 예측의 정확성이 담보될 수 없는 환경에서 예측 결과의 신뢰도 제고하는 방안으로 활용될 수 있다. 베이지안 기법 중 베이지안 계층화 모델(Bayesian hierarchical model)은 변수 간 상관성을 하위 계층(lower level)에서 규명하고, 이 상관성의 시공간적 변동성을 상위 계층(higher level)에서 규명하는 계층적 구조를 가지고 있기 때문에 학습된 모델을 새로운 시공간에 적용할 수 있는(transferable) 장점이 있다.

또한 베이지안 네트워크(Bayesian network)는 절점(node)과 화살표(arrow)로 구성된 모델의 구조를 시각화하여 표출한다. 모델 구조에서 각 절점은 개별 변수, 통합 지표(integrated index) 혹은 서브모델(sub-model) 등이 될 수 있고, 서브모델로는 구조기반 모델도 사용할 수 있어 베이지안 네트워크는 모델 융합·통합 플랫폼으로 사용될 수 있는 장점을 가진다. 또한 모델 구조에서 화살표는 변수 혹은 모델 간 의존성을 의미하기 때문에 자료기반 모델임에도 불구하고 결과에 대한 인과적 해석이 가능하다.

베이지안 모델의 국내 적용 사례를 살펴보면 먼저 Cha et al. (2014)은 베이지안 포아송 허들 모델(Bayesian Poisson hurdle model)을 사용하여 팔당호에서 녹조의 발생 여부와 발생 시 강도를 예측하였다. 또한 베이지안 계층화 모델을 적용 사례가 다수 있었는데 Cha et al. (2016)은 낙동강 상류에서 하류 흐름에 따른 클로로필-a와 유량 간 상관성의 점진적 변화를 규명하였고, Cha et al. (2017)은 4대강 16개 보에 걸쳐 녹조 발생에 미치는 수온과 체류시간의 상대적 중요도를 정량화하였다. 국내 적용 사례는 아니지만 베이지안 네트워크의 적용 사례를 살펴보면 Stow and Cha (2013)는 호소 내 총인과 클로로필-a 간 상관성에 대한 인과적 추론의 타당성 여부를 규명하였고, Lee et al. (2014)는 축산 폐수 처리용 인공 못에서 에스트로겐(estrogen) 호르몬의 거동을 설명하는 기작기반 융합 모델을 개발하였다.

3. Limitations and future prospects

문헌조사를 통해 확인한 것처럼 물환경 분야에서도 자료기반 모델링을 통해 얻은 연구 및 관리 성과가 축적되고 있으며, 향후에도 자료기반 모델링 기술 수요의 증가 추세는 지속될 전망이다. 사전 지식 및 개입의 최소화, 유용한 정보 추출, 뛰어난 예측 성능 등 자료기반 모델은 많은 장점을 지니고 있지만 해결해야 할 과제 또한 안고 있으며, 이는 자료기반 모델의 물환경 시스템 적용 시에도 부담으로 작용할 것이다. 따라서 현재까지 지적되어 온 자료기반 모델의 한계점을 통해 시사점을 도출하고, 과제를 극복하기 위한 방안을 향후 발전방향으로 제시하고자 한다.

먼저 자료기반 모델은 예측결과에 대한 해석력(interpretability)의 부재라는 큰 과제를 안고 있다(Hutchinson et al., 2019; Schuwirth et al., 2019). 자료기반 모델은 내부 구조에 접근하기 어려워 예측 결과에 대한 결정 근거를 이해하기 어렵기 때문에 흔히, 블랙박스(black box)라고 불린다(Castelvecchi, 2016). 통계적 회귀 모델의 경우 물질 수지(mass balance) 등 간단한 기작을 포함할 수 있고 단일 결정나무 모델은 의사결정의 과정을 시각화하기 때문에 어느 정도 해석이 가능하지만, 인공신경망과 같은 순수 자료기반 모델일수록 결과의 해석이 어려운 경향을 나타낸다(Papernot and McDaniel, 2018). 성능의 우수성이 담보된다고 하더라도 해석력이 낮다면 자료기반 모델이 수생태 시스템에 대한 새로운 통찰력이나 이해를 제공하기 어렵기 때문에 적용의 걸림돌로 작용할 수 있다. 이에 따라 자료기반 모델의 해석력을 높이기 위한 연구가 활발히 진행되고 있다(Alain and Bengio, 2016; Bau et al., 2017; Dabkowski and Gal, 2017). 이 중 개별 예측에 대해 각 입력변수의 상대적 기여도를 규명하기 위해 사용되는 local interpretable modelagnostic explanations (LIME)이나 Shapley additive explanations (SHAP)은 다른 입력변수의 값에 대한 조건부 해석이 가능하다는 점에서 물환경 분야에서 유용성이 높은 기법으로 전망된다(Lundberg and Lee, 2017; Ribeiro et al., 2016).

자료기반 모델의 또 다른 과제는 전이성(transferability) 문제이다. 전이성이란, 모델의 학습 자료와는 다른 자료, 즉 새로운 장소, 시간이나 생물 종에 모델을 적용했을 때 나타나는 예측력을 의미한다(Wenger and Olden, 2012; Yates et al., 2018). 기작기반 모델의 경우 이론적 뒷받침이 된다면 모델을 새로운 시스템에 적용할 수 있지만, 자료기반 모델은 별도의 작업이나 확인 과정을 거치치 않고는 전이성을 장담하기 어려운 측면이 있다. 자료기반 모델이 낮은 전이성을 보이는 원인으로는 허위 상관성(spurious correlations)에 기반해 예측이 이루어지거나, 변수 간 상관성의 시공간적 변동성을 고려하지 못하는 경우를 들 수 있다(Dormann et al., 2012). 하지만, 자료가 부족한 환경(새로운 장소, 시간, 생물종)에 대해서도 모델이 관리 의사결정을 지원하는 예측 정보를 제공한다면 실용적 가치를 제고할 수 있을 것이며, 이와 같은 맥락에서 자료기반 모델의 전이성을 확보하기 위한 연구가 활발히 진행되고 있다.

자료기반 모델의 전이성을 높이기 위한 기법으로는 전이학습(transfer learning)을 들 수 있는데, 전이 학습이란 어떤 다량의 자료로 모델을 학습시키고 이 과정에서 학습된 특징(learned features)을 새로운, 소량의 자료로 학습시킬 모델에 전이하는 기법이다. 다량의 자료로부터 학습된 특징이 보편성(generality)을 띤다면 전이 학습을 통해 모델의 전이성을 강화하고 자료 부족으로 인한 예측력 저하의 문제를 해결할 수 있다(Ma et al., 2019; Tian et al., 2019). 또한 자료기반 모델의 전이성을 제고하기 위한 기법으로 앙상블 학습(ensemble learning)과 모델 검증(validation)을 들 수 있다. 앙상블 학습은 다수의 단일 모델 결과를 통합하는 기법으로서, 단일 모델의 다양성이 클수록 전이성에 대한 앙상블 학습의 효과를 기대할 수 있다(Michielsen et al., 2016; Zhai and Chen, 2018). 모델 검증은 모델의 초매개변수(hyperparameter)를 최적화하여 과적합(overfitting)을 방지함으로서 전이성을 높이는 기법이며, 무작위 교차검증(random cross-validation) 등 다양한 검증 자료의 사용 방법에 따라 전이성이 달라질 수 있다(Ozesmi et al., 2006; Rohani et al., 2018).

한편 센서 및 센서 플랫폼 기술의 발달로 인해 물환경 분야에서도 자료 형태의 다양화가 가속화될 전망이다. 위성영상 자료 이외에도 무인 혹은 유인 항공기에 탑재된 센서로부터 생성된 초분광 자료의 가용성이 증대될 것이고, 실시간 모니터링 시스템은 확대 구축되어 수리·수문 및 다양한 수질 자료를 고빈도로 생산할 것이다(Schuwirth et al., 2019). 뿐만 아니라 대규모 오믹스(omics) 자료나 환경 DNA (environmental DNA, eDNA) 자료의 생산으로 인해 새로운

자료의 활용성 요구가 커질 전망이다(Peters et al., 2014; Schuwirth et al., 2019). 가용한 자료 용량의 증가는 대용량 자료에 강점을 보이는 딥러닝 등의 기술에 대한 개발 수요를 지속적으로 창출할 것이다(LeCun et al., 2015; Shen, 2018). 또한 자료 출처가 다양화됨에 따라 특정 목적을 위해 취합한 데이터셋 내 자료의 형태나 생성 주기도 다양해질 것이기 때문에 이를 통합할 수 있는 다중규모(multi-scale), 다중모드(multi-modal)의 모델링 기법, 예를 들어 베이지안 네트워크나 생성적 적대 신경망(generative adversarial network; GAN)의 활용이 증가할 것으로 예상할 수 있다(Robson et al., 2018; Shen, 2018).

우리나라 물환경 모델은 세계적인 기술 추세를 반영하는 동시에 국가의 물환경 정책 흐름을 수용하는 방향으로 발전할 것이다. 2018년 '물관리기본법'이 제정되어 유역 중심 물관리, 통합 물관리 등 국가 물환경 관리 정책의 방향이 설정되었으며, 이에 따라 수량-수질-수생태 통합 모델에 대한 개발 수요도 지속적으로 증대될 전망이다. 따라서 자료기반 수질 및 수생태 모델에서도 수리 · 수문 혹은 유역 모델과의 연계 · 통합 방안 마련의 중요성이 더욱 대두될 것이다. 더 나아가 불확실성 정량화에 대한 강점을 가진 자료기반 방법론을 고도화하여, 기후변화에 따른 불확실성 증가를 고려한 의사결정 및 리스크 평가에 활용함으로서 미래 물관리 위기대응 능력 제고에 이바지할 필요가 있다.

4. Conclusions

수질 및 수생태 모델은 수체에서 일어나는 현상에 대한 개념적 이해를 검증하거나 보완하는 학문적 가치뿐만 아니라 관리 목표 수립 및 달성 여부 평가를 위한 정량적 근거를 마련하는 등 실용적 가치 또한 지니고 있다. 모델이 물환경 분야에서 필수적인 역할을 담당해온 만큼 물환경 관리 선진화를 달성하기 위해서는 기작기반이나 자료기반 모델링 기술을 고도화하기 위한 노력을 경주해야 하지만, 동시에 어떠한 모델도 완전하지 않다는 사실을 직시함으로서 중요한 시사점을 얻을 수 있다. 모델이 고도화되더라도 모델링 방법 및 인식체계로부터 기인한 한계점을 완전히 극복할 수는 없기 때문에 방법론의 교류와 융합을 통해 기작기반과 자료기반 모델의 장점과 단점을 상호보완하는 조정 과정을 견지해야 한다. 자료기반 모델은 사전 지식이나 수동적 개입 없이도 자료로부터 상관성을 추출하고 예측을 수행하지만 이를 인과적으로 혹은 외삽(extrapolation)하여 해석할 경우 오류를 범할 수 있다. 그러나 자료기반 모델에서 발견된 상관성이 기작기반 모델에서 새로운 가설로서 검증 과정을 거친다면 해석력과 전이성을 겸비한 기술적 성과를 달성할 수 있을 것이다. 반대로 자료기반 모델에서 종종 발견되는 예상치 못한 상관성이나 특징은 새로운 이론 수립이나 사전 지식의 수정 및 보완을 위한 시작점으로 작용할 수 있으며, 이는 기작기반 모델에서 수식이나 가정의 수정을 통해 반영될 수 있다.

최근 물관리 일원화로 통합물관리를 위한 제도적 기반이 마련됨에 따라 수질 · 수생태 모델이 당면한 과제 해결의 시급성은 더 부각될 예정이다. 특히 기존 기작기반 모델은 모델에 수리 · 수문 기작으로부터 이화학적, 생물학적 기작을 더할수록 모델 성능이 현저히 저하되는 결과를 나타내어 왔다. 기작기반 모델의 복잡성 증가와 성능 향상 간의 관계가 불투명한 상태에서 방향을 전환하여 자료기반 모델 적용을 적극적으로 추진할 필요가 있으나, 이를 위해서는 수리 · 수문, 수질, 수생태 모듈을 아우르고 기작기반 모델과 융합할 수 있는 통합 플랫폼 개발이 선제되어야 한다. 이와 함께 자료기반 모델의 장점을 살려 예측의 불확실성과 단계적 예측에 따른 불확실성 전파를 정량화하여 수량-수질-수생태의 통합관리를 위해 보다 신뢰성 있는 정보를 제공할 필요가 있다.

Acknowledgements

본 결과물은 환경부의 재원으로 한국환경산업기술원 수생태계 건강성 확보 기술개발사업의 지원을 받아 연구되었습니다(과제번호: 202000305003).

References

Alain, G. and Bengio, Y. (2016). *Understanding intermediate layers using linear classifier probes*, arXiv preprint arXiv:1610.01644.

Altunkaynak, A. and Wang, K. H. (2011). A comparative study of hydrodynamic model and expert system related models for prediction of total suspended solids concentrations in Apalachicola Bay, *Journal of Hydrology*, 400(3-4), 353-363.

Arhonditsis, G. B. and Brett, M. T. (2004). Evaluation of the current state of mechanistic aquatic biogeochemical modeling, *Marine Ecology Progress Series*, 271, 13-26.

Arhonditsis, G. B., Adams-VanHarn, B. A., Nielsen, L., Stow, C. A., and Reckhow, K. H. (2006). Evaluation of the current state of mechanistic aquatic biogeochemical modeling: citation analysis and future perspectives, *Environmental science & technology*, 40(21), 6547-6554.

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

Bau, D., Zhou, B., Khosla, A., Oliva, A., and Torralba, A. (2017). Network dissection: Quantifying interpretability of deep visual representations, *Proceedings of the IEEE conference on computer vision and pattern recognition*, 6541-6549.

Castelvecchi, D. (2016). Can we open the black box of AI?, *Nature News*, 538(7623), 20.

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 research*, 124, 11-19.

Cha, Y., Park, S. S., Kim, K., Byeon, M., and Stow, C. A. (2014). Probabilistic prediction of cyanobacteria abundance

in a Korean reservoir using a Bayesian Poisson model, *Water Resources Research*, 50(3), 2518-2532.

Cha, Y., Park, S. S., Lee, H. W., and Stow, C. A. (2016). A Bayesian hierarchical approach to model seasonal algal variability along an upstream to downstream river gradient, *Water Resources Research*, 52(1), 348-357.

Choi, S. Y. and Seo, I. W. (2018). Prediction of fecal coliform using logistic regression and tree-based classification models in the North Han River, South Korea, *Journal of Hydro-environment Research*, 21, 96-108.

Cloern, J. E. and Jassby, A. D. (2010). Patterns and scales of phytoplankton variability in estuarine-coastal ecosystems, *Estuaries and coasts*, 33(2), 230-241.

Dabkowski, P. and Gal, Y. (2017). Real time image saliency for black box classifiers, *Advances in Neural Information Processing Systems,* 6967-6976.

Dormann, C. F., Schymanski, S. J., Cabral, J., Chuine, I., Graham, C., Hartig, F., Kearney, M., Morin, X., Römermann, C., Schröder, B., and Singer, A. (2012). Correlation and process in species distribution models: bridging a dichotomy, *Journal of Biogeography*, 39(12), 2119-2131.

Hutchinson, L., Steiert, B., Soubret, A., Wagg, J., Phipps, A., Peck, R., Charoin, J. E., and Ribba, B. (2019). Models and machines: how deep learning will take clinical pharmacology to the next level, *CPT: pharmacometrics & systems pharmacology*, 8(3), 131.

Jeong, S. U. (2012). The state of the art of lake water quality modeling and applications, *Magazine of the Korean Society of Agricultural Engineers*, 54(1), 56-69. [Korean Literature]

Kim, H. G., Hong, S., Jeong, K. S., Kim, D. K., and Joo, G. J. (2019). Determination of sensitive variables regardless of hydrological alteration in artificial neural network model of chlorophyll a: case study of Nakdong River, *Ecological Modelling*, 398, 67-76.

Kim, H. K., Cho, I. H., Hwang, E. A., Kim, Y. J., and Kim, B. H. (2019). Benthic diatom communities in Korean estuaries: Species appearances in relation to environmental variables, *International journal of environmental research and public health*, 16(15), 2681.

Kim, S. E. and Seo, I. W. (2015). Artificial neural network ensemble modeling with conjunctive data clustering for water quality prediction in rivers, *Journal of Hydro-Environment Research*, 9(3), 325-339.

Kim, S. E., Seo, I. W., and Choi, S. Y. (2017). Assessment of water quality variation of a monitoring network using exploratory factor analysis and empirical orthogonal function, *Environmental Modelling & Software*, 94, 21-35.

Kratzert, F., Klotz, D., Herrnegger, M., Sampson, A. K., Hochreiter, S., and Nearing, G. S. (2019). Toward improved predictions in ungauged basins: Exploiting the power of machine learning, *Water Resources Research*, 55(12), 11344-11354.

Kwon, Y. S., Bae, M. J., Hwang, S. J., Kim, S. H., and Park, Y. S. (2015). Predicting potential impacts of climate change on freshwater fish in Korea, *Ecological Informatics*, 29, 156-165.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning, N*ature*, 521(7553), 436-444.

Lee, B., Kullman, S. W., Yost, E., Meyer, M. T., Worley Davis, L., Williams, C. M., and Reckhow, K. H. (2014). A Bayesian network model for assessing natural estrogen fate and transport in a swine waste lagoon, *Integrated environmental assessment and management*, 10(4), 511-521.

Lundberg, S. M. and Lee, S. I. (2017). A unified approach to interpreting model predictions, *In Advances in neural information processing systems,* 4765-4774.

Ma, J., Cheng, J. C., Lin, C., Tan, Y., and Zhang, J. (2019). Improving air quality prediction accuracy at larger temporal resolutions using deep learning and transfer learning techniques, *Atmospheric Environment*, 214, 116885.

Michielsen, A., Kalantari, Z., Lyon, S. W., and Liljegren, E. (2016). Predicting and communicating flood risk of transport infrastructure based on watershed characteristics, *Journal of environmental management*, 182, 505-518.

Özesmi, S. L., Tan, C. O., and Özesmi, U. (2006). Methodological issues in building, training, and testing artificial neural networks in ecological applications, *Ecological Modelling*, 195(1-2), 83-93.

Papernot, N. and McDaniel, P. (2018). *Deep k-nearest neighbors: Towards confident, interpretable and robust deep learning*, arXiv preprint arXiv:1803.04765.

Park, Y., Cho, K. H., Park, J., Cha, S. M., and Kim, J. H. (2015). 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, 31-41.

Park, Y., Pyo, J., Kwon, Y. S., Cha, Y., Lee, H., Kang, T., and Cho, K. H. (2017). Evaluating physico-chemical influences on cyanobacterial blooms using hyperspectral images in inland water, Korea, *Water research*, 126, 319-328.

Peters, D. P., Havstad, K. M., Cushing, J., Tweedie, C., Fuentes, O., and Villanueva-Rosales, N. (2014). Harnessing the power of big data: infusing the scientific method with machine learning to transform ecology, *Ecosphere*, 5(6), 1-15.

Pyo, J., Duan, H., Ligaray, M., Kim, M., Baek, S., Kwon, Y. S., Lee, H., Kang, T., Kim, K., Cha, Y., and Cho, K. H. (2020). An integrative remote sensing application of stacked autoencoder for atmospheric correction and cyanobacteria estimation using hyperspectral imagery, *Remote Sensing*, 12(7), 1073.

Randolph, K., Wilson, J., Tedesco, L., Li, L., Pascual, D. L., and Soyeux, E. (2008). Hyperspectral remote sensing of cyanobacteria in turbid productive water using optically active pigments, chlorophyll a and phycocyanin, *Remote Sensing of Environment*, 112(11), 4009-4019.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier, *Proceedings of the 22nd ACM SIGKDD international*

conference on knowledge discovery and data mining, 1135-1144.

Robson, B. J., Arhonditsis, G. B., Baird, M. E., Brebion, J., Edwards, K. F., Geoffroy, L., Hébert, M. P., van Dongen-Vogels, V., Jones, E. M., Kruk, C., Mongin, M., Shimoda, Y., Skerratt, J. H., Trevathan-Tackett, S. M., Wild-Allen, K., Kong, X., and Steven, A. (2018). Towards evidence-based parameter values and priors for aquatic ecosystem modelling, *Environmental modelling & software*, 100, 74-81.

Rohani, A., Taki, M., and Abdollahpour, M. (2018). A novel soft computing model (Gaussian process regression with K-fold cross validation) for daily and monthly solar radiation forecasting (Part: I), *Renewable Energy*, 115, 411-422.

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

Shen, C. (2018). A transdisciplinary review of deep learning research and its relevance for water resources scientists, *Water Resources Research*, 54(11), 8558-8593.

Shin, J., Yoon, S., and Cha, Y. (2017). Prediction of cyanobacteria blooms in the lower Han River (South Korea) using ensemble learning algorithms, *Desalination and Water Treatment*, 84, 31-39.

Stow, C. A. and Cha, Y. (2013). Are chlorophyll a－total phosphorus correlations useful for inference and prediction?, *Environmental science & technology*, 47(8), 3768-3773.

Tian, W., Liao, Z., and Wang, X. (2019). Transfer learning for neural network model in chlorophyll-a dynamics prediction, *Environmental Science and Pollution Research*, 26(29), 29857-29871.

Wenger, S. J. and Olden, J. D. (2012). Assessing transferability of ecological models: an underappreciated aspect of statistical validation, *Methods in Ecology and Evolution*, 3(2), 260-267.

Yates, K. L., Bouchet, P. J., Caley, M. J., Mengersen, K., Randin, C. F., Parnell, S., Fielding, A. H., Bamford, A. J., Stephan, B., Barbosa, A. M., Dormann, C. F., Elith, J., Embling, C. B., Ervin, G. N., Fisher, R., Gould, S., Graf, R. F., Gregr, E. J., Halpin, P. N., Heikkinen, R. K., Heinänen, S., Mannocci, L., Mellin, C., Mesgaran, M. B., Moreno-Amat, E., Mormede, S., Novaczek, E., Oppel, S., Crespo, G. O., Peterson, A. T., Rapacciuolo, G., Roberts, J. J., Ross, R. E., Scales, K. L., Schoeman, D., Snelgrove, P., Sundblad, G., Thuiller, W., Torres, L. G., Verbruggen, H., Wang, L., Wenger, S., Whittingham, M. J., Zharikov, Y., Zurell, D., and Sequeira, A. M. (2018). Outstanding challenges in the transferability of ecological models, *Trends in ecology & evolution*, 33(10), 790-802.

Yim, I., Shin, J., Lee, H., Park, S., Nam, G., Kang, T., Cho, K. H., and Cha, Y. (2020). Deep learning-based retrieval of cyanobacteria pigment in inland water for in-situ and airborne hyperspectral data, *Ecological Indicators*, 110, 105879.

Zhai, B. and Chen, J. (2018). Development of a stacked ensemble model for forecasting and analyzing daily average PM2. 5 concentrations in Beijing, China, *Science of The Total Environment*, 635, 644-658.