Journal of Korean Society on Water Environment, Vol. 33, No. 3, pp. 326-333 (May, 2017)
pISSN 2289-0971 eISSN 2289-098X https://doi.org/10.15681/KSWE.2017.33.3.326

저수지 CO_2 배출량 산정을 위한 기계학습 모델의 적용

유지수 · 정세웅† · 박형석

충북대학교 환경공학과

Applications of Machine Learning Models for the Estimation of Reservoir CO_2 Emissions

Jisu Yoo · Se-Woong Chung† · Hyung-Seok Park

Department of Environmental Engineering, Chungbuk National University
(Received 28 February 2017, Revised 18 May 2017, Accepted 26 May 2017)

Abstract

The lakes and reservoirs have been reported as important sources of carbon emissions to the atmosphere in many countries. Although field experiments and theoretical investigations based on the fundamental gas exchange theory have proposed the quantitative amounts of Net Atmospheric Flux (NAF) in various climate regions, there are still large uncertainties at the global scale estimation. Mechanistic models can be used for understanding and estimating the temporal and spatial variations of the NAFs considering complicated hydrodynamic and biogeochemical processes in a reservoir, but these models require extensive and expensive datasets and model parameters. On the other hand, data driven machine learning (ML) algorithms are likely to be alternative tools to estimate the NAFs in responding to independent environmental variables. The objective of this study was to develop random forest (RF) and multi-layer artificial neural network (ANN) models for the estimation of the daily CO_2 NAFs in Daecheong Reservoir located in Geum River of Korea, and compare the models performance against the multiple linear regression (MLR) model that proposed in the previous study (Chung et al., 2016). As a result, the RF and ANN models showed much enhanced performance in the estimation of the high NAF values, while MLR model significantly under estimated them. Across validation with 10-fold random samplings was applied to evaluate the performance of three models, and indicated that the ANN model is best, and followed by RF and MLR models.

Key words : Artificial neural network, Carbon emission, Daecheong Reservoir, Machine learning, Random forest

1. Introduction

지난 수십 년간 육상 담수는 대기 중 이산화탄소(CO_2)와 메탄(CH4)의 주요 배출원으로 연구되어 왔다(Guerin et al., 2006; Kemenes et al., 2007; Rudd et al., 1993; Therrien et al., 2005). Cole et al. (2007)은 호소, 저수지와 같은 담수가 육상계의 자연적 및 인위적 배출원으로부터 연간 약 1.9 Pg C/yr의 탄소를 수용하며, 이중 최소한 0.8 Pg C/yr는 수체와 대기 경계면에서의 가스 교환을 통해 대기로 배출된다고 산정하였다. Louis et al. (2000)은 온대 저수지를 대상으로 수면-대기 경계면을 통해 배출되는 CO_2 플럭스를 직접 측정하는 연구를 수행하였으며, Enting (2002)은 관측된 대기 중의 CO_2 농도 자료를 역 모델링을 통해 탄소 플럭스를 추정하는 방법을 제안하였다. 그러나, 하천, 호소, 저수지 수면으로부터 대기 중으로 배출되는 탄소량은 위도, 기후, 수문 조건, 계절, 그리고 유역의 토지 이용특성과 해당 수체의 이화학적 특성 등 다양한 요소에 따라 변동이 크며, 전 지구적 규모의 산정결과에도 불확실성이 많이 내포되어 있다(UNESCO/IHA, 2010). 다양한 환경조건에서 담수로부터 배출되는 CO_2 순 대기 배출 플럭스(Net Atmospheric Flux, NAF)를 보다 정확히 산정하기 위해서는 현장 실험과 병행하여 효과적인 시뮬레이션 모델을 개발하는 것이 중요하다.

담수 수면과 대기 경계면을 통한 탄소 교환을 정량적으로 모의하는 방법은 크게 기계적 모델(mechanistic models)을 사용하는 것과 데이터 기반의 통계적 모델(statistical models)을 사용하는 것이 가능하다(Safaie et al., 2016; Wang et al., 2016). 기계적 모델은 질량, 운동량, 에너지 보존방정식을 수치적으로 해석하여 수체 내 수리동력학적, 생지화학적 과정을 정량적으로 해석하는 방법이며, 유역으로부터 입자상과 용존상 유기물의 유입, 침강, 분해 과정, 수체 내에서 1차 생산에 의해 생성되는 유기물의 대사과정과 다양한 경로를 통해 생성된 무기탄소의 대기-수체 교환과정을 정량적으로 해석하는데 매우 유용하다(Cole and

† To whom correspondence should be addressed.
schung@chungbuk.ac.kr

Wells, 2015). 그러나, 기계적 모델은 광범위한 경계조건 자료와 복잡한 모델의 매개변수를 요구하므로 자료가 불충분 하거나 모의기작이 단순화 될 경우, 모의결과의 불확실성이 증가할 수 있으며, 모델의 구축과 구동에 소요되는 비용도 상대적으로 큰 것이 단점이다. 반면, 통계적 모델링 방법은 다양한 독립변수로부터 종속변수를 통계적으로 산정하는 기법으로써 다중선형회귀모델이 가장 간단한 예이다(Dillon and Rigler, 1975; Recknagel et al., 1997; Whitehead and Hornberger, 1984). 최근에는 데이터 기반 기계학습(machine learning, ML) 알고리즘이 급속히 발전하면서, 이러한 기법을 활용할 경우 예측 성능이 우수하고 운영도 효율적인 모델을 손쉽게 구축할 수 있다는 장점이 있다.

본 연구의 목적은 Chung et al. (2016)이 금강의 대청호에서 일별 CO_2 NAF 산정을 위해 개발한 다중선형회귀모델의 한계점인 높은 NAF 값을 과소산정 하는 문제를 해결하기 위한 방안으로 기계학습(ML) 모델을 개발하고 그 성능을 평가하는데 있다. 선행연구에서 개발한 다중회귀모델(MLR)은 최적 매개변수 선정을 위해 단계적 전진 방법을 사용하였다. 모델의 독립변수로 선정된 pH, 수온(WT), 클로로필 a(Chl-a) 농도, 풍속(WS), 알칼리도(Alk)를 사용하여 모의한 결과, 2012년과 2013년에 NAF의 계절별 변동 특성을 각각 75.7%와 84.0%까지 설명하였으나, NAF가 크게 발생하는 시기에 실측값과의 편차가 매우 커 표준오차가 높아지는 단점을 보였다. 이러한 문제점을 해결하고자 본 연구에서 적용한 ML 알고리즘은 앙상블 의사결정 나무 모델의 한 종류인 랜덤포레스트(random forest, RF) 기법과 다층 퍼셉트론 인공신경망(artificial neural networks, ANN) 기법이며, 개발된 모형의 성능을 선행연구에서 제시한 MLR과 비교, 평가 하였다.

2. Materials and Methods

2.1. 연구대상 지역

대청호는 1981년 대청댐 준공과 함께 형성된 인공 저수지이며, 금강 하구로부터 약 150 km 상류에 위치한다. 상류 용담댐의 유역면적을 제외한 총 유역면적은 3,204 km^2이며, 금강 수계 전체 면적의 32.4%에 해당한다. 대청호는 유역 내 인근 도시들의 물 공급을 위한 최대 상수원으로 사용되고 있으며, 이외에도 관개용수, 수력발전, 홍수조절 등의 다목적으로 사용되고 있다. 대청호는 추동(R1), 댐 앞(R2), 문의취수탑(R3), 장계(R4), 회남(R5), 대정리(R6) 총 6개 지점에서 환경부 정기수질측정망이 운영되고 있다(Fig. 1). 이 중 수질자동측정망이 운영되고 있는 회남 대교(R5) 지점을 본 연구의 대표지점으로 선정하였다. 수질 자동측정망은 수온, pH, DO, 전기전도도(EC)와 같은 기본항목을 5분단위로 측정하며, TOC, Chl-a, T-N, T-P를 포함한 선택항목을 시간단위로 측정한다. Alk 자료는 K-water의 C 정수장에서 원수 수질관리를 위해 일별로 실측한 자료를 제공 받아 사용하였다.

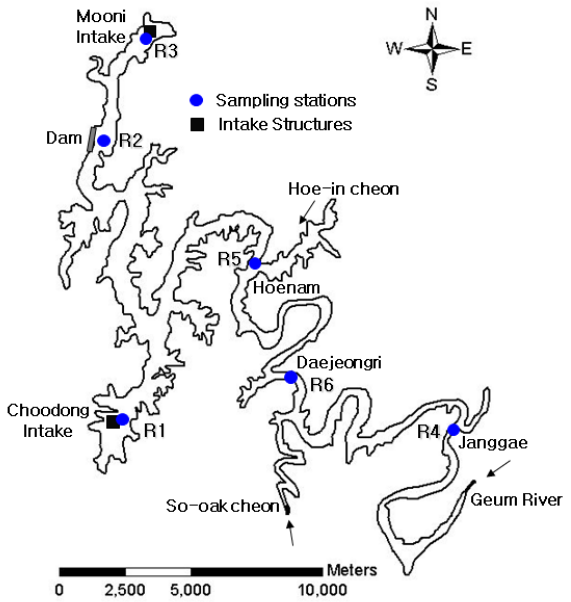

Fig. 1. Layout of Daecheong Reservoir and locations of monitoring stations (Chung et al., 2016).

2.2. 다중회귀 및 기계학습 모델

2.2.1. 다중회귀(MLR)모델

MLR모델(식 (1))은 주어진 하나의 종속변수(Y)와 여러 개의 독립변수들(X_1, X_2, ..., X_n) 간에 선형 관계(Linear Relationship)가 있음을 가정하고, 주어진 학습데이터를 바탕으로 각 독립변수의 예측력(영향력)인 회귀계수(β)를 추정하는 모형이다(Bae and Kim, 2016). 다중회귀분석은 데이터를 이용하여 최적의 회귀계수를 산정하는 과정이며, 독립변수들 간에 상관관계가 없고 오차항(ε)이 독립적이며 정규분포를 보이는 것을 가정한다. 따라서 이러한 가정이 위반될 때 회귀모델의 모의 결과는 신뢰성을 잃게 된다(Kim et al., 2016). 금번 연구에서 적용한 MLR모델은 R 프로그램의 glm 함수를 사용하여 개발하였다.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \cdots \beta n Xn + \varepsilon \quad (1)$$

2.2.2. 랜덤포레스트(Random forest)

랜덤 포레스트(RF) 모델은 앙상블 학습 기법을 사용한 모델이며, 의사결정나무(Decision Tree)의 확장 개념이다. 훈련 과정에서 구성한 다수의 결정트리로부터 분류 또는 평균 예측치(회귀분석)를 출력한다. 앙상블 학습은 주어진 데이터로부터 여러 개의 모델을 학습한 다음 예측 시 여러 모델의 결과를 종합해 정확도를 높이는 기법이며, 결정 트리는 계층구조로 이루어진 노드(node)와 에지(edge)들의 집합으로 복잡한 문제를 간단한 문제들로 이루어진 계층구조 형태로 나누기 위한 기술이다. 여기서 RF 모델의 노드는 임의성(randomness)에 의해 서로 조금씩 다른 특성을 가지는 트리를 구성한다. 임의화는 각 트리들의 훈련과정에서 진행되며, 임의 학습 데이터 추출 방법을 이용한 앙상블 학습법인 배깅(bagging)과 임의 노드 최적화(Randomized

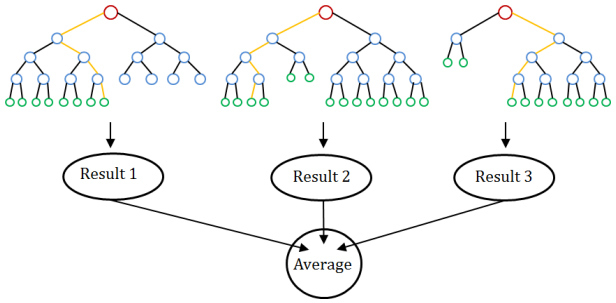

Fig. 2. A schematic description of random forest algorithm.

node optimization)방법이 있다. 배깅은 조금씩 다른 훈련데이터에 대해 기초 분류기를 결합시키는 방법이며, 임의노드최적화는 트리 내에 존재하는 노드마다 분할 함수를 이용해 각 노드의 정보 획득량을 최대가 되도록 하는 방식이다. 최종 결과는 각 의사결정나무의 예측결과를 평균 또는 과반수투표 방식을 이용해 결정한다. RF 모델의 성능을 조정하는 주요 매개변수는 성장하는 트리의 개수(ntree)와 매 분할시 선택하는 무작위로 추출하는 변수의 수(mtry)가 있다. ntree가 너무 적으면 트리들을 구성하고 테스트하는 시간이 줄어드는 반면에 일반화 능력이 떨어져 임의의 입력데이터에 대해 잘못 된 결과를 출력할 확률이 증가한다. 반면 트리의 수가 많을수록 훈련 및 테스트 시간이 증가하지만, 트리의 수가 적은 경우에 비해 비교적 연속적이며 일반화된 결과 값을 출력한다. mtry의 초기값은 분류 문제인 경우에 독립변수 개수(p)의 제곱근(sqrt(p)), 회귀문제인 경우에는 p/3의 값을 사용한다. mtry 값이 작을 경우 과소적합(underfitting)이 일어날 수 있고, 크면 과대적합(overfitting)이 발생하여 적절한 값의 설정이 필요하다. 금번 연구에서 적용한 RF 모델은 R 프로그램의 randomForest 패키지(Breiman and Cutler, 2015)를 사용하여 개발하였으며, ntree는 초기값을 사용하고 mtry의 최적값은 오차를 최소화하고 예측 성능이 가장 우수한 값으로 결정하였다.

2.2.3. 인공신경망(Artificial Neural Network, ANN)

ANN 모형은 통계적 학습 알고리즘의 하나로 생물학의 신경망을 모방하여 만들어 졌다. 인간의 두뇌를 구성하고 있는 기본단위인 뉴런(Neuron)의 인식과정을 수학적 모형으로 일반화시키기 위해 개발된 시스템이며, 함수 추론, 회귀분석, 시계열 예측, 근사 모델링 등에 사용될 수 있다(Brooks et al., 2016; Kim et al., 2016). ANN의 학습기법은 크게 지도학습, 자율학습, 준지도학습으로 나누어진다. 지도학습은 목표값을 포함한 훈련된 데이터로부터 함수를 추론하는 것이며, 주어진 변수사이의 관계를 학습함으로써 입력 값을 받으면 학습 결과를 토대로 새로운 예측 값을 제시한다. 이와 반대로 자율학습은 예측변수에 대한 어떠한 사전정보를 가지고 있지 않기 때문에, 예측 인자만을 이용하여 출력을 계산한다. 준지도학습은 지도학습과 자율학습의 중간 형태이며, 자율학습과정에 적은양의 훈련된 데이터를 입력하여 예측의 정확도 향상시키는 방법이다. 신경망에서 뉴런을 모방한 노드들은 입력층(Input Layer), 은닉층(Hidden Layer), 출력층(Output Layer) 으로 구분된다. 입력층노드(i)에 값이 주어지면 은닉층노드(j)로 전달되며, 은닉층의 노드는 주어진 입력에 따라 활성화된다. 활성화된 노드들이 출력값을 계산하며, 이는 모델의 최종 예측 결과가 된다(Fig. 3). 각 화살표에 가중치(연결강도)가 부여되고, 은닉층노드(j,k,l)에 전달되는 값을 넷 활성화(Net activation)라 한다.

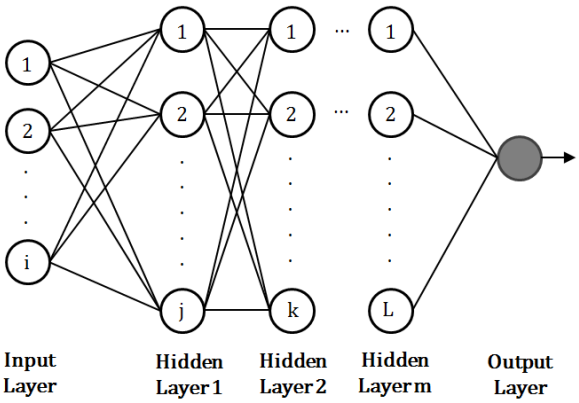

Fig. 3. A schematic description of multi-layer artificial neural network algorithm.

신경망 학습은 이 가중치를 조절하는 작업이다. 입력값 X와 원하는 출력값 Y가 있을 때 입력 값을 입력층에 준 다음 원하는 출력값이 나오는지를 확인하여 가중치를w를 조절한다. 모델 출력값은 에러의 제곱합(Sum of Squared Error, SSE)으로 평가한다(식 (2)). 여기서, $\hat{y_i}$는 출력노드 i의 출력값, y_i는 원하는 출력값을 의미한다. SSE가 0이 아닐 경우 해당 출력 노드에 연결된 은닉노드로, 은닉노드에서 입력노드로 거슬러 올라가며 가중치를 조절하여 0에 근사하게 한다. 이와 같은 과정을 수회 반복하면 적절한 가중치가 발견되며, 이러한 학습방법을 역전파 알고리즘(Back Propagation Algorithm)이라 한다.

$$SSE = \sum (y_i - \hat{y_i})^2 \quad (2)$$

금번 연구에서 적용한 ANN 모델은 딥러닝의 구현이 가능한 R 프로그램의 neuralnet 패키지를 사용하여 개발하였다(Fritsch et al., 2016).

2.3. CO_2 배출량 산정

본 연구에서 ML 모델의 개발에 사용한 독립변수와 종속변수는 선행연구(Chung et al., 2016)에서 다중선형회귀모델을 개발하는데 사용한 변수들과 동일하다. 독립변수는 선행연구에서 CO_2 NAF 변동성을 가장 우수하게 설명했던 수온, pH, Alk, 풍속, Chl-a 농도, 그리고 일별 pH 변화와 종속변수인 CO_2 NAF의 분산에 큰 영향을 미친 강수량(PRCP) 자료를 포함하였다. 기계학습모델의 학습과 성능평가에 사용한 종속변수인 대청호의 일별 CO_2 NAF는 Cole and Prairie (2009)이 제시한 방법을 사용하여 이론적으로 산정하였다. 이 방법은 대청호에서 일별로 측정한 실측 pH

와 Alk를 식 (3)에 입력하여 수중의 용존무기탄소(DIC)의 농도를 계산하고, 식 (4)를 이용하여 수중의 CO_2 농도를 산정한다. 그리고 실측 수온과 풍속의 함수로 기체전달 속도를 산정하는 Wanninkhof and Knox (1996) 방법으로 CO_2 기체전달 속도(k_g)를 산정한 후 식 (5)을 이용하여 대기-수표면 경계에서의 CO_2 배출량 또는 흡수량을 산정한다.

$$[DIC]=\frac{([Alk]-\frac{K_w}{[H^+]}+[H^+])\times([H^+]^2+K_1[H^+]+K_1K_2)}{K_1[H^+]+2K_1K_2} \quad (3)$$

$$[CO_2]=\frac{[DIC]\times[H^+]^2}{[H^+](K_1\times[H^+])+(K_1\times K_2)} \quad (4)$$

$$NAF=k_g\times K_H(pCO_2^{water}-pCO_2^{atm})=k_g(CO_2^{water}-CO_2^{atm}) \quad (5)$$

보다 상세한 CO_2 배출 및 흡수량 산정방법과 입력변수의 특성은 선행 연구논문(Chung et al., 2016)에 상세히 제시되어 있다.

2.4. 기계학습 모델 적용 절차

기계학습 알고리즘을 적용하기 위해서는 먼저 탐색적 데이터 분석(Exploratory Data Analysis, EDA) 단계가 필수적으로 수행되어야 한다. 이 단계에서는 데이터 수집 및 정렬, 결측값 처리 등의 전처리 작업이 이루어진다. 이러한 과정을 거친 후 해당 알고리즘의 학습이 이루어진다. 본 연구에서는 각각의 기계 학습 모델을 생성하기 위해 다음의 절차를 통해 결과를 추정하였으며, 전체 데이터의 75%를 학습 자료로 사용하고 나머지 25%를 검정 자료로 분할하여 실행하였다.

2.4.1. RF 모델

RF 모델링 절차도를 Fig. 4에 나타내었으며, 각 단계별 수행 내용은 다음과 같다.

STEP 1. 트리의 수(ntree)와 무작위추출 변수의 수(mtry) 결정

STEP 2. 오차가 가장 작은 mtry를 선정하여 학습용과 검정용 데이터로 분할

STEP 3. 학습 데이터를 무작위 복원추출하는 방식으로 실제 모집단과 가깝게 표본의 크기를 복원하는 과정을 반복함 → 다수의 부트스트랩 표본(X_i, i = 1, ..., n)을 생성 추출

STEP 4. 회귀 트리를 최대 크기로 성장할 때까지 ntree회 반복
회귀 앙상블 트리의 경우 연속형 변수인 최종 예측값을 평균하여 모델 생성

STEP 5. 생성된 RF 모델에 검정용 데이터를 사용하여 검정하며, 5회 반복실행
학습용과 검정용 데이터 각각의 RMSE의 평균값으로 오차 비교

2.4.2. ANN 모델

ANN 모델링 절차도는 Fig. 5에 나타내었으며, 각 단계별 수행 내용은 다음과 같다.

STEP 1. Scaling 기법을 사용한 데이터 표준화

STEP 2. 최적의 은닉층 노드수를 선정
여러 경우의 수를 설정 하여 예측 결과 정확도를

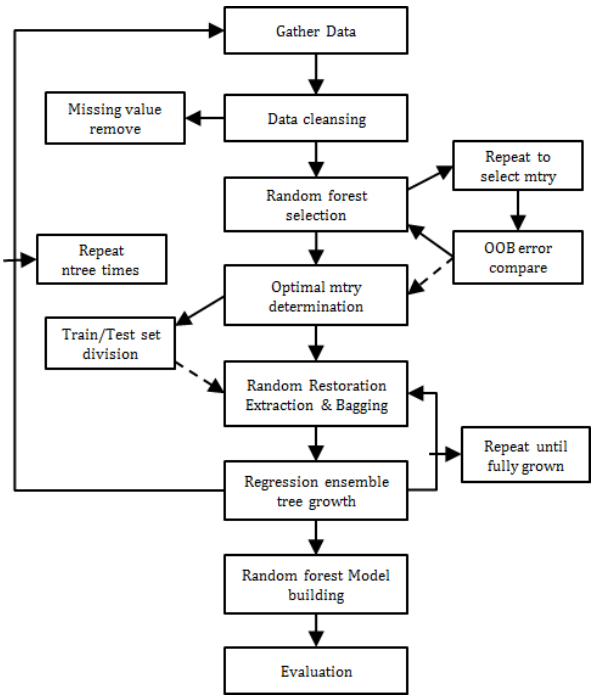

Fig. 4. Modeling procedure of the random forest model.

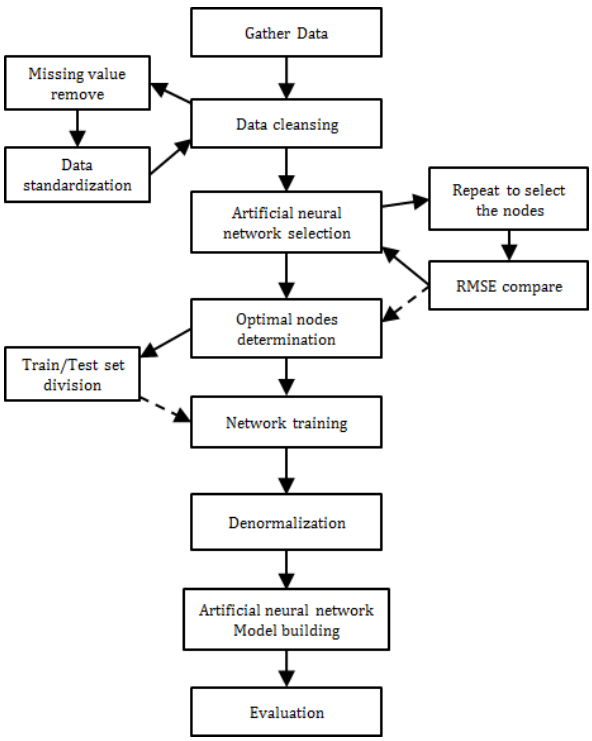

Fig. 5. Modeling procedure of the artificial neural network model.

평가하여 최적의 노드 수 선정

STEP 3. 선정된 최적 은닉층 노드수를 기준으로 학습용과 검정용 데이터 분할 후 학습데이터로 ANN 모델 생성

STEP 4. 생성된 ANN 모델에 검정용 데이터를 사용하여 검정하며, 5회 반복실행
학습용과 검정용 데이터 각각의 RMSE의 평균값으로 오차 비교

2.5. 모델 검정 및 평가

신규 구축한 기계학습 모델의 예측성능 평가를 위해 k-fold교차 검정(Cross Validation)을 실시하였다. 교차검정은 모델의 학습과정에서 발생할 수 있는 과적합 문제를 해결하는 하나의 방법이며, 학습과 훈련 데이터 선정과정에서 특정 데이터에서만 발견되는 결과를 일반화 시키는 오류가 있는지 검정하는 절차이다. 본 연구에서는 총 연구데이터의 90%를 학습데이터, 10%를 검정 데이터로 선택하여 교차검정을 수행하였으며, 검정절차는 다음과 같다.

STEP 1. 데이터의 일부를 정해진 분할로 학습 데이터(training set)와 검정 데이터(validation set)로 분리

STEP 2. 분할된 학습 데이터를 무작위로 k개의 fold로 구분

STEP 3. 학습 데이터로부터 모델을 생성 후, 검정 데이터를 적용하여 성능 평가

STEP 4. 위 과정을 반복하여 k번 실시하여 error을 기록

STEP 5. 기록된 error의 평균을 교차 검정 오차(cross-validation error)로 계산함

모델의 평가는 평균제곱근 오차인 RMSE (Root Mean Squared Error)와 실측 평균값에 대한 상대오차(%error)로 평가하였다. 계산식은 Table 2에 나타내었으며, 여기서 n은 샘플 수, A_i는 실측값, P_i는 예측값이다.

Table 1. Error statistics used to assess the models' performance

Notation	RMSE	%error
Equation	$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(A_i - P_i)^2}$	$\frac{\mathrm{RMSE}}{MEAN} \times 100$

Table 2. The RMSE and relative percent error between the multiple regression model and observed values.
Unit: mg CO_2 $m^{-2}day^{-1}$

Data set	Training data		Testing data	
	RMSE	%error	RMSE	%error
#1	1413.6	84.1%	1522.3	90.6%
#2	1432.5	85.2%	1477.0	87.9%
#3	1436.2	85.5%	1460.6	86.9%
#4	1476.0	87.8%	1342.4	79.9%
#5	1436.1	85.5%	1458.3	86.8%
Average	1438.9	85.6%	1452.1	86.4%

3. Results and Discussion

3.1. MLR 모델

MLR 모델은 전체 데이터의 75%를 학습 자료로 사용하고 나머지 25%를 검정 자료로 분할하여 총 5회 반복 실행하였다. Table 2에 MLR 분석 모델 분석 결과를 제시하였다. 학습 데이터와 검정 데이터의 평균 RMSE는 각각 1438.9 mg CO_2 $m^{-2}day^{-1}$와 1452.1 CO_2 $m^{-2}day^{-1}$이었으며, 실측 평균값에 대한 상대오차는 각각 85.6%와 86.4%이었다.

선행연구에서 도출한 대청호의 평균 NAF 산정값인 1680.5 CO_2 $m^{-2}day^{-1}$와 비교하여 학습 데이터와 검정 데이터 모두 매우 큰 오차범위를 나타냈다.

3.2. RF 모의 결과

RF 모델의 ntree 값은 Breiman and Cutler (2015)가 제시한 기본값인 500으로 설정하였으며, mtry의 값은 Liaw and Wiener (2002)의 연구결과를 참고하여 결정하였다. mtry의 개수를 2에서 6까지 모두 적용하여 모델링을 수행하였으며, 그 결과를 Table 3에 나타내었다. 모의결과 NAF 예측오차는 mtry가 증가할수록 감소하였으며, mtry가 6개인 모델에서 RMSE 값이 172.1 mg CO_2 $m^{-2}day^{-1}$, 상대오차는 10.2%로써 가장 낮은 편차를 보여, 최종 매개변수로 선정하였다.

최종 모델(mtry = 6)을 5회 반복 실행 하였으며, 그 결과를 대청호 실측 NAF대비 오차로 분석하여 Table 4에 나타내었다. 학습데이터와 검정용 데이터의 평균 오차는 각각

Table 3. The changes of RMSE according to mtry values in random forest model Unit : CO_2 $m^{-2}day^{-1}$

Data set	RMSE				
	mtry = 2	mtry = 3	mtry = 4	mtry = 5	mtry = 6
#1	247.0	198.1	178.3	171.1	171.1
#2	242.8	197.0	183.2	176.4	170.7
#3	244.2	196.2	179.3	172.0	175.4
#4	239.4	191.1	176.8	175.4	171.7
#5	246.3	198.3	180.8	169.3	171.5
Average	243.9	196.1	179.7	172.8	172.1

Table 4. The RMSE and relative error between the random forest model (mtry = 6) and observed values.
Unit : CO_2 $m^{-2}day^{-1}$

Data set	Training data		Testing data	
	RMSE	%error	RMSE	%error
#1	177.7	10.6%	426.4	25.4%
#2	180.3	10.7%	429.2	25.5%
#3	179.4	10.7%	428.4	25.5%
#4	184.3	11.0%	429.9	25.6%
#5	184.1	11.0%	431.7	25.7%
Average	181.2	10.8%	429.1	25.5%

181.2 mg CO_2 $m^{-2}day^{-1}$와 429.1 mg CO_2 $m^{-2}day^{-1}$로 나타났으며, MLR 모델에 비해 상대적으로 높은 예측 성능을 보였다.

3.3. ANN 모의 결과

본 연구의 ANN 모델은 2개의 은닉층으로 구성하였다. 첫 번째 은닉층 노드 수를 5개로 고정하고, 두 번째 은닉층의 노드 수는 Zhang et al. (1998)이 권장하는 방법에 따라 설정하여 시행 착오법을 거쳐 최적 모델을 선정하였다. 최종적으로 5-5-3-1, 5-5-6-1, 5-5-12-1, 5-5-13-1의 구조의 다층 퍼셉트론을 구성하였으며, 각 모델들의 평균 RMSE 비교 결과를 Table 5에 나타내었다. 두 번째 은닉층 노드 수가 최소인 모델(n = 3개)에서 RMSE는 109.9 mg CO_2 $m^{-2}day^{-1}$로 6.5%의 낮은 상대오차를 나타냈으며, 노드수가 최대인 모델(n = 12)에서 RMSE는 137.7 mg CO_2 $m^{-2}day^{-1}$로 나타났다. 또한 노드의 수가 많을수록 과대적합이 발생한 것으로 나타났다. 이는 ANN 모델의 성능을 향상시키기 위해 적절한 은닉층의 수를 설정해야하며, 많은 노드를 사용하는 것이 모델의 성능을 향상시키는 것이 아님을 의미한다.

최종적으로 선택된 ANN 모델(5-5-3-1)을 5회 반복 실행하였으며, 그 결과를 대청호 실측 NAF대비 오차로 분석하여 Table 6에 나타내었다. 학습 데이터의 평균 예측 오차는 136.3 mg CO_2 $m^{-2}day^{-1}$이고 검정 데이터의 평균 RMSE는 158.7 mg CO_2 $m^{-2}day^{-1}$로 각각 8.1%, 9.4%의 상대오차를 나타냈다.

Table 5. The changes of RMSEs according to the number of hidden layer nodes Unit: mg CO_2 $m^{-2}day^{-1}$

Data set	RMSE			
	c(5,3)	c(5,6)	c(5,12)	c(5,13)
#1	98.0	112.2	119.6	163.3
#2	131.0	122.7	172.1	127.1
#3	99.5	140.9	139.4	112.5
#4	118.1	108.9	133.0	144.1
#5	102.6	125.9	124.4	110.1
Average	109.9	122.1	137.7	131.4

Table 6. The results of artificial neural network according to the number of selected nodes

Unit: mg CO_2 $m^{-2}day^{-1}$

Data set	Training data		Testing data	
	RMSE	%error	RMSE	%error
#1	147.5	8.8%	157.6	9.4%
#2	168.7	10.0%	177.9	10.6%
#3	114.0	6.8%	135.3	8.1%
#4	109.9	6.5%	144.9	8.6%
#5	141.2	8.4%	177.6	10.6%
Average	136.3	8.1%	158.7	9.4%

3.4. 교차검정

MLR모델과 기계학습 모델의 개발과정에 특정 자료를 사용한 학습으로 인해 발생할 수 있는 오류를 점검하고 생성된 모델들에 대한 보다 객관적인 성능 평가를 위해 10-flod 교차검정을 수행하였고, 각 flod 마다 생성된 모델을 검정자료에 적합시켜 발생한 RMSE의 평균과 표준편차를 Table 7에 제시하였다.

10-flod 교차검정에서 얻은 MLR 모델의 평균 RMSE는 Table 2에 제시된 5회 임의추출 모델의 평균 RMSE보다 27.3 mg CO_2 $m^{-2}day^{-1}$만큼 감소, RF는 31.7 mg CO_2 $m^{-2}day^{-1}$ 증가, ANN은 6.8 mg CO_2 $m^{-2}day^{-1}$ 감소하는 것으로 나타났다. 이는 75% 임의추출로 학습한 MLR, RF, ANN 모의결과 대비 -1.88%, 7.38%, -4.28%의 변동률에 해당한다. 기계학습 모델 개발 과정에서 학습자료의 추출 방법은 어느 정도 모델 성능 평가에 영향을 미치는 것으로 보이나, 전반적으로 평가할 때 기계학습 모델인 ANN과 RF가 MLR보다 우수한 성능을 보였다.

3.5. 계절별 CO_2 NAF 산정 성능 평가

선행연구에서 MLR 모델은 CO_2 배출량이 크게 발생하는 시기에 예측값이 과소하게 평가되는 문제점이 발생하였다. 금번 연구에서 개발한 기계학습 모델의 CO_2 NAF 예측성능을 평가하기 위해 학습데이터와 검정데이터에 대한 NAF 산정 시계열 모의결과를 선행연구에서 산정한 NAF 실측값과 비교하여 Fig. 6에 나타내었으며, 각각의 모델에 대한 결정계수(R^2)를 분석하여 Table 8에 나타내었다. MLR 모

Table 7. The RMSEs obtained from 10-fold cross validation for each model Unit: mg CO_2 $m^{-2}day^{-1}$

k-fold	MLR	RF	ANN
1	1495.5	284.1	164.8
2	1271.7	405.6	171.4
3	1725.5	443.4	99.0
4	1585.3	672.3	185.9
5	1338.2	651.8	89.7
6	1242.3	184.4	158.0
7	1276.2	588.6	135.2
8	1368.6	537.3	192.5
9	1512.8	248.0	138.3
10	1352.9	272.0	149.1
Average	1424.8	460.8	151.9
Standard deviation	680.6	395.4	98.0

Table 8. Determination of coefficient (R^2) between observed and simulated NAFs

	Total	training	testing
MLR	0.728	0.735	0.699
RF	0.988	0.995	0.975
ANN	0.998	0.998	0.997

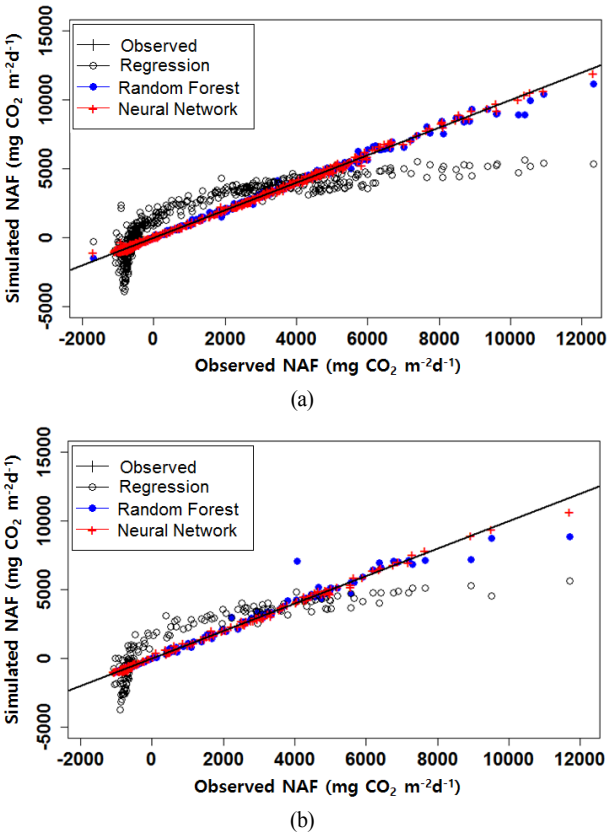

Fig. 6. Linear relationships between observed and simulated NAFs for each model: (a) training data, (b) testing data.

델은 실측값과 상당히 큰 분산을 보였으며, ANN 모델이 실측값에 가장 근접한 것으로 나타났다. 전체 데이터에 대한 RF 모델과 ANN 모델의 결정계수는 각각 0.988과 0.998로 실측값과 높은 상관관계를 보였으며, Fig. 7과 같이 2012년부터 2013년까지의 NAF 시계열변동 특성을 적절하게 재현하였다.

4. Conclusions

본 연구에서는 2012년도와 2013년도의 대청호 CO_2 배출량을 MLR 모델과 기계학습모델을 구축하여 산정하였으며, 각 모델별 예측성능을 평가하였다. 본 연구를 통해 도출된 주요 결론은 다음과 같다.

1) MLR 모델에 사용한 동일한 독립변수를 사용하여 RF와 ANN 모델을 학습시킨 결과, MLR 모델에서 나타났던 높은 CO_2 NAF 값에서 과소평가하는 문제를 해결할 수 있었다.

2) 기계학습 모델들의 매개변수 최적 선정 결과, RF 모델은 무작위추출변수(mtry)의 수가 6일 때, ANN 모델은 두 번째 은닉층의 노드 수가 3일 때 가장 높은 예측 정확도를 보였다.

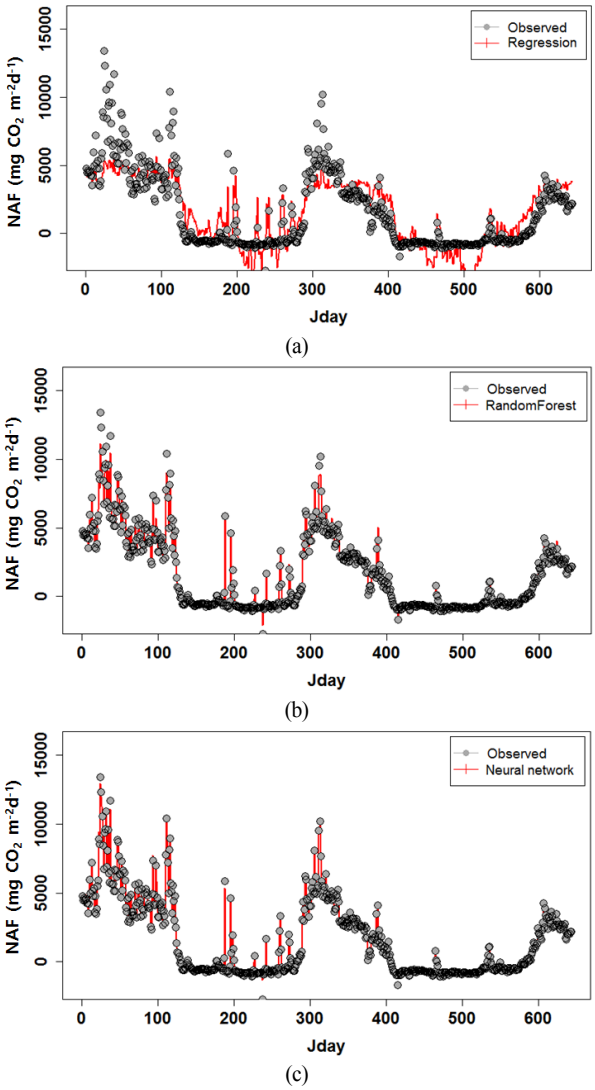

Fig. 7. Comparisons of daily CO_2 NAF time series during 2012 and 2013 with (a) multiple regression model, (b) random forest model, and (c) artificial neural network model.

3) MLR, RF, ANN 모델에 적용한 검정데이터의 평균 RMSE 값은 각각 1452.1 mg CO_2 $m^{-2}day^{-1}$, 429.1 CO_2 $m^{-2}day^{-1}$, 158.7 CO_2 $m^{-2}day^{-1}$, R^2 값은 각각 0.699, 0.975, 0.997로 산정되어 ANN 모델의 예측 성능이 가장 좋은 것으로 평가되었다.

4) 임의 추출방식으로 학습한 모델과 10-fold 교차검정으로 학습한 모델의 성능을 비교한 결과, 기계학습 모델 개발 과정에서 학습자료의 추출 방법은 어느 정도 모델 성능에 영향을 미치는 것으로 나타났으나, 본 연구에서는 그 영향이 크지 않았다.

5) 데이터 기반 기계학습 모델은 고전적인 보존법칙에 기반한 기계적 모델에 비해 상대적으로 적은 입력자료를 사용하여 저수지의 계절별 CO_2 배출량 및 흡수량 변화를 모의하는 대안 모델링 기술로 활용 가능하다고 판단된다.

Acknowledgement

이 논문은 2017년 대한민국 교육부와 한국연구재단의 개인기초연구지원사업의 지원을 받아 수행된 연구임(한국연구재단-2016-R1D1A3B03-2016131042).

References

Bae, K. T. and Kim, C. T. (2016). An Agricultural Estimate Price Model of Artificial Neural Network by Optimizing Hidden Layer, *Journal of Journal of Advanced Information Technology and Convergence*, 14(12), 161-169.

Breiman, L. and Cutler, A. (2015). Breiman and Cutler's Random Forests for Classification and Regression.

Brooks, W., Corsi, S., Fienen, M., and Carvin, R. (2016). Predicting Recreational Water Quality Advisories: A Comparison of Statistical Methods, *Environmental Modelling & Software,* 76, 81-94.

Chung, S. W., Yoo, J. S., Park, H. S., and Schladow, S. G. (2016). Estimation of CO_2 Emission from a Eutrophic Reservoir in Temperate Region, *Journal of Korean Society on Water Environment*, 32(5), 433-441. [Korean Literature]

Cole, J. J. and Prairie, Y. T. (2009). Dissolved CO_2, *Encyclopedia of Inland Water*, 33-34.

Cole, J. J., Prairie, Y. T., Caraco, N. F., McDowell, W. H., Tranvik, L. J., Striegl, R. G., Duarte, C. M., Kortelainen, P., Downing, J. A., Middelburg, J. J., and Melack, J. (2007). Plumbing the Global Carbon Cycle: Integrating Inland Waters into the Terrestrial Carbon Budget, *Ecosystems*, 10, 171-184.

Cole, T. M. and Wells, S. A. (2015). *CE-QUAL-W2: A Two-dimensional, Laterally Averaged, Hydrodynamic and Water Quality Model, Version 3.72*.

Dillon, P. J. and Rigler, F. H. (1975). A Simple Method for Predicting the Capacity of a Lake for Development Based on Lake Trophic Status, *Journal of the Fisheries Research Board of Canada,*, 32, 1519-1531.

Enting, I. G. (2002). *Inverse Problems in Atmospheric Constituent Transport,* Cambridge University Press, New York.

Fritsch, S., Cuenther, F., Suling, M., and Mueller, S. M. (2016). *Training of Neural Networks*.

Guérin, F., Abril, G., Richard, S., Burban, B., Reynouard, C., Seyler, P., and Delmas, R. (2006). Methane and Carbon Dioxide Emissions from Tropical Reservoirs: Significance of Downstream Rivers, *Geophysical Research Letters 33:L21407*, doi: 10.1029/2006GL027929.

Kemenes, A., Forsberg, B. R., and Melack, J. M. (2007). Methane Release Below a Tropical Hydroelectric Dam, *Geophysical Research Letters 34:L12809*, doi: 10.1029/2007 GL029479.

Kim, J. H., Lee, S. W., and Cha, S. M. (2016). *Environmental Statistics & Data Analysis*. [Korean Literature]

Liaw, A. and Matthew W. (2002). Classification and Regression by Random Forest. *R news 2.3*, 18-22.

Louis, V. L. S., Kelly, C. A., Duchemin, E., Rudd, J. W. M., and Rosenberg, D. M. (2000). Reservoir Surfaces as Sources of Greenhouse Gases to the Atmosphere: A Global Estimation. *BioScience*, 50(9), 766-775.

Recknagel, F., French, M., Harkonen, P., and Yabunaka, I. I. (1997). Artificial Neural Network Approach for Modelling and Prediction of Algal Blooms, *Ecological Modelling*, 96, 11-28.

Rudd, J. W .M., Harris, R., Kelly, C. A., and Hecky, R. E. (1993). Are Hydroelectric Reservoirs Significant Sources of Greenhouse Gases?, *Ambio*, 22, 246-248.

Safaie, A., Wendzel, A., Ge, Z., Nevers, M. B., Whitman, R. L., Corsi, S. R., and Phanikumar, M. S. (2016). Comparative Evaluation of Statistical and Mechanistic Models of *Escherichia coli* at Beaches in Southern Lake Michigan, *Environmental Science & Technology*, doi:10.1021/acs. est.5b05378.

Therrien, J., Tremblay, A., and Jacques, A. (2005). CO_2 Emissions from Semi-Arid Reservoirs and Natural Aquatic Ecosystems. In: Tremblay, A., Varfalvy, L., Roehm, C., Garneau, M. (eds.). Greenhouse Gas Emissions: Fluxes and Processes, Hydroelectric Reservoirs and Natural Environments. *Environmental Science Series*, Springer, New York, 233-250.

UNESCO/IHA. (2010). *GHG Measurement Guidelines for Freshwater Reservoirs*, International Hydropower association.

Wang, B., Oldham, C., and Hipsey, M. R. (2016). Comparison of Machine Learning Techniques and Variables for Groundwater Dissolved Organic Nitrogen Prediction in an Urban Area, *12th International Conference on Hydro- informatics*, *HIC 2016.*

Wanninkhof, R. and Knox, M. (1996). Chemical Enhancement of CO_2 Exchange in Natural Waters, *Limnology and Oceanography*, 41(4), 689-697

Whitehead, P. and Hornberger, G. (1984). Modelling Algal Behaviour in the River Thames, *Water Research*, 18, 945-53.

Zhang, G., Patuwo, B. E., and Hu, M. Y. (1998). Forecasting with Artificial Neural Networks: The State of the Art, *International Journal of Forecasting*, 14, 35-62.