Korean Journal Of Agricultural Science

pISSN : 2466-2402
eISSN : 2466-2410

ENGINEERING

Prediction of pollution loads in the Geum River upstream using the recurrent neural network algorithm

Heesung Lim[1], Hyunuk An[1,*], Haedo Kim[2,*], Jeaju Lee[2]

[1] Agricultural and Rural Engineering, Chungnam National University, Daejeon 34134, Korea
[2] Rural research institute, Korea Rural Community Corporation, Ansan 15634, Korea

*Corresponding authors: hyunuk@cnu.ac.kr; searoad@ekr.or.kr

Abstract

The purpose of this study was to predict the water quality using the RNN (recurrent neutral network) and LSTM (long short-term memory). These are advanced forms of machine learning algorithms that are better suited for time series learning compared to artificial neural networks; however, they have not been investigated before for water quality prediction. Three water quality indexes, the BOD (biochemical oxygen demand), COD (chemical oxygen demand), and SS (suspended solids) are predicted by the RNN and LSTM. TensorFlow, an open source library developed by Google, was used to implement the machine learning algorithm. The Okcheon observation point in the Geum River basin in the Republic of Korea was selected as the target point for the prediction of the water quality. Ten years of daily observed meteorological (daily temperature and daily wind speed) and hydrological (water level and flow discharge) data were used as the inputs, and irregularly observed water quality (BOD, COD, and SS) data were used as the learning materials. The irregularly observed water quality data were converted into daily data with the linear interpolation method. The water quality after one day was predicted by the machine learning algorithm, and it was found that a water quality prediction is possible with high accuracy compared to existing physical modeling results in the prediction of the BOD, COD, and SS, which are very non-linear. The sequence length and iteration were changed to compare the performances of the algorithms.

Keywords: LSTM (long short-term memory), machine learning, RNN (recurrent neural networks), Tensorflow, water pollution prediction

Citation: Lim H, An H, Kim H, Lee J. 2019. Prediction of pollution loads in the Geum River upstream using the recurrent neural network algorithm . Korean Journal of Agricultural Science. https://doi.org/10.7744/kjoas.20180085

DOI: https://doi.org/10.7744/kjoas.20180085

Received: October 25, 2018

Revised: November 11, 2018

Accepted: November 13, 2018

Introduction

최근 산업화와 공업의 발전에 의해 인간의 생활수준이 향상됨에 따라 환경오염은 점점 더 심각해지고 있으며, 특히 하천의 수질오염 문제는 사회·경제적으로 그 중요성이 점점 더 커지고 있다. 수질오염 문제에 있어 경제적이며 합리적인 수질관리를 위해서는 타당성 있고 현실적인 수질기준을 정하여 경제적 손실을 최소화 하는 작업과 함께 자연적 정화 능력인 하천의 자정작용을 최대한 활용하는 것이 중요하다. 이를 위해서는 모니터링 자료를 기초로 한 수질 예측을 바탕으로 적절한

대응책을 마련할 필요가 있으나 하수처리장, 하수관로, 가축분뇨 공공처리시설 등에 많은 시설 투자가 이루어지고 있는 반면, 상대적으로 모니터링 자료의 축척과 수질 예측에 대한 연구는 아직 부족한 것이 현실이다.

수질 예측을 위한 방법은 물리적 기반의 모형을 이용하는 방법과 데이터 기반의 기계학습을 이용한 방법으로 크게 구분할 수 있다. 국내에 널리 적용되고 있는 수질예측모형으로는 QUAL2E 모형(Brown and Barnwell, 1987), HSPF모형(EPA, Washington D.C., USA), WASP5모형 등이 있다. Shin and Kwun (1997)은 WASP5 모형을 사용하여 복하천의 수질 예측을 수행하여 계절에 따른 유달율 변화의 반영이 수질 예측의 정확성 향상에 도움이 되는 것을 밝혀냈다. Seo et al. (2004)은 QUALE2를 이용하여 금강 하류의 수질을 모델링하여 예측한 결과 유량의 급격한 변화를 모형에서 제대로 반영하지 못한 것을 오차에 대한 주원인으로 분석하고 수질 예측의 정확도를 높이기 위해서는 유량 예측 정확도가 확보되어야 한다고 결론 내렸다. Shin and Kim (2016)은 낙동강 주요 지류의 수질 예측을 위하여 HSPF 유역 모델을 이용하여 수질 예측 정확성을 평가하였는데 대부분의 지류에서 수질 모의 결과가 관측값을 잘 재현하였다고 하였는데, 수질 예측 향상을 위해서는 수치 모델의 개선, 기상 및 점오염원 등 모델 경계 조건의 정확도 향상 등 다각적인 노력이 필요할 것이라고 결론 내렸다. 물리적인 모형은 미계측 지역 또는 데이터가 충분히 확보되지 않은 지점에서도 적용이 가능하다는 장점이 있는 반면 데이터가 충분히 확보된 지역의 경우 데이터 기반의 방법에 비해 예측의 정확도가 떨어지는 경향이 있다.

기계학습을 이용한 국내 수질 예측 방법으로는 주로 인공신경망을 이용한 연구가 수행되었다. Jeong et al. (2001)은 10년간의 월별 수질 자료를 이용하여 인공신경망을 구축 하고 월별 pH, DO (dissolved oxygen), BOD (biochemical oxygen demand) 항목을 예측하여 양호한 결과를 얻었다.

Seo and Yun (2016)은 인공신경망모델을 이용하여 8개의 수질인자 실측값과 예보값을 비교해본 결과 평균적으로 10% 미만의 에러값을 얻었으며, 인공신경망이 수질데이터가 가지는 불확실성 및 복잡한 상관성에 효과적으로 대응할 수 있는 데이터기반 모델인 것으로 결론지었다.

Lee et al. (2001)은 수질 농도를 예측하기 위해 신경망의 역전파 학습 알고리즘을 적용하여 나주와 함평의 BOD, COD (chemical oxygen demand), T-N, T-P 수질 농도를 예측 하였으며 예측한 결과 예측치가 실측치를 잘 반영하고 있음을 알 수 있으나 자료의 수가 너무 적고 월 자료의 학습으로 인해 예측의 정확도가 떨어지는 것을 지적하였다. Oh et al. (2002)은 입력자료로는 수질 항목 DO, BOD, T-N, T-P 농도와 하천의 유출량 및 수온의 월 평균 자료를 이용하여 신경망 모형을 이용하여 하천수의 수질 예측을 하였는데 예측한 결과 월 자료만의 활용으로 자료의 빈약성에도 불구하고 우수한 모형을 개발했다고 결론을 내렸고 더 우수한 결과를 얻기 위해서는 일자료 또는 시자료의 구축이 필요할 것이라고 판단하였다. Park et al. (2000)은 DO, BOD 농도 예측을 위해 인공신경망 이론을 적용하였는데 입력 자료로 월 자료를 이용하였음에도 불구하고 우수한 모형을 개발하였는데, 일 자료 또는 5일 간격의 자료 구축이 선행된다면 더 우수한 모형을 개발할 수 있었을 것이라 결론 내렸다. Kim and Han (2002)은 단기 하천 수질 예측을 위해 신경망이론을 이용하였는데 연구 결과는 비교적 양호한 결과는 내놓았는데 연속된 수질 자료가 있었다면 예측능력을 충분히 검증할 수 있었을 것이라고 결론 내렸다. 수질 예측에 있어 자료의 부족이 많은 문제점을 가지고 있었는데, 기계학습을 이용한 많은 논문들을 살펴보면 빈약한 자료를 이용하였어도 우수한 결과를 나타내고 있었다. 그러나 더 우수한 결과를 얻기 위해는 시 자료 또는 일 자료의 구축이 선행되어야 된다고 판단하여 연구를 위해 일 자료 또는 시 자료의 구축이 필요할 것으로 판단하였다.

본 연구에서는 인공신경망의 발전된 형태인 순환신경망 알고리즘을 활용하여 일단위 수질 예측을 하고자 하였다. 일단위의 예측을 위해서는 일 단위 수질 측정 자료가 필요하나 수질 항목에 대한 일 단위의 측정은 비용 및 제반 여건 상 거의 불가능한 것이 현실이다. 수질 측정은 각 지역 보건환경연구원에서 월 1회 이상의 조사를 하고 있으며 일반적으로 월 2 - 4회 정도 비정기적으로 측정되고 있다. 본 연구에서는 이러한 자료를 선형보간 하여 일 자료를 구축한 뒤, 이를 순환신경망 알고리즘을 통해 학습하여 일 단위 수질 예측을 하고자 하였다. 본 연구에서는 Google에 개발한 오픈소스 라이브러리인 텐서플로우를 활용하여 연구를 수행하였다.

Materials and methods

기계학습 (Machine learning)

기계학습은 인공지능의 한 분야로 컴퓨터가 모델 생성을 자동화 하여 데이터를 바탕으로 학습하고 패턴을 찾아내는 기술이다. 기계학습방법 중 가장 널리 사용되는 알고리즘 중 하나인 인공신경망(artificial neural network, ANN) 알고리즘은 인간의 뇌가 패턴을 인식하는 방식을 모사한 알고리즘으로 금융, 경제 등 사회과학 및 과학 전반에 걸쳐 폭넓게 사용되고 있다(Roh et al, 2005). 순환신경망(recurrent neural network, RNN)은 인공 신경망의 한 종류로, 유닛간의 연결이 순환적 구조를 갖는 특징을 갖고 있으며 대상 시스템의 시변적 동적 특징을 잘 예측할 수 있는 심층학습(deep learning) 알고리즘의 일종이다. 순환신경망은 강우예측(Jeong et al., 2016) 및 수위예측(Tran and Song, 2017)에 성공적으로 적용된 바 있다.

순환신경망 (RNN)

순환신경망은 sequence data 순서대로 데이터를 처리하는 모델로, 과거의 데이터가 미래에 영향을 주는 루프 반복형 구조로 되어있다. Fig. 1와 같이 데이터 시퀀스 상의 특정 시점 t에서 X_t는 입력층(input layer), h_t는 은닉층(hidden layer), A_t는 입력층과 은닉층 간의 가중치(weight) 라고 하면 순환신경망에서는 신경망을 연결하는 과정에서 지난 시점(t-1, t-2,...)의 은닉층을 현 시점의 은닉층에 누적시켜 계산하게 된다. 순환신경망은 이러한 특징으로 인해 이전의 계산 결과가 다음 계산에 영향을 미치기 때문에 데이터의 순서가 정해져 있는 sequence date, 특히 시계열 데이터를 처리하는데 적합하다.

순환신경망의 가장 기본적인 형태인 vanilla RNN은 다음과 같은 기초식을 사용한다.

$$h_t = f_w(h_{t-1}, x_t) \tag{1}$$

여기서, f_w는 활성함수이다. 위 수식을 보면 현 시점의 은닉층 벡터 h_t는 이전 시점의 은닉층 벡터 h_{t-1}와 매 타임마다 적용되는 입력벡터 x_t의 활성함수로 표현된다는 것을 알 수 있다. 일반적으로 vanilla RNN은 활성함수(activation function)으로 tanh함수를 사용하는데 이를 식(1)에 적용하면 다음과 같다.

$$h_t = \tanh(W_{hh}h_{t-1} + W_{xh}x_t) \tag{2}$$

여기서, W_{hh}, W_{xh}는 가중치행렬이다. 식(2)를 통하여 산정된 은닉층 벡터의 값은 가중치를 통해 다음식과 같이 y_t출력층으로 변환된다.

$$y_t = W_{hy}h_t \tag{3}$$

이하 본 논문에서 사용하는 RNN은 vanilla RNN을 뜻한다.

LSTM (long short-term memory)

순환신경망모형의 장점은 이전의 정보를 현재의 문제 해결에 활용할 수 있다는 점으로 시계열 데이터처리에 특화된 알고리즘으로 알려져 있다. 그러나 루프 반복형 구조로 인해 sequence data의 순서가 멀어지면 가중치가 작아져 소멸해버리는 가중치 소실(vanishing gradient)문제가 발생하게 된다. 이러한 vanilla RNN의 단점을 극복하기 위해 Hochreiter and Schmidhuber (1997)은 장기 의존성을 학습할 수 있도록 순환신경망 모형을 수정한 LSTM메모리 네트워크 기법을 제안하였다.

LSTM에서는 Fig. 2와 같이 가중치 반영 및 활성함수 변환을 통하여 입력값에서 출력값으로 변환하는 단계를 하나의 셀

(cell)로 보고 셀 내부의 상태량인 셀 스테이트(cell state)를 입력, 망각, 출력게이트를 이용하여 총 4단계의 계산과정을 통해 가중치 소실문제가 발생하지 않도록 조절한다. 첫 번째 단계에서는 특정 정보의 제거여부를 망각게이트 (forget gate, f_t)를 통하여 다음 식과 같이 결정한다.

$$f_t = \sigma(W_f[h_{t-1}, x_t] + b_f) \quad (4)$$

여기서, f_t는 0 또는 1의 값을 가지는 망각게이트 값, W_f는 망각게이트 가중치, b_f는 망각게이트 편향값, 는 Sigmoid 활성화 함수이다. 두 번째 단계에서는 입력게이트(input gate, i_t)를 통하여 새로운 정보의 저장여부를 다음 식과 같이 결정한다.

$$i_t = \sigma(w_i[h_{t-1}, x_t] + b_i) \quad (5)$$

여기서, i_t는 0 또는 1의 값을 가지는 입력게이트 값, W_c는 입력게이트 가중치, b_c는 입력게이트 편향값이다. 세 번째 단계에서는 입력게이트와 출력게이트의 값을 이용하여 셀 스테이트를 다음 식과 같이 업데이트 한다.

$$C_t = \tanh\,(W_C[h_{t-1}, x_t] + b_c) \quad (6)$$

$$C_t = f_t \cdot C_{t-1} + i_t \cdot \bar{C_t} \quad (7)$$

여기서, C_t는 tanh로 구성되어있는 셀 스테이트 중간값이며, C_t는 업데이트 된 시점 t에서의 셀 스테이트를 나타낸다. 마지막 단계는 출력게이트와 셀 스테이트를 이용하여 출력값을 다음 식과 같이 계산한다.

$$o_t = \sigma(W_0[h_{t-1}, x_t] + b_0) \quad (8)$$

$$h_t = o_t \cdot \tanh\,(C_t) \quad (9)$$

여기서, o_t는 0 또는 1의 값을 가지는 출력게이트 값, W_o는 출력게이트 가중치, b_0는 출력게이트 편향값이다.

대상지점 및 자료

본 연구의 대상 지점은 Fig. 3과 같이 충청남도 옥천군 이원면 36'14'31.0'N, 127'40'02.6'E 이원대교에 위치한 옥천관측소이다. BOD, COD, SS (suspended solids) 자료는 물환경정보시스템의 자료를 활용하였고, 수위, 유량등의 자료는 국가수자

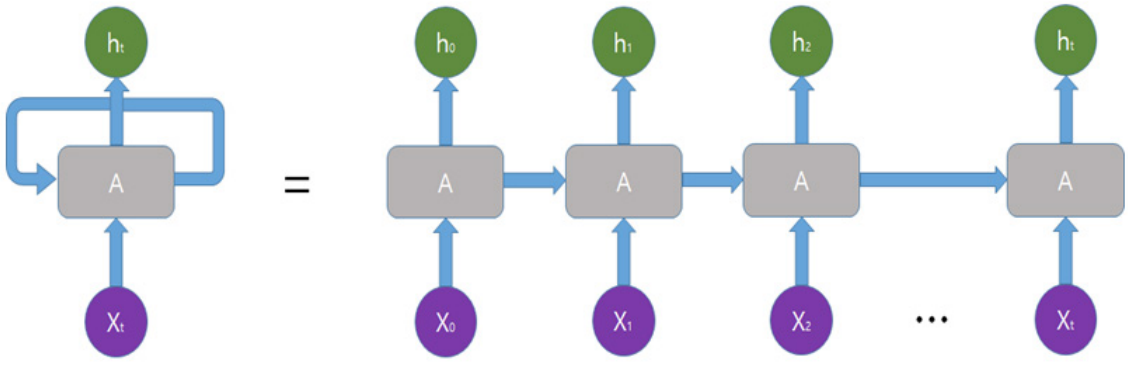

Fig. 1. Structure of recurrent neural networks. A, weight; h_t, hidden layer; h_0, hidden layer 0; h_1, hidden layer 1; h_2, hidden layer 2; X_t, input layer; X_0, input layer 0; X_1, input layer 1; X_2, input layer 2.

원관리종합정보시스템에서 자료를 수집하였다.

물환경정보시스템에서 취득한 10년동안 관측된 수질(BOD, COD, 부유물질)측정 자료는 총 411건으로 측정시점이 일정하지 않았으며 약 9일마다 한번꼴로 측정되었다. 본 연구의 목적은 일별 수질예측으로 학습을 위해서 원자료를 선형보간하여 일별 기상 및 수질 자료를 산정하여 입력자료로 사용하였다. 2008 - 2014년 7년간의 자료를 학습 자료로 사용하였고 2015 - 2017년 3년간의 자료는 모형의 검증을 위해 사용하였다.

RNN, LSTM 모형의 입력자료로는 평균기온, 평균풍속, 하천의 수위, 유량의 자료를 사용하였으며, 은닉층 개수는 수질오염의 복잡성을 고려하여 10개층으로 정하였다. SL (sequence length)는 총 3가지 경우(SL = 3, 5, 7)로 설정하였으며, 반복시행(iteration)수는 3,000, 5,000, 10,000번으로 3가지 경우를 설정하여 조합에 따른 결과를 분석하였다. 활성화 함수로는 RNN과 LSTM에 가장 널리 쓰이는 hyperbolic tangent (tanh) 함수를 사용하였으며 역전파(backpropagation) 알고리즘을 이용하여 신경망을 학습하였다.

예측결과가 실측자료와 비교하여 얼마나 정확한 것인가 하는 문제는 단기 수질모형에 대한 신경이론의 적용타당성과

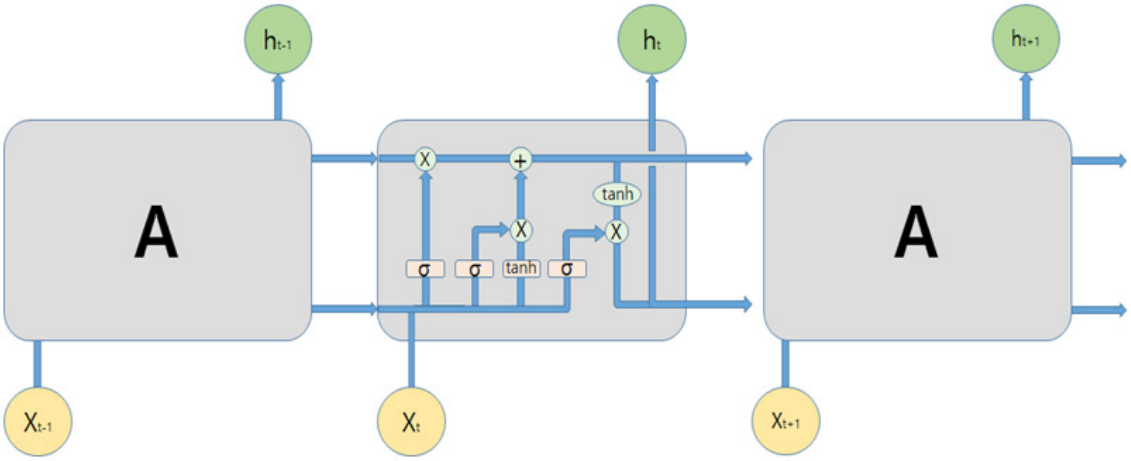

Fig. 2. Structure of LSTM (long short-term memory) network. A, weight; h_{t-1}, hidden layer (t-1); h_t, hidden layer; h_{t+1}, hidden layer (t+1); X_{t-1}, input layer (t-1); X_t, input layer; X_{t+1}, input layer (t+1); σ, sigmoid layer; tanh, hyperbolic tangent; X, multiplication; +, plus.

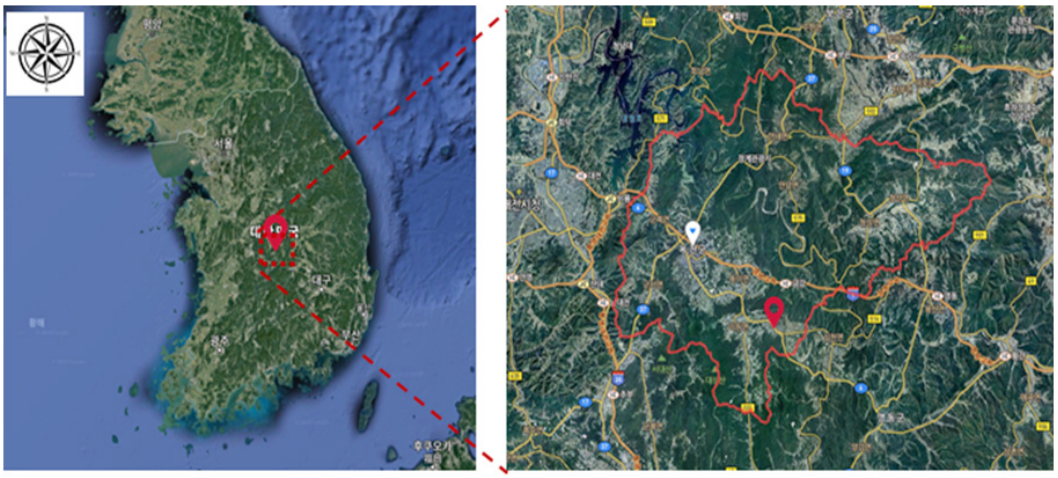

Fig. 3. Study area (Okcheon-gun, Chungcheongbuk-do, Korea).

관련된다. 본 연구에서는 예측문제에 적용되고 있는 통계적 검증방법으로 각 모형에서 계산된 오차와 비교 검토하여 예측에 사용될 최적 모형을 선정하기 위해 R^2와 RMSE (root mean square error)을 이용하여 예측결과를 분석하였으며 이들 인덱스의 식은 Table 1에 제시하였다.

Results and Discussion

RNN 알고리즘 결과 분석

RNN 알고리즘으로 BOD 농도값을 예측한 결과값을 Table 2에 제시하였다. R^2은 0.960 - 0.417의 범위를 나타내었으며, RMSE 0.0000026 - 0.0005666의 범위를 나타내었다. Case 중에서 BR5, BR4, BR2의 R^2가 각각 0.960, 0.942, 0.915로 높은 정확도를 나타내었다. SL은 3, 5, 7 순으로 높은 정확도를 보였으며, iteration 수는 3,000, 5,000번 순으로 높은 정확도를 보이는 것으로 나타났다. BR3, BR6, BR9의 case에서 알 수 있듯이 iteration 수가 과도한 경우(10,000번), 과적합(overfitting)이 발생하여 오히려 예측 정확도가 떨어지는 현상이 발생하였다.

COD 농도값을 예측한 결과값을 분석한 결과를 Table 3에 제시하였다. R^2은 0.982 - 0.944의 범위를 나타내었으며 RMSE 0.0001680 - 0.0016531의 범위를 나타내어 BOD에 비해 비교적 정확한 예측이 가능하였다. Case들 중 CR2, CR1, CR7의 R^2가 각각 0.982, 0.981, 0.970으로 높은 정확도를 나타내었으며, sequence length는 BOD의 경우와 마찬가지로 3이, iteration 수는 3,000번이 가장 좋은 결과를 내는 것으로 나타났다. BOD의 경우와 마찬가지로 과도한 iteration 수는 과적합을 유발하여 오히려 예측정확도를 저하하는 것으로 나타났다. 그러나 COD의 경우는 저하된 R^2의 값들도 0.94 이상으로 매우 높은 것을 확인할 수 있다.

SS를 예측한 결과값을 Table 4에 나타내었다. R^2은 0.985 - 0.516의 범위를 나타내었으며 RMSE 0.0185378 - 19.0433669의 범위를 나타내었다. Case들 중 SR5, SR1, SR7의 R^2가 각각 0.985, 0.976, 0.973으로 높은 정확도를 나타내었으며, SL이 3일 때, iteration 수는 3,000일 때 비교적 정확한 예측결과를 얻을 수 있는 것으로 나타났다. SR3의 경우 극단적인 과적합으로 매우 낮은 정확도(R^2 = 0.516)를 나타내는 경우도 있으며, SL이 짧을 경우 과도한 반복학습 회수는 매우 낮은 정확도를 유발할 수 있는 것으로 판단된다.

Table 2 - 4를 종합적으로 분석해 보면 일단위 수질오염 지표 예측에 있어서 SL에 따른 결과의 차이는 크지 않았으며, 이에 비해 iteration 수에 의한 결과의 차이는 비교적 크다는 것을 알 수 있다. 본 연구의 경우 과도한 3,000회 이상의 iteration 수의 경우 과적합 현상이 발생하여 결과의 정확도가 떨어지는 것을 확인할 수 있었다. 예측지표를 비교해 보면 COD, SS, BOD 순으로 높은 정확도로 예측이 가능하였으며, 적절한 SL와 iteration 수를 적용하였을 경우 전반적으로 세 지표 모두 R^2 0.95 이상의 높은 정확도로 예측이 가능하였다. Fig. 4에서 보여주는 바와 같이 RNN 모형 대부분은 농도값 예측은 우수하고 안정적으로 예측하는 것으로 분석되었다.

Table 1. Meaning of R^2 and RMSE (root mean square error) index.

Index	Relation equation
R^2	$R^2 = 1 - \frac{\sum(Error)^2}{\sum(Deviation)^2}$
RMSE	$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(t_i - p_i)}$

R^2, R square; t_i, true value; p_i, prediction value; $\overline{t_i}$, average value; Error, ti - pi; Deviation, ti - $\overline{t_i}$.

LSTM 알고리즘 결과 분석

LSTM 알고리즘도 RNN 알고리즘에 대하여 수행한 분석과 마찬가지로 각각 3가지 경우의 LS과 iteration 수를 조합하여 이에 따른 결과를 분석하였다. LSTM 알고리즘으로 BOD 농도값을 예측한 결과값을 분석한 결과를 Table 5에 제시하였다. R^2은 0.984 - 0.762의 범위를 나타내었으며, RMSE은 0.0000004 - 0.0000940의 범위를 나타내었다. Table 2와 비교해 보면 LSTM의 결과가 RNN에 비하여 전반적으로 좋은 것을 확인할 수 있으며, 가장 높은 정확도(R^2 = 0.984)의 경우도 RNN의 경우(R^2 = 0.960)에 비해 높은 것을 알 수 있다. LSTM의 경우 RNN의 결과와 달리 iteration 수가 커질수록 과적합 현상이 발생하는 것이 아니라 정확도가 높아지는 경향을 보이며 SL이 5일 때 반복학습 회수가 10,000의 경우에만 과적합 현상이 관찰된다.

Table 2. Model performance results of the BOD (biochemical oxygen demand) RNN (recurrent neural network) model.

Case	Sequence length	Iterations	R^2	RMSE
BR1	3	3,000	0.831	0.0000474
BR2		5,000	0.915	0.0000121
BR3		10,000	0.417	0.0005666
BR4	5	3,000	0.942	0.0000055
BR5		5,000	0.960	0.0000026
BR6		10,000	0.905	0.0000150
BR7	7	3,000	0.808	0.0000615
BR8		5,000	0.827	0.0000500
BR9		10,000	0.780	0.0000809

R^2, R square; RMSE, root mean square error.

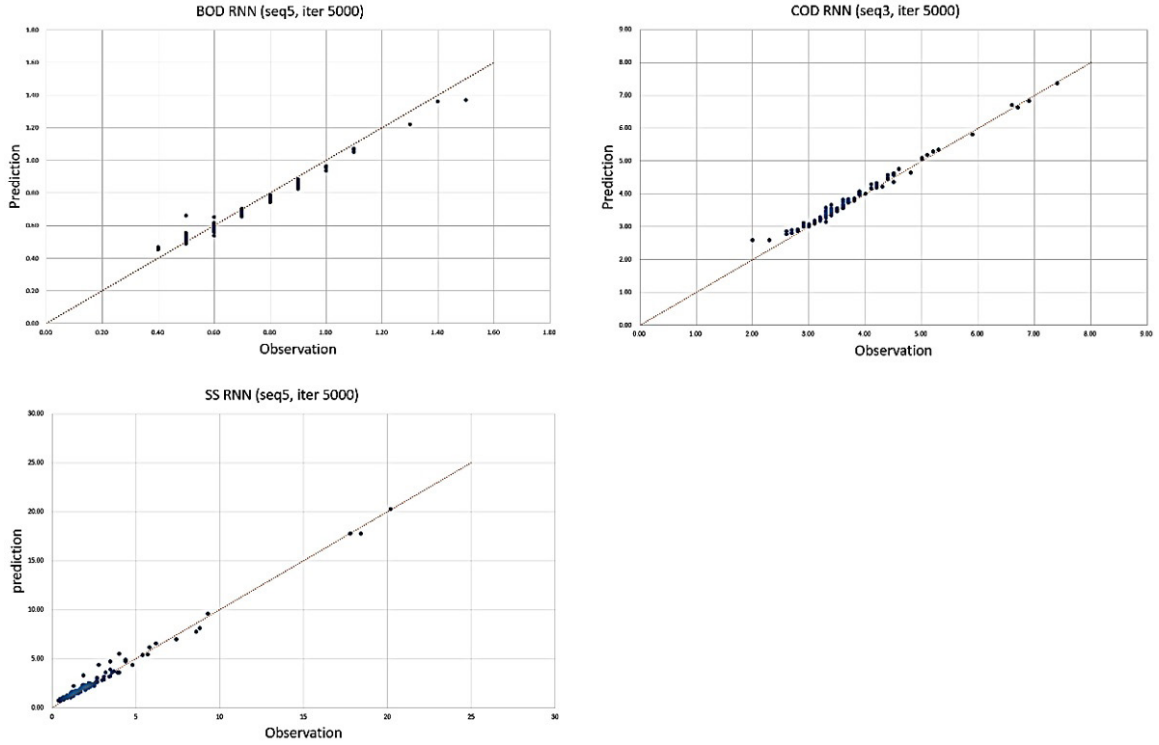

Fig. 4. RNN (recurrent neural network) scatter plot of water quality. BOD, biochemical oxygen demand; COD, chemical oxygen demand; SS, suspended solids.

LSTM 알고리즘으로 COD 농도값을 예측한 결과을 Table 6에 제시하였다. R^2은 0.994 - 0.978의 범위를 나타내었으며, RMSE 0.0000182 - 0.0002481의 범위를 나타내었다. RNN의 경우와 마찬가지로 COD의 예측 정확도는 BOD보다 높았으며, LSTM의 예측정확도는 RNN과 비교하여 높은 것으로 나타났다. BOD의 경우와 마찬가지로 LSTM의 경우 과적합 현상이 발생하지 않았으며 많은 반복학습 회수의 경우 비교적 높은 정확도의 결과를 나타내었다.

LSTM 알고리즘으로 부유물질 농도값을 예측한 결과값을 Table 7에 제시하였다. R^2은 0.991 - 0.896의 범위를 나타내었

Table 3. Model performance results of the COD (chemical oxygen demand) RNN (recurrent neural network) model.

Case	Sequence length	Iterations	R^2	RMSE
CR1		3,000	0.981	0.0001900
CR2	3	5,000	0.982	0.0001680
CR3		10,000	0.944	0.0016531
CR4		3,000	0.967	0.0005675
CR5	5	5,000	0.956	0.0010397
CR6		10,000	0.946	0.0015353
CR7		3,000	0.970	0.0004717
CR8	7	5,000	0.958	0.0009313
CR9		10,000	0.955	0.0010522

R^2, R square; RMSE, root mean square error.

Table 4. Model performance results of the SS (suspended solids) RNN (recurrent neural network) model.

Case	Sequence length	Iterations	R^2	RMSE
SR1		3,000	0.976	0.0462721
SR2	3	5,000	0.954	0.1715917
SR3		10,000	0.516	19.0433669
SR4		3,000	0.940	0.2954386
SR5	5	5,000	0.985	0.0185378
SR6		10,000	0.923	0.4764045
SR7		3,000	0.973	0.0573869
SR8	7	5,000	0.823	2.5308037
SR9		10,000	0.959	0.1376899

R^2, R square; RMSE, root mean square error.

Table 5. Model performance results of the BOD (biochemical oxygen demand) LSTM (long short-term memory) model.

Case	Sequence length	Iterations	R^2	RMSE
BL1		3,000	0.867	0.0000295
BL2	3	5,000	0.975	0.0000010
BL3		10,000	0.984	0.0000004
BL4		3,000	0.873	0.0000270
BL5	5	5,000	0.968	0.0000017
BL6		10,000	0.902	0.0000160
BL7		3,000	0.762	0.0000940
BL8	7	5,000	0.871	0.0000278
BL9		10,000	0.931	0.0000080

R^2, R square; RMSE, root mean square error.

으며, RMSE 0.0071847 - 0.8769174의 범위를 나타내었다. BOD 및 COD의 경우와 마찬가지로 LSTM알고리즘을 적용한 경우 과적합 현상이 발생하지 않았으며 iteration 수가 높을수록 높은 정확도의 예측결과를 얻을 수 있었다.

전반적으로 LSTM을 적용한 경우에 RNN을 적용한 경우보다 안정적으로 정확한 결과를 얻을 수 있었으며, iteration 수가 높을수록 예측이 정확해 지는 경향을 확인할 수 있었다. RNN의 경우와 마찬가지로 SL의 향상에 따른 뚜렷한 경향성은 확인할 수 없었으며, 계산시간을 고려해 보면 3 - 5 정도의 SL의 설정이 적절한 것으로 판단된다. 적절한 SL과 iteration 수를 적용하였을 경우 BOD, COD, SS 지표에 대하여 각각 0.984, 0.994, 0.991이라는 매우 높은 R^2값을 얻을 수 있었다. Fig. 5에서 보여주는 바와 같이 LSTM 모형 대부분은 농도값 예측은 우수하고 안정적으로 높은 정확도를 나타내는 것으로 분석되었다.

Conclusion

본 연구에서는 기계학습 딥러닝 알고리즘의 일종인 RNN 및 LSTM 알고리즘을 이용하여 일단위 BOD, COD, SS를 예측하였다. RNN 및 LSTM 알고리즘을 구현하기 위하여 텐서플로우를 활용하였으며, 신경망의 학습에는 역전파 알고리즘을 이용하였다. 대청댐 상류부의 금강유역 옥천관측소 지점의 자료를 활용하였으며, 2008년부터 2017년까지의 10년간의 일단위 기상(수위, 유량, 평균풍속, 평균온도) 데이터와 평균 10일 간격으로 관측된 수질(BOD, COD, SS)자료를 일단위로 선형보간하여 사용하였다. 모형의 학습에는 2008 - 2014년(7개년)의 자료를 활용하였으며 2015 - 2017년(3개년)에 대하여

Table 6. Model performance results of the COD (chemical oxygen demand) LSTM (long short-term memory) model.

Case	Sequence length	Iterations	R^2	RMSE
CL1		3,000	0.985	0.0001256
CL2	3	5,000	0.992	0.0000354
CL3		10,000	0.994	0.0000182
CL4		3,000	0.992	0.0000358
CL5	5	5,000	0.978	0.0002481
CL6		10,000	0.980	0.0002098
CL7		3,000	0.986	0.0000976
CL8	7	5,000	0.992	0.0000362
CL9		10,000	0.994	0.0000186

R^2, R square; RMSE, root mean square error.

Table 7. Model performance results of the SS (suspended solids) LSTM (long short-term memory) model.

Case	sequence length	Iterations	R^2	RMSE
SL1		3,000	0.896	0.8769174
SL2	3	5,000	0.941	0.2856273
SL3		10,000	0.968	0.0812743
SL4		3,000	0.950	0.2027885
SL5	5	5,000	0.954	0.1683504
SL6		10,000	0.991	0.0071847
SL7		3,000	0.989	0.0099223
SL8	7	5,000	0.972	0.0632135
SL9		10,000	0.988	0.0122544

R^2, R square; RMSE, root mean square error.

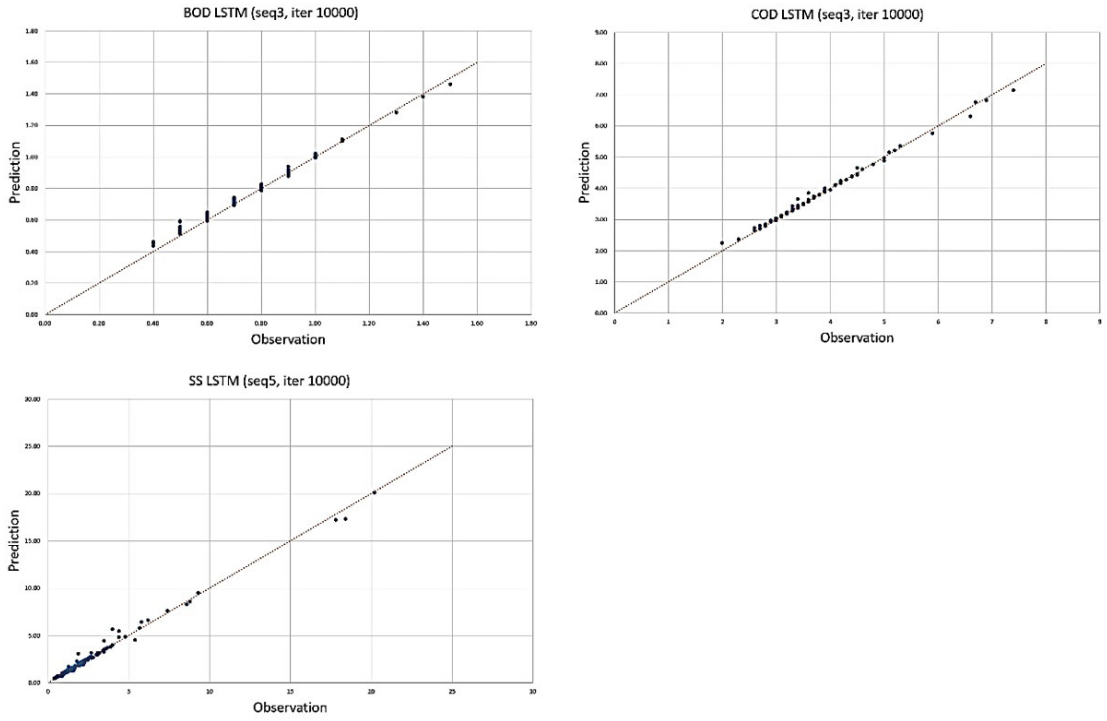

Fig. 5. LSTM (long short-term memory) scatter plot of water quality. BOD, biochemical oxygen demand; COD, chemical oxygen demand; SS, suspended solids.

모형을 검증하여 분석하였다. 본 논문의 결과를 정리하면 다음과 같다.

RNN 알고리즘을 활용한 수질예측 결과 SL의 길이는 3 이상일 경우 결과의 정확도에 미치는 영향이 크지 않은 것으로 나타났으며, iteration 수의 경우 10,000번 이상이 되면 대체로 과적합 현상이 나타나는 것이 판명되었다. 적절한 SL과 iteration 수가 적용되었을 경우 RNN 알고리즘은 세 항목(BOD, COD, SS)의 수질지표에 대하여 모두 R^2가 0.95 이상의 높은 정확도로 예측이 가능하였다.

LSTM 알고리즘을 활용한 수질예측 결과, RNN에 비하여 보다 정확한 수질예측 결과를 얻을 수 있었으며, 과적합 현상도 거의 발생하지 않았다. RNN의 경우와 마찬가지로 SL의 길이는 3 이상일 경우 결과의 정확도에 미치는 영향이 크지 않은 것으로 나타났다. 적절한 SL과 iteration 수를 적용하였을 경우 LSTM 알고리즘을 활용하여 BOD, COD, SS 지표에 대하여 각각 0.984, 0.994, 0.991이라는 매우 높은 정확도(R^2)의 예측을 할 수 있었다.

본 연구는 기계학습으로 기상자료를 활용하여 하루 뒤의 수질예측을 하고자 하였으며, 기존의 물리적 모델링 결과들과 비교하여 높은 정확도로 수질예측이 가능한 것을 확인하였다. 향후 연구의 보다 실용적인 활용을 위해서 3, 5, 10일 수질예측에 대한 검증을 수행할 예정이다.

Acknowledgements

본 연구는 농림축산식품부 농촌기반기술연구사업(농업생산기반시설 성능개선 및 자율학습 물관리 기술개발)의 지원을 받아 수행된 연구입니다.

Authors Information

Hyunuk An, https://orcid.org/0000-0002-4566-5159

Heesung Lim, Chungnam National University, Master

Haedo Kim, Korea rural community Corporation, Researcher

Jeaju Lee, Korea rural community Corporation, Researcher

References

Brown LC, Barnwell TO. 1987. The enhanced stream water quality models QUAL2E and QUAL2E-UNCAS. US Environmental Protection Agency, Georgia, USA.

Hochreiter S, Schmidhuber J. 1997. Long short-term memory. Neural computation 9:1735-1780.

Jeong HJ, Lee SJ, Lee HK. 2002. Water quality forecasting of Chungju Lake using artificial neural network algorithm. Journal of the Environmental Sciences 11:201-207. [in Korean]

Jeong HW, Ki SJ, Jeon DJ, Kim JH. 2016. Development of system based on weather radar images for predicting rainfall events using machine learning models in watershed of Yeong-san River. pp. 283-284. Proceedings of 2016 Joint Conference of Korean Society on Water Environment and Korean Society of Water & Wastewater, Korea. [in Korean]

Kim MS, Han JS. 2002. Artificial neural networks for forecasting of short-term river water quality. Journal of the Korean Geo-Environmental Society 3:11-17. [in Korean]

Lee GH, Kim IH, Moon BS. 2001. Water quality prediction of river using intelligent model. pp. 179-182. Proceedings of 2001 Joint Conference of Korean Society of Water & Wastewater and Korean Society on Water Environment, Korea. [in Korean]

Oh CR, Park SC, Lee HM, Pyo YP. 2002. A forecasting of water quality in the Youngsan River using neural network. Jouranl of The Korean Society of Civil Engineers 22:371-382. [in Korean]

Park SC, Lee HM, Oh CR. 2000. The application of artificial neural network for forecasting of DO, BOD concentration. Journal of Environmental Research 5:31-48. [in Korean]

Roh TH, Lee TH, Han IG. 2005. Forecasting the volatility of KOSPI 200 using neural network-financial time series model. Korean Management 34:683-713. [in Korean]

Seo IW, Yun SH. 2016. Forecasting water quality by ANN model at the downstream of Cheongpyeong Dam. pp. 41-42. Korean Society of Civil Engineers (KSCE) 2017 Convention, Seoul, Korea. [in Korean]

Shin DS, Kwun SK. 1997. Water quality modeling for Bokha Stream by WASP5 model. Korean Journal of Environmental Agriculture 16:233-238. [in Korean]

Seo DI, Lee JH, Lee EH, Ko IH. 2004. Analysis on errors of water quality modeling for the Geum River down stream areas using QUAL2E. Korean Society of Environmental Engineers 26:933-940. [in Korean]

Shin CM, Kim KH. 2016. Operational water quality forecast for the Nakdong River basin using HSPF watershed model. Journal of Korean Society on Water Environment 32:570-581. [in Korean]

Tran Q, Song S. 2017. Water level forecasting based on deep learning: A use case of Trinity River-Taxas-The United States. Journal of Korean Institute of Information Scientists Engineers 44:607-612. [in Korean]