한국해안 · 해양공학회논문집/ISSN 1976-8192(Print), ISSN 2288-2227(Online)
Journal of Korean Society of Coastal and Ocean Engineers 34(3), pp. 47~57, June 2022
https://doi.org/10.9765/KSCOE.2022.34.3.47

기계학습 모형 기반 진해만 용존산소농도 및 빈산소수괴 발생 예측
Prediction in Dissolved Oxygen Concentration and Occurrence of Hypoxia Water Mass in Jinhae Bay Based on Machine Learning Model

박성식* · 김병국** · 김경회***
Seongsik Park*, Byeong Kuk Kim** and Kyunghoi Kim***

요 지 : 본 연구에서는 진해만의 단일 정점 장기 모니터링 자료를 사용하여 LSTM 모형을 이용한 DO 농도 예측 및 결정 트리 모형을 이용한 빈산소수괴 발생 예측 연구를 수행하였다. LSTM을 이용한 DO 농도 예측 결과, Hidden node의 수가 증가할수록 모형의 복잡도가 증가하여 많은 Epoch을 요구하는 모습을 보였으며, 예측 시간 간격이 증가할수록 긴 Sequence length에서 높은 정확도를 보였다. 결정 트리를 이용한 빈산소수괴 발생 예측 결과, 30 day 예측에서 빈산소수괴 미발생 예측 정확도는 66.1%로 발생 예측 정확도의 37.5%보다 상대적으로 높게 나타났다. 이는 결정 트리 모형이 DO 농도를 관측값보다 고평가하여 나타난 결과로 판단된다.

핵심용어 : 기계학습, LSTM, 결정 트리, 용존산소, 빈산소수괴, 진해만

Abstract : We carried out studies on prediction in concentration of dissolved oxygen (DO) with LSTM model and prediction in occurrence of hypoxia water mass (HWM) with decision tree. As results of study on prediction in DO concentration, a large number of Hidden node caused high complexity of model and required enough Epoch. And it was high accuracy in long Sequence length as prediction time step increased. The results of prediction in occurrence of HWM showed that the accuracy of nonHWM case was 66.1% in 30 day prediction, it was higher than 37.5% of HWM case. The reason is that the decision tree might overestimate DO concentration.

Keywords : machine learning, LSTM(long short-term memory), decision tree, dissolved oxygen, hypoxia water mass, Jinhae Bay

1. 서 론

빈산소수괴(hypoxia water mass, HWM)란 용존산소(dissolved oxygen, DO) 농도가 3 mg L^{-1} 이하인 수괴를 말한다. DO 농도가 3.6 mg L^{-1} 이하로 내려가게 되면 저서동물의 폐사가 시작되며, 3 mg L^{-1}의 DO 농도는 해양 생물체 생존의 최저 인내한계로 작용한다. DO 농도가 2 mg L^{-1} 이하로 내려가게 되면 해양 생물체의 폐사가 시작되며 더 이상 해양 생물체가 살 수 없는 환경이 되어 해양생물자원이 감소하게 된다(Baden et al., 1990; Breitburg et al., 2018; Pearson and Rosenberg, 1978; Yin et al., 2004). 따라서 연안 수산자원의 보존과 지속 가능한 이용을 위해 DO 농도 및 빈산소수괴 발생의 예측·예보는 매우 중요한 과제 중 하나이다.

빈산소수괴는 산소 소비가 공급보다 많아 용존산소가 고갈되어 발생하며, 이러한 이유로 빈산소수괴는 반폐쇄성 연안 해역의 저층에서 주로 발생한다. 여름철 반폐쇄성 연안 해역은 외해와의 해수 교환이 억제되는 그 지리적 특성과 성층이 형성되는 계절적 영향으로 인해 저층으로의 산소 공급이 차단된다. 또한, 육지로부터 유입되는 다량의 유기물 및 영양염의 과잉 공급으로 인한 조류의 과증식은 저층 퇴적물로 유기물 부하를 증가시켜 산소 소비를 촉진한다. 이러한 이유로 국내의 대표적인 반폐쇄성 해역인 진해만에서는 매년 여름철 빈산소수괴가 발생하여 문제가 되고 있다.

최근 DO 농도의 예측 연구에는 딥러닝 모형인 Long Short-Term Memory(LSTM) 모형이 활발히 사용되고 있으며 우수한 예측 성능을 보이고 있다(Li et al., 2021b; Lim et al., 2020; Park and Kim, 2021). 그러나 대부분의 연구가 하천과 같은 1차원적인 흐름이나 정적인 연못에서 수행되었으며 단기 예측만을 고찰하고 있다. 특히 국내 반폐쇄성 연안 해역을 대상으로 한 DO 농도 예측에 관한 연구는 매우 미흡한 실정이다.

본 연구에서는 진해만을 대상으로 LSTM 모형을 이용한

*부경대학교 해양공학과 대학원생(Graduate student, Department of Ocean Engineering, Pukyong National University)
**한국가스공사 안전환경부 과장(Manager, Tongyeong Terminal Division, Korea Gas Corporation)
***부경대학교 해양공학과 교수(Corresponding author: Kyunghoi Kim, Professor, Department of Ocean Engineering, Pukyong National University, 45 Yongso-ro, Nam-Gu Busan 48513, Korea, Tel: +82-51-629-6583, hoikim@pknu.ac.kr)

DO 농도의 장·단기 예측 및 결정 트리(Decision tree) 모형을 이용한 빈산소수괴 발생 예측 연구를 수행하였다. 상기 연구 결과를 토대로 DO 농도 및 빈산소수괴 발생 예측에 대한 기계학습 모형의 성능을 평가하였다.

2. 재료 및 방법

2.1 Data analysis

본 연구에서는 진해만 내에 있는 국립수산과학원의 '실시간 해양환경 어장정보시스템'에서 제공하는 '통영 안정' 정점(34.9395°N, 128.4456°E) 자료를 취득하여 사용하였다(Fig. 1; NIFS, 2022). 항목은 표·저층의 수온, 염분, 탁도, Chl.a, 용존산소포화도(DO-S), 용존산소농도이며, 42개월간(2015.07.01.~2018.12.31.) 1시간 간격으로 관측되었다. 결측치(missing)와 이상치(outlier)는 선형보간하여 자료의 약 70%를 기계학습에 이용하였고 나머지 30%를 모형 검증에 사용하였다. 각 항목의 평균, 표준편차, 최대, 최소, 중앙값을 Table 1에 제시하였다. 저층 DO 농도를 종속변수, 그 외 항목들을 독립변수로 하여 다중선형회귀 및 t-검정(t-test)를 진행하여 그 결과를 Table 2에 나타내었으며, DO 농도와 각 항목의 상관분석

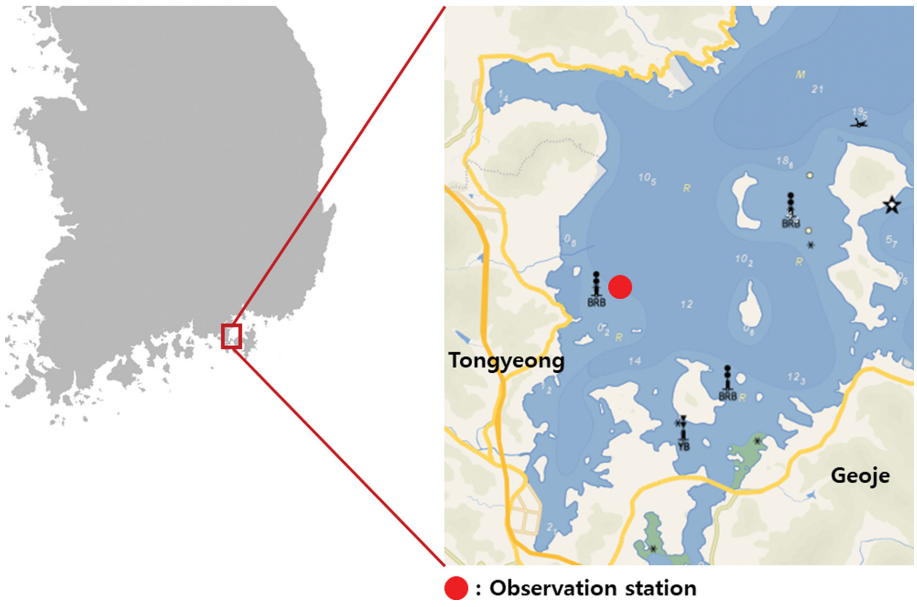

Fig. 1. Location of observation station in Jinhae Bay.

Table 1. The mean, standard deviation, min, max, median of features

	Temp. (°C)		Salinity (psu)		Turbidity (FTU)		Chl.a (μg/L)		DO-S (%)		DO (mg/L)	
	Sur.	Bot.	Sur.	Bot.	Sur.	Bot.	Sur.	Bot.	Sur.	Bot.	Sur.	Bot.
Mean	15.1	13.7	32.1	32.6	0.8	1.0	4.1	7.5	109.8	89.4	9.2	7.8
Std.	7.2	6.4	1.0	0.8	0.4	0.4	3.3	6.3	10.1	18.9	1.3	2.2
Max	30.5	27.5	33.6	33.8	4.8	4.0	27.0	144.3	176.1	120.7	12.4	11.9
Min	3.8	3.2	24.3	29.4	0.2	0.3	0.1	0.6	61.1	14.0	4.7	1.1
Median	14.7	13.0	32.3	32.7	0.6	1.0	3.2	5.8	108.7	96.9	9.3	8.2

Table 2. The t-test results of linear regression between features and bottom DO concentration

	Estimate	Std. error	t-stat	p-value
Intercept	5.237	0.25	21.29	0.000 < 0.05
Sur. temperature	−0.037	0.00	−9.79	0.000 < 0.05
Bot. temperature	−0.128	0.00	−31.66	0.000 < 0.05
Sur. salinity	−0.052	0.01	−7.41	0.000 < 0.05
Bot. salinity	−0.017	0.01	−2.07	0.039 < 0.05
Sur. turbidity	−0.051	0.01	−3.81	0.000 < 0.05
Bot. turbidity	0.053	0.01	3.83	0.000 < 0.05
Sur. Chl.a	−0.008	0.00	−4.58	0.000 < 0.05
Bot. Chl.a	−0.004	0.00	−3.90	0.000 < 0.05
Sur. DO-S	0.002	0.00	5.11	0.000 < 0.05
Bot. DO-S	0.077	0.00	218.37	0.000 < 0.05

Table 3. The $|R|$ values between DO concentration and features

$\|R\|$ between	Temp.		Salinity		Turbidity		Chl.a		DO-S	
	Sur.	Bot.	Sur.	Bot.	Sur.	Bot.	Sur.	Bot.	Sur.	Bot.
Sur. DO	0.80	0.82	0.20	0.33	0.12	0.21	0.21	0.04	0.32	0.55
Bot. DO	0.87	0.81	0.27	0.26	0.12	0.05	0.11	0.02	0.16	0.91

(correlation analysis)을 진행하여 그 결과를 Table 3에 나타내었다. 선형회귀의 t-test는 회귀예측에서 예측변수를 선별하는 방법의 하나로 '독립변수의 계수(또는 가중치)는 0이다'라는 귀무가설 검정을 진행한다. 검정 결과 그 확률(p-value)이 0.05 이하일 경우 가설은 기각되며 해당 항목은 예측변수로 선별된다. t-test 결과 모든 항목의 p-value가 0.05 이하로 나타나 모두 예측변수로 선별하였다. 상관분석에서는 상관계수(correlation coefficient, R)의 절대값($|R|$)에 따라 두 변수의 상관성을 평가할 수 있다. 항목 중 수온이 DO 농도와 가장 높은 상관성을 보였으며, 탁도와 Chl.a 농도는 상대적으로 낮은 상관성을 보였다.

LSTM을 이용한 DO 농도의 장·단기 예측에는 전체 자료 중 2017.01.01.~2018.07.01. 자료를 사용하였다. 시간 예측에는 기존 관측자료를 사용하였으며, 일 예측에는 자료를 일 평균하여 사용하였다. 자료의 수(Data point)는 시간 예측과 일 예측 각각 n_{hour} = 13,101, n_{day} = 546이다. 자료는 평균=0, 표준편차 = 1로 표준화(standardization)하였다.

결정 트리를 이용한 빈산소수괴 발생 예측에는 전체 자료(2015.07.01.~2018.12.31.)를 일 평균하여 일 예측만을 진행하였다(n = 1280). DO 농도는 3 mg L^{-1}를 기준으로 빈산소수괴 발생(HWM) 및 미발생(nonHWM)으로 범주화(categorical)하여 반응변수(response variable)로 사용하였다. 단, 전체 자료 중 빈산소수괴 발생 사례와 미발생 사례의 데이터 포인트는 각각 n_{HWM} = 97, n_{nonHWM} = 1183으로 데이터의 불균형이 존재한다. 이 불균형을 해결하기 위해 빈산소수괴 발생 사례에 대한 데이터 증식(augmentation)을 진행하여 미발생 사례와 데이터 수를 일치시켰다. 데이터 증식은 예측변수(predictor)인 수온, 염분, 탁도, Chl.a, DO-S에 –0.1~0.1 사이의 무작위 잡음을 추가하여 수행하였다.

2.2 LSTM

Long Short-Term Memory(LSTM)은 Recurrent Neural Network(RNN) 계열의 모형으로 기존 인공신경망 모형과 달리 Hidden cell간의 순환구조를 이루고 있다(Fig. 2). 이러한 순환구조로 LSTM은 sequence data 처리에 적합하다고 평가받는다(Dupond, 2019; Tealab, 2018). LSTM의 하이퍼파라미터(hyperparameter)는 Epoch, Number of hidden node, sequence length가 있다. 하이퍼파라미터의 값에 따라 정확도와 필요비용이 결정되기 때문에 비용과 성능을 고려하여 하이퍼파라미터를 최적화(optimization)하는 것이 중요하다. LSTM은 입력자료의 길이(sequence length)와 출력의 개수에 따라 다음과 같은 4가지 구조를 갖는다. one-to-one, one-to-many, many-to-many 그리고 many-to-one. 본 연구의 LSTM 모형은 sequence length가 1인 경우를 제외하고 many-to-one의 구조를 갖는다. 즉, 본 연구에서는 과거 일정 기간의 시계열 자료를 입력받아 수일 후의 DO 농도를 예측하였으며, case study를 통해 적합한 하이퍼파라미터 조건을 찾는 것을 목적으로 한다. LSTM의 각 gate와 state에 대한 식은 아래와 같다.

$$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + b_f) \quad (1)$$

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + b_i) \quad (2)$$

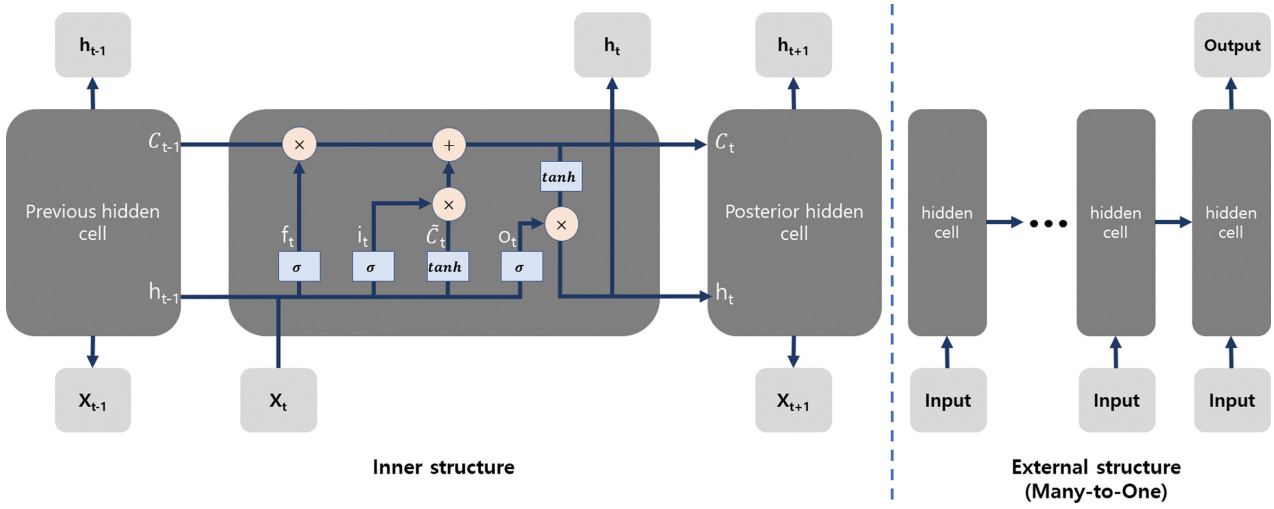

Fig. 2. The inner and external structure of LSTM.

$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + b_o) \quad (3)$$

$$\tilde{C}_t = \tanh(W_{x\tilde{C}}x_t + W_{h\tilde{C}}h_{t-1} + b_{\tilde{C}}) \quad (4)$$

$$C_t = C_{t-1} \cdot f_t + i_t \cdot \tilde{C}_t \quad (5)$$

$$h_t = o_t \cdot \tanh(C_t) \quad (6)$$

여기서, f_t는 forget gate, i_t는 input gate, o_t는 output gate, C_t는 Cell state, h_t는 hidden state, W는 가중치(weight), b는 편향(bias)이다.

2.3 Decision tree

결정 트리(Decision tree)는 기계학습 기반의 데이터 분류(classification) 기법이다. 결정 트리는 나무 형태의 구조로 뿌리 노드(root node)로부터 분기된 노드들이 모여 의사결정나무를 형성한다. 각각의 노드는 분기 조건을 가지며 최종적으로 말단 노드(leaf node)에서 객체의 클래스가 결정된다. 각 분기마다 불순도가 감소하는 방향으로 노드를 형성하여 의사결정나무를 완성한다. 이후 모형의 과대적합(overfitting)을 방지하기 위해 부적절한 분기 조건을 갖는 노드를 제거하는 가지치기 과정을 거친다. 결정 트리에 대한 구조를 Fig. 3에 나타내었다.

2.4 Experiment model condition

LSTM 모형을 이용한 DO 농도의 장·단기 예측 연구에서는 모형의 하이퍼파라미터에 따른 예측 정확도를 비교하였다. 하이퍼파라미터 항목으로는 LSTM 모형의 number of Hidden node, Epoch, Sequence length이다. 예측 시간 간격은 시간 예측에서 1, 6, 12, 24, 36, 48 hour, 일 예측에서 1, 3, 6, 10, 15, 30 day를 고려하였다. 위 실험을 통해 DO 농도의 장·단기 예측에 적합한 LSTM 모형의 하이퍼파라미터 값을 찾고자 하였다.

결정 트리를 이용한 빈산소수괴 발생 예측에서는 일 예측만을 진행하였으며, 예측 시간 간격은 1, 3, 6, 10, 15, 30 day를 고려하였다. 트리의 분기 조건에는 지니 지수(Gini index)를 사용하였으며, 예측 결과는 빈산소수괴 발생(HWM)과 미발생(nonHWM) 2가지 범주형 변수로 표현하였다. 본 연구에서 고려된 LSTM 모형 조건과 결정 트리 모형 조건을 Table 4에 제시하였다. 연구의 진행도를 Fig. 4에 요약하여 나타내었다.

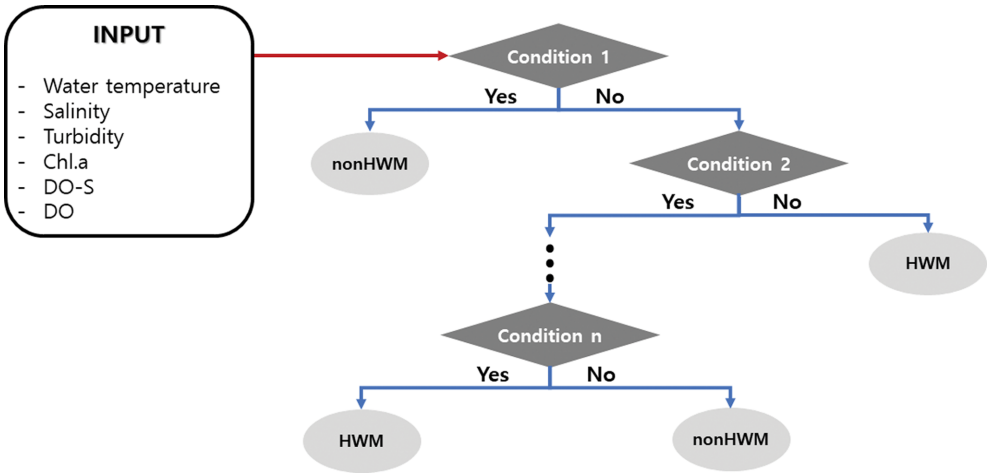

Fig. 3. The structure of decision tree.

Table 4. Experiment model condition

LSTM - prediction in DO concentration			
Time step	Hyperparameter		
	Sequence length	Epoch	Hidden node
1, 6, 12, 24, 36, 48 (hours) 1, 3, 6, 10, 15, 30 (days)	1, 6, 12, 24, 36, 48 1, 3, 5, 10, 15, 30	300, 600, 900, 1200, 1500	10, 20, 30, 40, 50

Decision tree - prediction in hypoxia water mass		
Time step	Split criterion	Response variable
1, 3, 6, 10, 15, 30 (days)	Gini's diversity index	HWM or nonHWM (categorical)

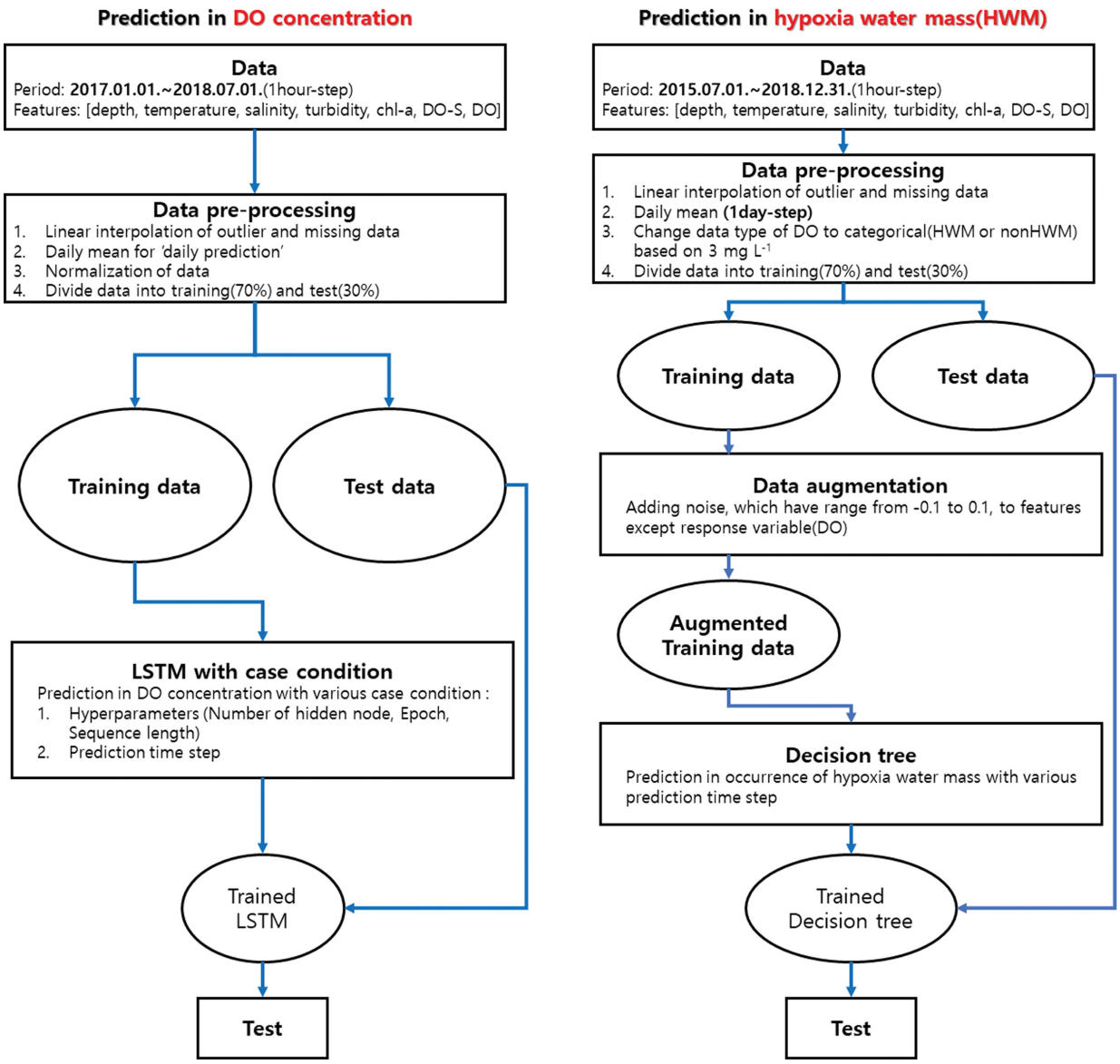

Fig. 4. The flowchart of prediction in DO concentration and hypoxia water mass.

3. 결 과

3.1 Prediction in DO concentration based on LSTM

본 연구에서는 진해만 DO 농도의 장·단기 예측에 적합한 LSTM 모형의 하이퍼파라미터값을 찾기 위한 case study를 진행하였다. 여러 하이퍼파라미터값 조건에서 DO 농도를 장·단기 예측하였으며, 그 결과를 결정계수(R^2) 값으로 비교·평가하였다. Hidden node의 개수에 따른 예측 정확도를 Fig. 5에 나타내었으며, Sequence length와 Epoch에 따른 예측 정확도를 heat map으로 Fig. 6과 Fig. 7에 나타내었다. 그 이후에 모형의 시계열 예측성을 평가하기 위해 [1 hour, 48 hour]과 [1 day, 30 day] 예측 시간 간격에 대한 예측값과 관측값의 시계열 비교 그래프를 Fig. 8에 나타내었다.

Short Sequence length, Less Epoch과 Long Sequence length, Enough Epoch 조건에서 Hidden node의 개수에 따른 예측 정확도를 R^2 값으로 나타내었다(Fig. 5). Short Sequence length, Less Epoch 조건의 시간 예측에서 Hidden node의 개수에 따른 정확도 차이는 나타나지 않았다. 일 예측에서는 Hidden node의 개수가 많아질수록 R^2 값은 감소하는 결과를 보였다. 저층 DO 농도 예측에서 Hidden node = 10일 때 6 day와 10 day 예측의 R^2 값은 각각 0.75, 0.74에서 Hidden node가 증가함에 따라 0.36, 0.50으로 감소하였다. 이는 Hidden node의 개수 증가로 모형의 복잡도가 증가하면서 Epoch = 300 조건에서 모형이 과소적합(Underfitting)된 것으로 판단된다. 반면에 Long Sequence length, Enough Epoch 조건의 일 예측에서는 Hidden node의 개수가 많아질수록 R^2 값은 증가하는 결과를 보였다. 표층 DO 농도 예측에서 Hidden node = 10일 때 모든 예측 시간 간격의 평균 R^2 값은 0.44에서 Hidden node가 50으로 증가함에 따라 0.71로 증가하였으며, 저층 DO 농도 예측에서도 평균 R^2 값이 0.60

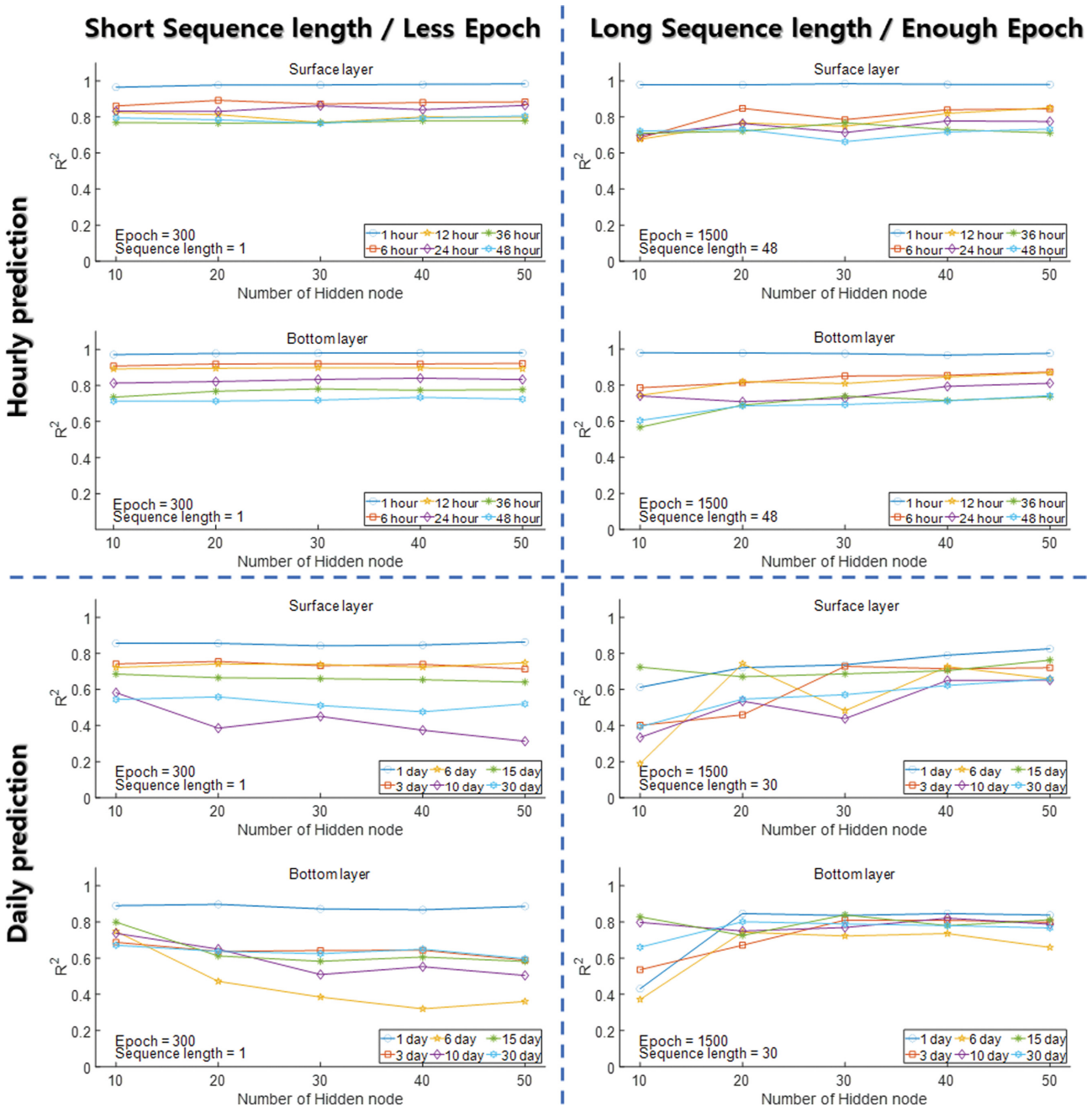

Fig. 5. DO prediction result R^2 values with respect to number of hidden node.

에서 0.78로 증가하는 결과를 보였다. 이는 Hidden node = 10일 때 Epoch = 1500 조건에서 과대적합(Overfitting)이 발생한 것으로 판단되며, Hidden node = 50에서 모형의 복잡도가 증가함에 따라 요구되는 Epoch도 증가하여 과대적합이 해소된 것으로 판단된다.

Sequence length와 Epoch에 따른 예측 정확도를 비교하기 위해 Hidden node의 수를 30으로 고정하고 R^2 값을 heat map으로 나타내었다(Fig. 6; Fig. 7). 시간 예측 결과, 표·저층 DO 농도의 1 hour 예측에서 하이퍼파라미터 조건에 따른 R^2 값의 유의미한 차이는 보이지 않았다. Epoch에 따른 시간 예측 정확도는 Epoch = 300 또는 Epoch = 600의 낮은 Epoch 조건에서 상대적으로 높은 R^2 값을 보였다. 낮은 Epoch 조건에서 상대적으로 높은 R^2 값을 보인 것은 900 이상의 Epoch 조건에서 과대적합이 발생한 것으로 판단된다. Sequence length에 따른 예측 정확도는 표층의 12 hour 예측을 제외하고 표·저층 예측의 모든 예측 시간 간격에서 Sequence length = 1일 때 가장 높은 R^2 값을 보였다. Sequence length = 1에서 높은 R^2 값을 보인 것은 상대적으로 짧은 예측 시간 간격에 의한 결과로 판단된다. 시간 예측에서 고려된 예측 시간 간격은 최대 48 hour로 DO 농도의 시계열 변화 특성이 나타나지 않아 긴 Sequence length 조건에서 낮은 R^2 값을 보인 것으로 판단된다. 예측 시간 간격이 1 day에서 최대 30 day까

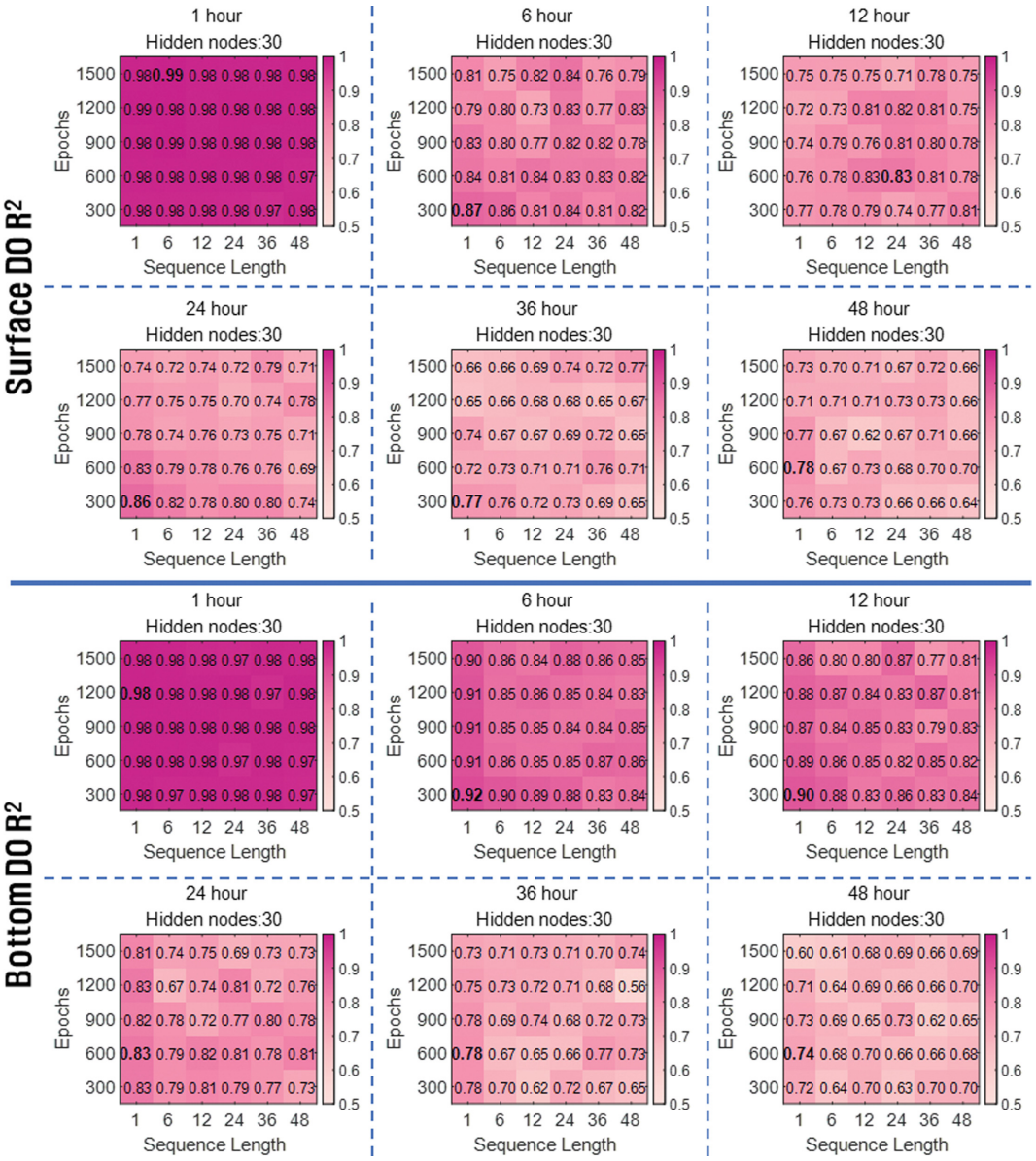

Fig. 6. DO prediction R^2 values with respect to Epoch and Sequence length in hourly prediction.

지 고려된 일 예측에서는 예측 시간 간격이 길어짐에 따라 긴 Sequence length 조건에서 상대적으로 높은 R^2 값을 보였다. 저층 DO 농도의 1 day 예측에서 R^2의 최댓값은 0.88로 Sequence length = 1 조건에서 나타났다. 이후 3 day 예측에서는 Sequence length = 15 조건에서 0.84의 R^2 최댓값이 나타났으며, 15 day와 30 day 예측에서는 Sequence length = 30 조건에서 각각 0.85, 0.79의 R^2 최댓값을 보였다. 표층 DO 농도 예측에서도 R^2의 최댓값이 나타난 Sequence length 조건은 [1, 3, 6, 10, 15, 30] day 예측에서 각각 [3, 6, 30, 15, 15, 30]으로 나타나 예측 시간 간격이 길어질수록 긴 Sequence length를 요구하는 경향을 보였다. 이는 예측 시간 간격이 길수록 DO 농도의 시계열 변화 특성이 강하게 반영되어 나타난 결과로 판단된다.

LSTM 모형의 시계열 예측 성능을 평가하기 위해 시간 예측과 일 예측의 장·단기 예측 결과를 관측값과 함께 시계열 그래프로 나타내었다(Fig. 8). 시간 예측과 일 예측의 예측 기간은 각각 약 1개월과 5개월이며, 장·단기 예측 시간 간격은 다음과 같다. 시간 예측: [단기, 장기] = [1 hour, 48 hour]/일 예측: [단기, 장기] = [1 day, 30 day]. 모형 조건은 각각의 예측에서 R^2 값이 가장 높게 나온 하이퍼파라미터 조건을 사용하였다. 단기 예측(Short-term prediction)의 경우 시간 예측과 일 예측의 예측값 모두 관측값을 잘 재현하는 것으로 나타났다. 단기-시간 예측의 표층과 저층 R^2 값은 각각 0.99, 0.98로 나타났으며, 단기-일 예측의 경우 표층과 저층 각각 0.88, 0.90으로 나타났다. 장기-시간 예측에서 표층의 경우 06-27 이후에 급격히 변하는 DO 농도 변화는 재현하지 못했다. 저층의 경우 관측값과 비교했을 때 예측값이 일정 시간 간격 뒤로 밀리는 모습을 보였다. 이러한 결과는 Li et

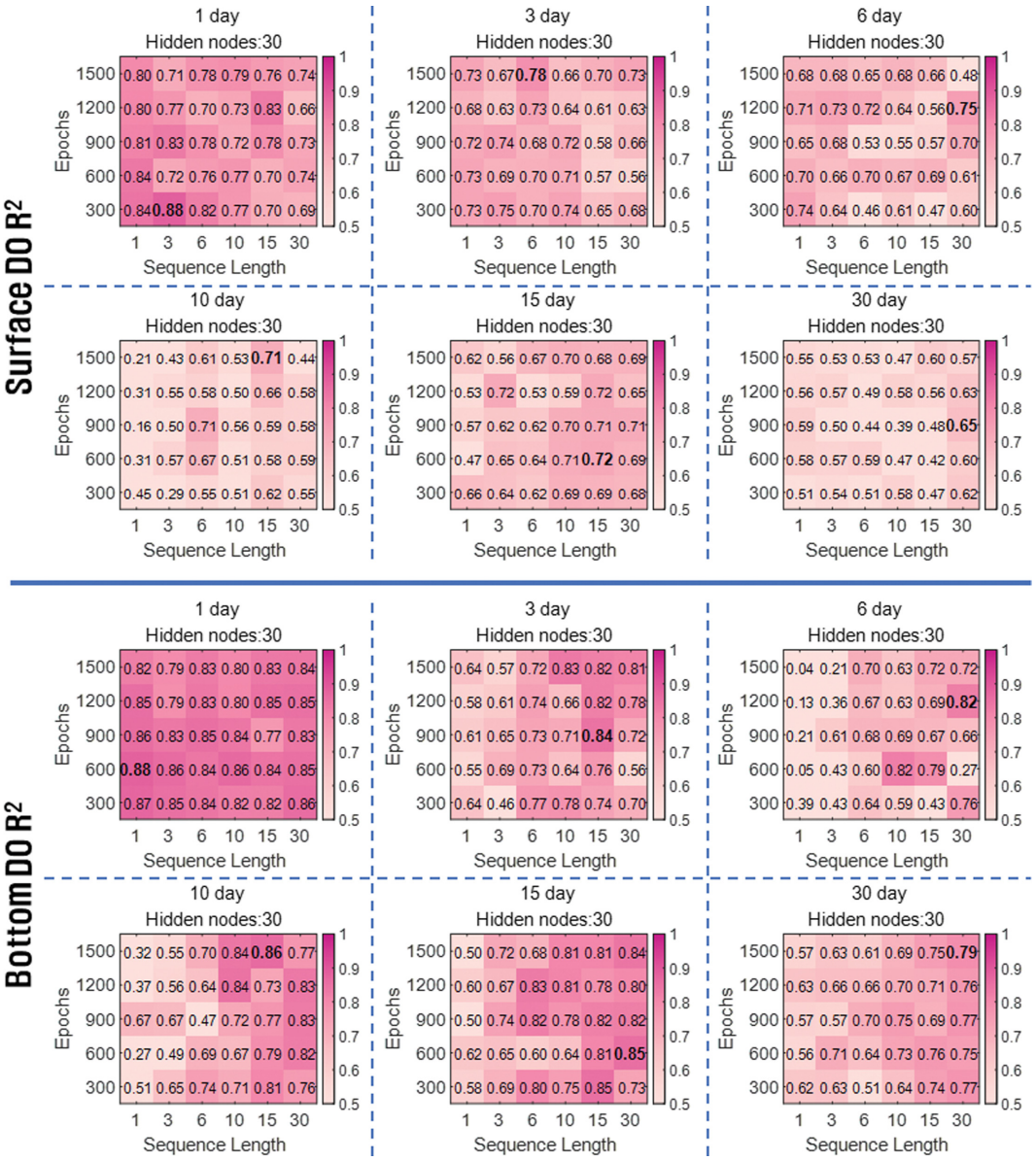

Fig. 7. DO prediction R^2 values with respect to Epoch and Sequence length in daily prediction.

al.(2021a)의 연구에서도 확인할 수 있으며, 이는 DO 농도의 비정상성(non-stationary)과 복잡성에 기인한 것으로 판단된다. 장기-일 예측의 경우 표층과 저층 모두 단기적인 재현성은 떨어지는 모습을 보였다. 그러나 30 day라는 긴 예측 시간 간격을 고려하여 그 목적을 장기적인 추세 예측으로 제한했을 때 높은 재현성을 보였다.

3.2 Prediction in hypoxia water mass based on decision tree

본 연구에서는 결정 트리를 이용한 저층의 빈산소수괴 발생 예측 연구를 진행하였다. 그 결과를 정오분류표(confusion matrix)로 Fig. 9에 나타내었으며, 예측 시간 간격별 예측 정확도 그래프로 Fig. 10에 요약하여 나타내었다. 반응변수는 빈산소수괴 발생(HWM) 또는 미발생(nonHWM)의 범주형 변수이다. 예측 시간 간격으로는 1, 3, 6, 10, 15, 30 day를 고려하였다. HWM 케이스의 예측 결과, 예측 시간 간격 10 day 이하의 단기 예측에서 약 80% 이상의 높은 예측 정확도를 보였다. 이후 예측 시간 간격이 늘어남에 따라 예측 정확도는 급격히 감소하였다. HWM 케이스의 예측 정확도는 [1, 3, 6, 10, 15, 30] day 예측에서 각각 [83.3, 87.5, 79.2, 83.3, 62.5, 37.5]%로 나타났다. nonHWM 예측에서도 단기 예측에서 장기 예측으로 갈수록 예측 정확도는 감소하는 결과를 보였다. 1 day 예측의 정확도는 88.3%이었으며, 30 day 예측에서 66.1%로 감소하였다. 단, [10, 15, 30] day 예측만을 봤을 때는 오히려 정확도가 증가하는 결과를 보였다. 장기 예측에서 HWM 예측 정확도의 급격한 감소와 nonHWM 예측 정확도가 증가하는 결과를 봤을 때, 결정 트리 모형이 DO 농도를 관측치보다 고평가(overestimate)하는 경향이 있는 것으로 판단된다. 전체 케이스에 대한 정확도는 nonHWM 케이스의 정확도와 큰 차이를 보이지 않았다(Fig. 10). 이는

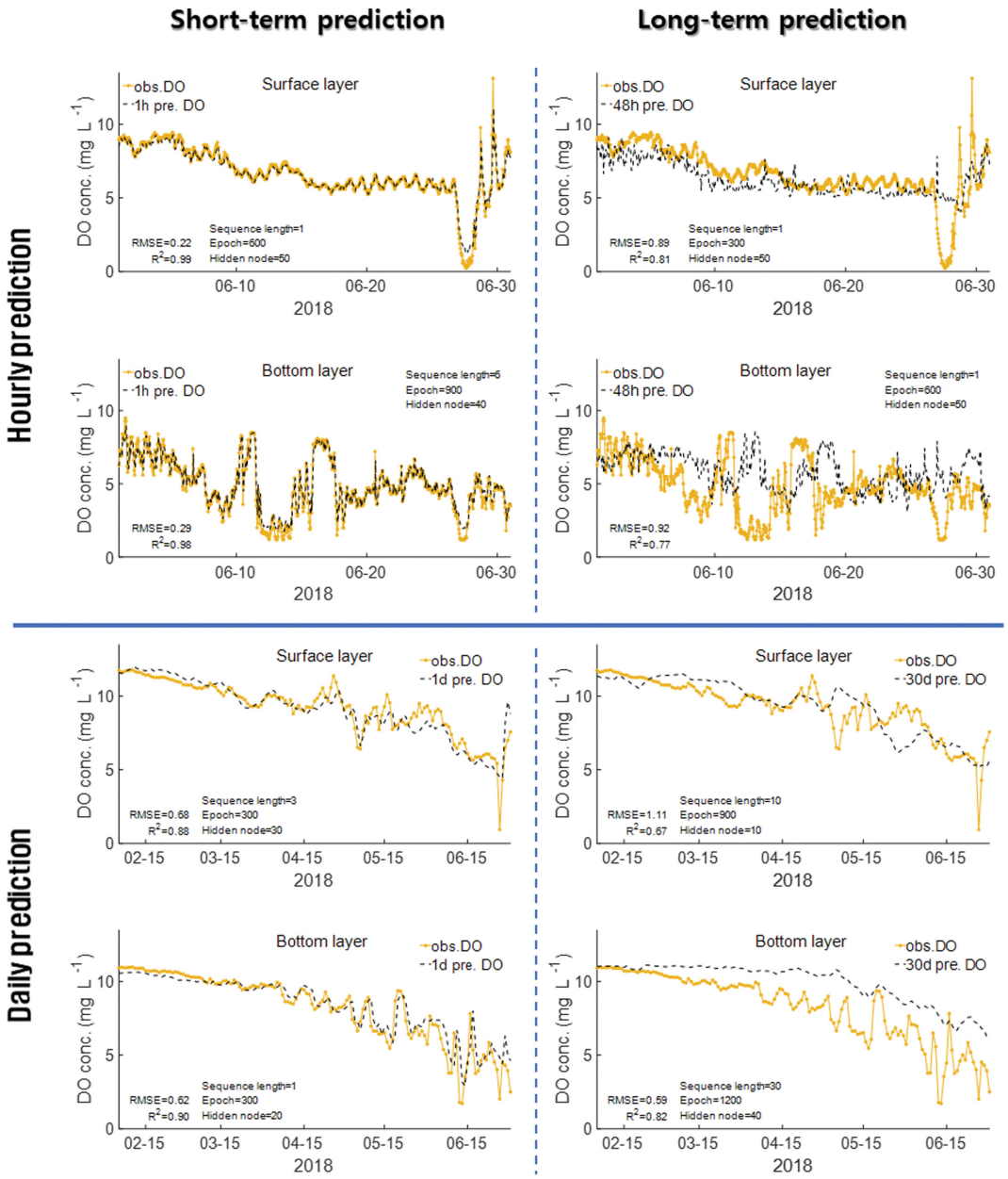

Fig. 8. Time series of observed (line) and predicted (dotted line) DO concentration.

전체 데이터 포인트 수(n = 1280)에서 nonHWM 케이스의 데이터 수(n = 1183)가 차지하는 비중이 커서 나타난 결과로 판단된다.

4. 결 론

본 연구에서는 통영 연안의 한 정점을 대상으로 LSTM 모형을 이용한 DO 농도의 장·단기 예측 및 결정 트리 모형을 이용한 빈산소수괴 발생 예측 연구를 수행하였다. 연구 결과를 토대로 DO 농도 및 빈산소수괴 발생 예측에 대한 기계학습 모형의 성능을 평가하였다.

LSTM 모형을 이용한 DO 농도 예측 연구 결과, Hidden node의 수가 적은 경우 모형의 복잡도가 낮아서 높은 Epoch에서 과대적합이 발생할 우려가 있다. 반면에 Hidden node의 수가 증가할수록 모형의 복잡도도 증가하면서 많은 Epoch을 요구하는 모습을 보였다. 단기-시간 예측에서 장기-일 예측으로 갈수록 DO 농도의 시계열 변화 특성이 강하게 반영되어 긴 Sequence length에서 높은 정확도를 보였다. 따라서 LSTM 모형을 구축할 때는 가진 자료의 수와 모형의 목적 그리고 비용과 성능을 고려하여 하이퍼파라미터를 결정해야 한다.

결정 트리를 이용한 빈산소수괴 발생 예측 연구 결과, 1 day 발생(HWM) 예측시 약 83.3%의 정확도를 보였다. 이후 예측 시간 간격이 증가할수록 정확도는 점차 감소했으며, 30 day 발생 예측에서는 37.5%의 정확도를 보였다. 반면에 30 day 미발생(nonHWM) 예측 정확도는 66.1%로 상대적으로 높게 나타났다. 이는 결정 트리 모형이 DO 농도를 관측값보

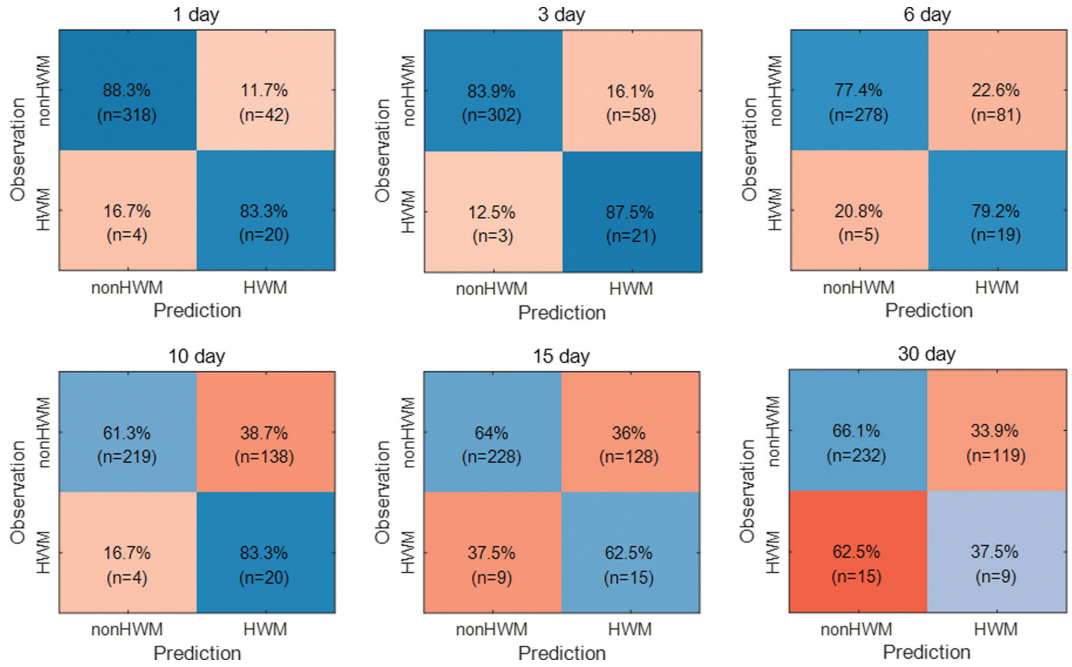

Fig. 9. Confusion matrix of prediction accuracy (%) in HWM and nonHWM.

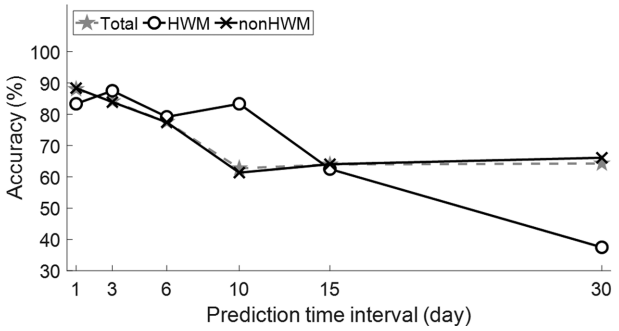

Fig. 10. Prediction accuracy in hypoxia water mass for HWM, non-HWM and total cases.

다 고평가(overestimate)하는 경향이 있는 것으로 판단되며, 이에 관한 후속연구가 필요할 것으로 사료된다.

감사의 글

이 논문은 2022년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(Grant 2021R1I1A3060374126178206434010 2).

References

Breitburg, D., Levin, L.A., Oschlies, A., Grégoire, M., Chavez, F.P., Conley, D.J., Garçon, V., Gilbert, D., Gutiérrez, D., Isensee, K., Jacinto, G.S., Limburg, K.E., Montes, I., Naqvi, S.W.A., Pitcher, G.C., Rabalais, N.N., Roman, M.R., Rose, K.A., Seibel, B.A., Telszewski, M., Yasuhara, M. and Zhang, J. (2018). Declining oxygen in the global ocean and coastal waters. Science, 359 (6371).

Baden, S.P., Loo, L.O., Pihl, L. and Rosenberg, R. (1990). Effects of eutrophication on benthic communities including fish: Swedish west coast. AMBIO A Journal of the Human Environment, 13(3), 113-122.

Dupond, S. (2019). A thorough review on the current advance of neural network structures. Annual Reviews in Control, 14, 200-230.

Li, Q., Xang, X., Wang, J. and Zhou, Y. (2021a). Prediction of dissolved oxygen content in water based on EEMD-Pearson and LSTM hybrid models. Earth and Environmental Science, 760(1).

Li, W., Wu, H., Zhu, N., Jiang, Y., Tan, J. and Guo, Y. (2021b). Prediction of dissolved oxygen in a fishery pond based on gated recurrent unit (GRU). Information Processing in Agriculture, 8(1), 185-193.

Lim, H., An, H., Choi, E. and Kim, Y. (2020). Prediction of the DO concentration using the machine learning algorithm: case study in Oncheoncheon, Republic of Korea. Korean Journal of Agricultural Science, 47, 1029-1037 (in Korean).

NIFS (National Institute of Fisheries Science) (2022). real-time marine environment fishing ground information system, https://www.nifs.go.kr/risa/main.risa (in Korean).

Park, S. and Kim, K. (2021). Prediction of DO concentration in nakdong river estuary through case study based on long short term memory model. Journal of Korean Society of Coastal and Ocean Engineers, 33(6), 1-8 (in Korean).

Pearson, T.H. and Rosenberg, R. (1978). Macrobenthic succession on relation to organic enrichment and pollution of the marine environment. Oceanography and Marine Biology, 16, 229-311.

Tealab, A. (2018). Time series forecasting using artificial neural

networks methodologies: A systematic review. Future Computing and Informatics Journal, 3(2), 334-340.

Yin, K., Lin, Z. and Ke, Z. (2004). Temporal and spatial distribution of dissolved oxygen in the Pearl River Estuary and adjacent coastal waters. Continental Shelf Research, 24(16), 1935-1948.

Received 28 April, 2022
Revised 25 May, 2022
Accepted 7 June, 2022