한국해양환경·에너지학회지
J. Korean Soc. Mar. Environ. Energy
Vol. 27, No. 3, 153-166, August 2024

https://doi.org/10.7846/JKOSMEE.2024.27.3.153
ISSN 2288-0089(Print) / ISSN 2288-081X(Online)

Original Article

간헐적 시계열자료를 활용한 인공지능 모델 기반 마산만 저층 용존산소 농도 장기 변화 예측

김수빈[1] · 이재성[2,3] · 김성한[2,3] · 김경태[2,†]
[1]한국해양과학기술원 해양환경연구부 기술원
[2]한국해양과학기술원 해양환경연구부 책임연구원
[3]해양과학기술전문대학원 해양과학기술융합과 겸임교수

AI Enhanced Prediction of Long Term Change in Bottom Dissolved Oxygen Concentration Using Intermittent Time Series Data Observed at Masan Bay

Soobin Kim[1], Jaeseong Lee[2,3], Sunghan Kim[2,3], and Kyungtae Kim[2,†]
[1]Research Specialist, Marine Environmental Research Department, Korea Institute of Ocean Science & Technology (KIOST), Busan 49111, Korea
[2]Principal Research Scientist, Marine Environmental Research Department, Korea Institute of Ocean Science & Technology (KIOST), Busan 49111, Korea
[3]Adjunct Professor, Department of Convergence Study on the Ocean Science & Technology, School of Ocean Science & Technology (OST), Busan 49112, Korea

요 약

최근 기후변화, 육상 기원 오염원 증가, 생지화학적 및 물리적 요인 등으로 인해 해수 중 용존산소(dissolved oxygen, DO) 농도가 감소하여 빈산소 수괴(hypoxia water mass)가 발생하고 있다. 빈산소 수괴의 발생은 해양생태계를 급격히 변화시키고 사회·경제적인 피해를 가져올 수 있다. 해양 수질 및 생태계의 효율적인 관리를 위해 DO 농도 변화 예측과 빈산소 수괴 발생 영향인자를 파악할 필요가 있다. 이 연구에서는 최근 빈산소 수괴 발생이 우려되는 마산만 인근 연안의 저층 DO 농도를 예측하고자 한다. 해양환경측정망 및 어장환경모니터링 자료 중 조사일시, 수심, 투명도, 표층 및 저층의 수온, 염분, 수소이온농도, DO, 화학적산소요구량, 암모니아성질소, 아질산성질소, 질산성질소, 용존무기질소, 총질소, 용존무기인, 총인, 규산성규소, 부유물질, 엽록소 a 자료를 수집하여 결합하였다. 조위, 풍향, 풍속, 기온, 기압자료는 조위관측소, 유속, 유향자료는 해양관측부이에서 측정된 자료를 사용하였다. 일강수량 자료는 종관기상관측 자료를 활용하였고 일사량 자료는 해양수질자동측정망 자료에서 추출하였다. 유입 하천수질 자료는 수질측정망 자료, 하수처리시설 방류수 수질자료는 전국오염원조사 자료에서 수집하였다. 수집한 자료를 모두 결합하여 결측치를 제거하고 신경망(neural network) 모델 학습 기반 자료 합성(data synthesis) 방법을 이용하여 자료의 수를 증식(augmentation)시켰다. 자료의 비정상성(non-stationarity)을 경감하기 위해 경험웨이블릿변환(empirical wavelet transform)으로 자료를 분해(decomposition)하고 교차웨이블릿변환(cross wavelet transform)을 정규화하여 얻은 웨이블릿 긴밀도(wavelet coherence)를 비교하여 모델 입력변수를 선택하였다. 모델의 출력변수는 저층 DO 농도로 설정하였다. 랜덤포레스트 회귀(random forest regression), 계절성 자기회귀누적이동평균(seasonal autoregressive integrated moving average), 장단기메모리(long short-term memory)신경망 알고리즘으로 모델 학습을 하였다. 모델의 성능 평가를 위해 평균제곱근오차(root mean squared error), 평균절대비오차(mean absolute percentage error), 수정 결정계수(adjusted coefficient of determination), 상관계수(correlation coefficient)를 사용하였다. 평가 결과 모델의 예측 성능이 상이하였으나 저층 DO 농도 변화가 급격한 구간에서는 성능이 낮았고 계절 변동성만 근사하게 예측하였다. 결과적으로 이 연구는 반폐쇄성 내만의 저층 DO 농도 변화에 영향을 미치는 인자를 확인하였고 향후 개선을 통해 실시간 빈산소 수괴 발생 예측을 통한 사전 경보, 해양환경 복원·규제 정책 지원, 연안오염총량관리 오염부하 삭감 이행평가 등에 활용될 수 있다. 나아가 기존 수치모델과 연계하였을 때 예측 결과에 대한 정밀한 이론적 해석이 가능하며 자료 처리 방식과 자료의 양과 질이 보완된다면 저층 DO 농도 변화의 정확한 예측이 가능할 것이다.

†Corresponding author: ktkim@kiost.ac.kr

Abstract – Recently, climate change, the increase in land-based sources of pollution, and various biogeochemical and physical factors have led to a decrease in the concentration of dissolved oxygen (DO) in seawater, resulting in the formation of hypoxic water masses. The occurrence of hypoxic water masses can drastically alter marine ecosystems and lead to significant socio-economic damage. For the effective management of marine water quality and ecosystems, it is essential to predict changes in DO concentrations and identify the factors influencing the occurrence of hypoxic water masses. This study aims to predict the bottom DO concentrations in the coastal waters near Masan Bay, where the occurrence of hypoxic water masses has recently become a concern. Marine environmental data, hydroclimatic data, ocean current data, and land-based pollution source data were collected for model training. All collected data were combined, missing values were removed, and the data were augmented using a neural network model-based data synthesis method. To mitigate the non-stationarity of the data, the empirical wavelet transform (EWT) was used for data decomposition, and cross wavelet transform (CWT) was normalized to obtain wavelet coherence. These wavelet coherences were compared to select the model input variables. The output variable of the model was set to the bottom DO concentration. The model was trained using random forest regression, seasonal autoregressive integrated moving average (SARIMA), and long short-term memory (LSTM) neural network algorithms. To evaluate the performance of the model, root mean squared error (RMSE), mean absolute percentage error (MAPE), adjusted coefficient of determination (adjusted R^2), and correlation coefficient were used. The evaluation results showed that the predictive performance of the models varied; however, the performance was lower in regions with abrupt changes in bottom DO concentrations, and the models only approximated the seasonal variability. In conclusion, this study identified the factors influencing changes in bottom DO concentrations in semi-enclosed coastal bays. Future improvements could enable real-time prediction of hypoxic water mass occurrences, providing early warnings and supporting marine environmental restoration and regulation policies, as well as evaluating the reduction of pollution loads in coastal total maximum daily load (TMDL) management. Furthermore, when integrated with existing numerical models, precise theoretical interpretation of the prediction results is possible. If the data processing methods and the quantity and quality of data are improved, accurate prediction of changes in bottom DO concentrations will be achievable.

Keywords: 빈산소 수괴(Hypoxic water mass), 경험 웨이블릿 변환(EWT), 계절성 자기회귀누적이동평균(SARIMA), 랜덤포레스트 회귀(RFR), 장단기메모리(LSTM), 마산만(Masan bay)

1. 서 론

해양환경 중 용존산소(dissolved oxygen, DO)는 해양생물 및 생태계에 매우 중요한 인자이다. 해양의 용존산소 농도는 물리(physical), 역학(mechanical), 생지화학(biogeochemical), 수문기상학적(hydroclimatic) 요인에 의해 변화한다(Zhang *et al.*[2010]). 해양환경 중 용존산소 농도가 2 $mg \cdot L^{-1}$ 이하가 되면 저서동물(benthic fauna)이 이탈하고, 0.5 $mg \cdot L^{-1}$ 이하에서는 대부분의 해양생물에 치명적인 영향을 미친다(Diaz and Rosenberg[2008]). 우리나라에서는 해양환경 중 용존산소 농도가 3 $mg \cdot L^{-1}$ 이하로 감소할 때 빈산소 수괴(hypoxia water mass)로 정의하고 있다(NIFS[2009]). 최근 기후변화, 육상 기원 오염원 유입, 연안 이용 및 개발 등으로 인해 빈산소 수괴가 우리나라 연안에서 빈번히 발생하고 있다(Breitburg *et al.*[2018]).

하계에 우리나라 대표적인 반폐쇄성 내만(semi-closed bay)으로 알려진 경상남도 창원시 일대의 마산만(Masan bay)에서도 수심 5 m 이하의 저층에서 빈산소 수괴가 발생하고 있다(Park[2020]). 반폐쇄성 내만에서는 해수교환 불량, 수온·염분차에 의한 밀도성층(density stratification) 형성에 의해 저층에 산소 공급이 차단된다. 실제로 마산만 내측은 하계에 간조와 만조의 DO 농도차가 외측보다 작게 나타났다(Yoo and Kim[2019]).

우리나라에서는 지속가능한 해양환경 관리를 위해 해양환경을 지속적으로 모니터링(monitoring)하고 있다. 하지만, 불충분한 모니터링 빈도(sparsity), 측정 오차(measurement error), 측정 장비의 노후화 및 유지상태 불량, 과도한 비용 및 노동력 부족 등으로 인해 해양환경 변화의 정확한 파악이 어렵다. 그러므로 최근 환경자료의 인공지능 모델 학습을 통해 환경변화를 예측하는 기술이 제안되고 있다. Coopersmith *et al.*[2011]은 멕시코만(Gulf of Mexico) 인근 Corpus Christi Bay에서 저층 DO 연속관측 자료의 연속 정규화(sequential normalization)를 통해 추세 제거(detrending)를 한 후 1일 후 저층 DO, 염분, 수온, 풍속, 풍향 자료와 결합하여 K-최근접 이웃(K-nearest neighbor, KNN) 알고리즘으로 학습하였다. 저층 DO 농도를 예측하였고 공간 보간(spatial interpolation)을 통해 빈산소 발생 확률을 공간적으로 예측하여 추가 모니터링이 필요한 정점 위치를 제안하였다. Muller and Muller[2015]는 미국 Chesapeake Bay에서 기후학적(climatological) 자료(Oceanic Nino Index, Susquehanna River index, Cross-Bay wind data)를 이용하여 빈산소 용적지수(hypoxic volume index)를 외생변수 비선형 자기회귀 신경망(nonlinear autoregressive neural network with exogenous inputs, NARX) 모델로 예측하였고, 연속 웨이블릿 변환(continuous wavelet transform, CWT), 교차 웨이블릿 변환(cross-wavelet transform, XWT), 웨이블릿 긴밀도(wavelet coherence, WC)를 통해 자료의 주기성(periodicity), 교차상관관계 분석을 수행하여 빈산소 발생에 영향을 미치는 기후학적 변수를 추정하였다. Ross and Stock[2019]은 Chesapeake Bay에서 연직 밀도차(vertical density), 수온, 수심, 평균해수면(mean sea level), 풍속, 총질소, 경위도, 유입하천 유량 자료를 이용하여 월평균 수층별 최소 DO

농도를 model tree 알고리즘을 이용하여 예측하였다. 자료의 계절성(seasonality)을 배제하기 위하여 이상치(anomalies)로 변환한 자료를 학습에 사용하였고, 개별 조건부 기대치(individual conditional expectation, ICE), 부분의존도(partial dependence, PDP)를 통해 모델의 민감도(sensitivity) 분석을 하였다. Yu *et al.*[2020]은 영양염류 부하(nutrient loading), 강의 유량, 기온, 일사량(solar radiation), 풍속, 풍향 자료를 시간 지연(time-lag)과 축적(accumulation)을 고려하여 변환하고 DO 자료를 경험고유함수(empirical orthogonal function, EOF)로 시공간적 요소(spatial-temporal component)로 분해(decomposition)하여 차원 축소(dimension reduction)한 DO 자료를 이용하여 시공간적 DO 변화를 신경망 모델로 예측하였다. Valera *et al.*[2020]은 시간, 수온, 염분, 수심, 용승 바람 응력(upwelling wind stress) 자료를 랜덤포레스트 회귀(random forest regression, RFR), 서포트 벡터 회귀(support vector regression, SVR) 알고리즘으로 학습하여 미국 San Luis Obispo Bay의 연안(nearshore)에서는 자료의 수와 정점에 따른 DO 농도 예측에 대한 모델 민감도를 분석하였고, 외해(offshore)에서는 수심에 따른 DO 농도를 예측하였다. Lui *et al.*[2021]은 하와이 제도(Hawaiian Islands) Hilo 인근 해역 정점에서 수집한 수온, 염분, 탁도(turbidity), 엽록소(chlorophyll), 산소 포화도(oxygen saturation) 시계열 자료에 경험 웨이블릿 변환(empirical wavelet transform, EWT)을 적용한 자료를 이용하여 Broyden-Fletcher-Goldfarb-Shanno(BFGS), elman neural network(ENN), general regression neural network(GRNN), outlier-robust extreme learning machine(ORELM) 모델의 결과를 조합(combination)한 다중 모델 앙상블(multi-model ensemble) 방법으로 DO를 예측하였다. Park *et al.*[2022]은 국립수산과학원 실시간 해양환경 어장정보시스템의 진해만(통영안정) 자료(수심, 표·저층 수온, 염분, 탁도, 엽록소 a, DO 포화도, DO 농도)를 활용하여 장단기메모리(long short-term memory, LSTM) 모델로 표·저층 DO 농도의 시간(1, 6, 12, 24, 36, 48시간) 및 일(1, 3, 6, 10, 15, 30일) 예측을 수행하였고, 의사결정나무(decision tree, DT) 모델로 빈산소 수괴 발생(DO 농도 ≤ 3 mg·L^{-1}) 여부를 일 예측하였다. Kim *et al.*[2024]은 당동만의 층별 1시간 간격 연속관측자료(수온, 염분, DO, 기온, 바람, 강수량)를 사용하여 LSTM, 게이트 순환유닛(gate recurrent unit, GRU), 1차원 합성곱신경망(1dimension-convolutional neural network, 1D-CNN) 모델로 1-72시간 후 저층 DO 농도를 예측하였다.

2. 연구방법

2.1 대상지역

연구지역은「해양환경관리법」제15조 1항에 따라 지정된 마산만 특별관리해역 내측이다(Fig. 1). 특별관리해역은 해양환경기준 유지가 어려운 해역 혹은 해양환경 및 생태계의 보전에 현저한 장애가 있거나 장애를 초래할 우려가 있는 해역을 의미한다. 이 연구에서 진해만, 행암만, 마산만 내측에 위치한 정점의 해양환경측정망 및 어장환경모니터링 자료 중 저층 용존산소 농도가 3 mg·L^{-1} 이하인 자료의 빈도를 비교한 결과 마산만에서 가장 높았다(약 14%). Park[2020]은 10년(2010~2019년)간 우리나라 연안의 해양환경측정망 자료를 분석한 결과 빈산소 수괴는 총 117회 발생했으며, 그 중 약 74%(86회)가 마산만에서 발생했으며, 특히 마산만 내측은 해수 유통이 불량하며 육상 기원 오염원부하의 유입이 많아 빈산소 수괴가 빈번히 발생한다고 보고한 바 있다. 마산만 특별관리해역은 육상 기원

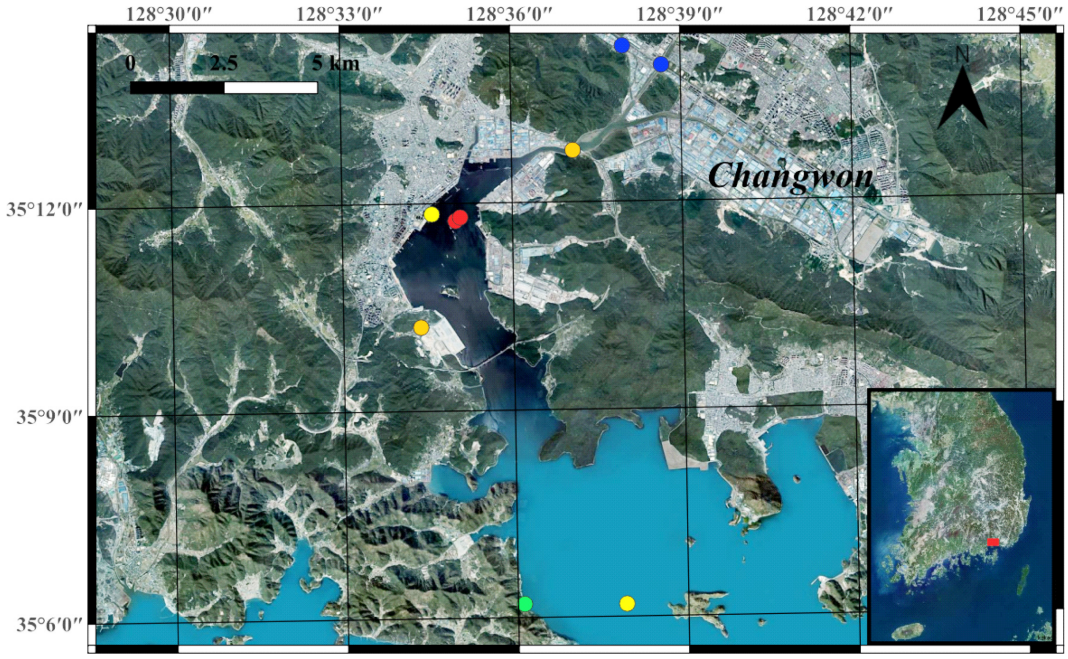

Fig. 1. A map showing the monitoring stations (red: marine environment, blue: stream environment, yellow: hydrodynamic, orange: climate, and lime: effluent) in Masan Bay.

오염부하 관리를 위해 2007년부터 연안오염총량관리를 시행하여 과거에 비해 육상에서 유입되는 오염부하는 감소하였으나, 마산만 내측의 오염물질의 체류시간이 증가하여 외측보다 내측의 집중적으로 관리가 필요하다(Park *et al.*[2018]).

2.2 모델 학습 자료

2.2.1 자료 수집 및 전처리

해양수산부 해양환경정보포털 및 국립수산과학원에서 제공하는 해양환경측정망 및 어장환경모니터링 자료를 해양수질자료로 이용하였다. 모델 학습을 위한 충분한 양의 자료를 확보하기 위해 해양환경측정망 인근 정점(직선거리 : 약 160 m)의 어장환경모니터링 자료도 활용하였다. 해수유동을 고려하기 위해 조위(tidal level), 유속(flow velocity), 유향(flow direction) 자료도 활용하였다. 기상 변화에 의한 영향을 고려하기 위해 풍속(wind speed), 풍향(wind direction), 기온(air temp), 기압(air pressure), 강수량, 일사량 자료를 수집하였다. 육상 기원 오염원의 영향을 고려하기 위해 유입하천들의 월 수질자료와 하수처리시설 일일 방류수 수질자료를 활용하였다. 수집 자료 정보는 Fig. 1과 Table S1에 나타내었다. 최종적으로 수집한 자료에서 해양수질자료 측정시간과 동일한 시간의 자료만을 추출하여 해양수질자료와 결합한 시계열(time-series) 자료를 생성하였다. 내만의 체류시간을 고려하여 60일 누적강우량을 사용하였고 육상 기원 오염부하의 유달 시간지연(time-lag) 효과를 고려하기 위해 유입하천(freshwater, FW)과 하수처리시설 방류수(sewage disposal plant effluent, SDE) 수질자료는 30, 60일 전 자료(FW_COD_30, FW_TN_30, FW_TP_30, FW_COD_60, FW_TN_60, FW_TP_60, SDE_COD_30, SDE_TN_30, SDE_TP_30, SDE_COD_60, SDE_TN_60, SDE_TP_60)도 추가하였다.

결합한 자료는 변수 별 결측치(missing value) 수를 파악하여 결측치를 포함하는 시간의 자료를 제거하기에 앞서 자료의 손실을 최소화하는 방식으로 결측치를 제거하며 변수(variable)를 선택하여 6세트(set)의 변수조합(combination) 표본(sample)을 재구성하였다. Fig. 2는 이 연구의 순서도이다.

2.2.2 자료 증식

이 연구에서 수집한 자료는 시간적으로 연속성을 가지지 않으며 희소성을 가지기 때문에 자료의 편향(bias)를 피하기 위해 모델 학습에 충분한 양의 합성 자료(synthetic data)를 생성하여 자료의 수를 증식(augmentation)시켰다. Patki *et al.*[2016]은 원본(original) 자료를 학습하여 통계적으로 유사한 자료를 합성할 수 있는 생성형(generative) 모델을 제안하였다. Zhang *et al.*[2022]은 이를 발전시켜 순차 자료(sequential data)를 합성할 수 있는 조건형 확률(conditional probabilistic) 자기회귀 신경망 모델을 고안하였다. 이는 행 간 의존성(inter-row dependency)을 유지하면서 시퀀스(sequence) 내 다음 행을 생성하기 위해 필요한 모수(parameter)들의 분포(distribution)를 출력하고 그 분포에서 새로운 행 자료 표본을 추출한다. 이 방법을 통해 앞서 결측치를 제거한 6개의 시계열 자료세트를 동일하게 1300개(행)의 자료로 증식시켰다. 자료 증식 결과, 변수조합에 따른 합성자료의 품질 점수는 변수조합3(81.13) > 변수조합2(78.32) > 변수조합4(74.82) > 변수조합6(74.70) > 변수조합5(73.59) > 변수조합1(71.59) 순이었다. 품질 점수는 실제자료와 합성자료의 분포 유사도(similarity) 및 변수 간 상관관계를 종합한 점수이다. 최종적으로 원 자료의 특성을 재현한 변수조합3 합성자료(해수유동 변수 제외)를 모델 학습에 사용하였다(Table 1).

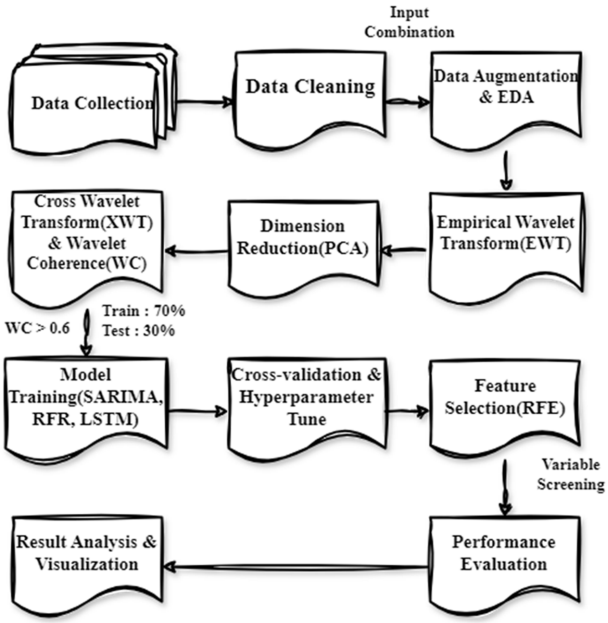

Fig. 2. A schematic flowchart showing outline of this study.

2.2.3 웨이블릿 변환 및 변수 선택

이 연구에서는 신호(signal)의 여러 모드(mode) 혹은 푸리에 스펙트럼(Fourier spectrum)을 추출하여 적응성(adaptive) 웨이블릿을

Table 1. Information on selected variables after data preprocessing and augmentation

Combination No.	Water quality variables	Hydroclimatic variables	Land source variables
3	Date, depth, SD, temp_sur, temp_bot, salinity_sur, salinity_bot, pH_sur, pH_bot, DO_sur, DO_bot, COD_sur, COD_bot, NH_4-N_sur, NH_4-N_bot, NO_2-N_sur, NO_2-N_bot, NO_3-N_sur, NO_3-N_bot, DIN_sur, DIN_bot, TN_sur, TN_bot, DIP_sur, DIP_bot, TP_sur, TP_bot, SiO_2-Si_sur, SiO_2-Si_bot, SS_sur, SS_bot, Chl-a_sur, Chl-a_bot	Wind speed, wind direction, air temp, air pressure, 60d cumulative rainfall	FW_COD_0, FW_TN_0, FW_TP_0, FW_COD_30, FW_TN_30, FW_TP_30, FW_COD_60, FW_TN_60, FW_TP_60, SDE_COD_0, SDE_TN_0, SDE_TP_0, SDE_COD_30, SDE_TN_30, SDE_TP_30, SDE_COD_60, SDE_TN_60, SDE_TP_60

생성하는 경험 웨이블릿 변환(empirical wavelet transform, EWT)을 사용하여 시계열 자료를 분해(decomposition)하였다. 저층 DO 시계열 자료와 같이 비정상성(non-stationarity)을 보유한 자료는 예측하기 어려워 안정적으로 만들기 위해 분해하여 훈련자료로 사용한다(Lui *et al.*[2021]). 기존의 분해 방식은 분해층(decomposition layer) 수를 설정하여야 했으나, EWT에서는 전체 신호 스펙트럼에 포함된 정보를 고려하여 분해층 수를 결정한다. EWT는 Huang *et al.*[1998]이 제안한 경험 모드 분해(empirical mode decomposition, EMD) 알고리즘을 기반으로 필터 뱅크(filter bank)를 적용하여 푸리에 스펙트럼을 분해한다. 푸리에 서포트(Fourier support)는 조각(segment)으로 연속적으로 분해되고 경험 웨이블릿은 개별 조각에 대한 통과대역(passband) 필터로 정의된다. 이 연구에서는 Littlewood-Paley and Meyer의 웨이블릿 구성 방식을 사용하였다(Gilles[2013]).

분해된 개별 변수의 웨이블릿 자료는 주성분분석(principal component analysis, PCA)를 통해 차원 축소(dimension reduction)하여 단일대역 웨이블릿으로 변환하였다. 그리고 저층 DO 웨이블릿과 개별 변수에 대해 웨이블릿의 교차 웨이블릿 변환(cross-wavelet transform, XWT)을 수행했다(Eq. 1). 여기서 W는 웨이블릿, X, Y는 시계열, *는 켤레 복소수(complex conjugation)를 의미한다.

$$W^{XY} = W^{X}W^{Y*} \tag{1}$$

XWT를 통하여 두 시계열 자료의 공통 출력을 찾고 시간-주파수 공간에서 국소 상관계수로 정의된 웨이블릿 긴밀도(wavelet coherence, WC)를 계산하여 학습에 사용한 입력변수(input variable)를 선택하였다(Grinsted *et al.*[2004])(Eq. 2). 여기서 s는 스케일(scale), n(=1,...,N)은 시간, S는 평활 연산자(smoothing operator)이다. 이 연구에서는 WC가 0.6보다 큰 상관관계를 나타내는 입력변수만을 선택하여 모델 학습 자료를 구성하였다.

$$R_n^2(s) = \frac{\left|S(s^{-1}W_n^{XY}(s))\right|^2}{S(s^{-1}\left|W_n^{X}(s)\right|^2)\cdot S(s^{-1}\left|W_n^{Y}(s)\right|^2)} \tag{2}$$

2.3 모델 알고리즘

2.3.1 Seasonal autoregressive integrated moving average(SARIMA)

자기회귀누적이동평균(autoregressive integrated moving average, ARIMA)은 Eq. 3의 자기회귀(autoregressive, AR)와 Eq. 4의 이동평균(moving average, MA)을 선형 조합(linear combination)한 이론으로 비정상성(non-stationarity)을 지닌 시계열(time-series) 자료의 예측 모델에 적합한 알고리즘(algorithm)이다(Box *et al.*[2015]). 여기서 t는 시간, z_t는 시계열(time-series), ϕ는 AR 모수(parameter), θ는 MA 모수, p는 AR 모수의 차수(order), q는 MA 모수의 차수이다. a_t는 백색소음(white noise)이라고 부르는 단위 무작위 정규편차(unit random normal deviates)를 의미한다.

$$z_t = \phi_1 z_{t-1} + \phi_2 z_{t-2} + \cdots + \phi_2 z_{t-p} + a_t \tag{3}$$

$$z_t = a_t + \theta_1 a_{t-1} + \theta_2 a_{t-2} + \cdots + \theta_q a_{t-q} \tag{4}$$

후방 이동 연산자(backward shift or lag operator)를 이용하여 ($Bz_t = z_{t-1}$, $Ba_t = a_{t-1}$) 자기회귀 다항식(polynomial)과 이동평균 다항식을 변형하면 각각 Eq. 5-6과 같다. 여기서 B는 후방 이동 연산자, $\phi(B)$는 자기회귀 연산자(autoregressive operator), $\theta(B)$는 이동평균 연산자(moving average operator)를 의미한다.

$$a_t = (1-\phi_1 B - \cdots - \phi_p B^p) z_t = \phi_p(B) z_t \tag{5}$$

$$z_t = (1-\theta_1 B - \ldots - \theta_p B^q) a_t = \theta_q(B) a_t \tag{6}$$

자기회귀와 이동평균을 결합한 ARIMA는 Eq. 7-8과 같이 나타낸다. 여기서 d는 시계열 z_t의 차분 차수(differencing order)를 의미한다.

$$z_t = \phi_1 z_{t-1} + \phi_2 z_{t-2} + \cdots + \phi_p z_{t-p} + a_t + \theta_1 a_{t-1} + \theta_2 a_{t-2} + \cdots + \theta_t a_{t-q} \tag{7}$$

$$\phi_p(B)(1-B)^d z_t = \theta_q(B) a_t \tag{8}$$

차분 연산자(∇=1−B)를 적용하여 Eq. 9와 같이 나타내고, 차분 $w_t = \nabla^d z_t$를 대입하면 Eq. 9-10과 같다. ∇는 차분 연산자(differencing operator), w_t는 시계열 z_t의 d번째 차분(difference)을 의미한다.

$$\phi_p(B)\nabla^d z_t = \theta_q(B) a_t \tag{9}$$

$$\phi_p(B) w_t = \theta_q(B) a_t \tag{10}$$

하지만, ARIMA는 시계열 추세 차분(trend difference)만 고려하며 자료의 계절성(seasonality) 및 주기성(periodicity)을 반영할 수 없어 계절적 특성을 지닌 저층 DO 농도의 변이 패턴(pattern)을 학습하기 어렵다. 이 연구에서는 이를 고려한 계절성 자기회귀누적이동평균(seasonal ARIMA, SARIMA) 알고리즘을 사용하였다. 계절적 자기회귀 및 이동평균 연산자는 각각 Eq. 11-12와 같이 나타낸다. 여기서 P는 계절 자기회귀(seasonal AR, SAR) 모수의 차수, Q는 계절 이동평균(seasonal MA, SMA) 모수의 차수, S는 계절성 차수이다.

$$\Phi_P(B^S) = 1 - \Phi_1 B - \cdots - \Phi_P B^P \tag{11}$$

$$\Theta_Q(B^S) = 1 - \Theta_1 B - \cdots - \Theta_Q B^Q \tag{12}$$

결론적으로 SARIMA를 Eq. 13와 같이 나타낼 수 있다. 여기서 D는 계절 차분(seasonal difference)의 차수이다.

$$\Phi_P(B^S)\phi_p(B)(1-B)^d(1-B^S)^D z_t = \Theta_Q(B^S)\theta_q(B) a_t \tag{13}$$

이 연구에서는 Box-Jenkins 분석 방법(Durbin and Koopman[2012])을 통해 시계열 자료의 비정상성 및 계절성을 확인하여 계절-추세 분해(season-trend decomposition)를 통해 자료의 안정성을 확보하고 모델의 식별(identification), 추정(estimation), 진단(diagnostic checking) 과정을 반복하여 최종 확정된 모델로 예측(forecasting)을 하였다.

2.3.2 Random forest regression(RFR)

랜덤포레스트(random forest, RF)는 Breiman[2001]이 제안한 의

사결정나무(decision tree, DT) 기반 앙상블(ensemble) 기계학습 알고리즘이다. 기존의 앙상블 방법인 배깅(bagging)과 랜덤 서브스페이스(random subspace)를 통해 훈련자료(training datasets) 중 무작위로 선택된 다수의 표본(sample)을 학습한 개별 의사결정나무의 출력(output)을 평균하여 단일 훈련자료를 학습한 의사결정나무 보다 나무들 간의 상관성(correlation)을 줄이고 모델의 편향(bias)과 분산(variance)을 감소시킬 수 있다.

2.3.3 Long short-term memory(LSTM)

장단기메모리(long short-term memory, LSTM)는 기존 시간 진행 역전사(back-propagation through time, BPTT) 방식이 순환신경망(recurrent neural network)의 가중치 조정(update) 시 기울기 소실 및 폭주(gradient vanishing and exploding)를 발생시킬 수 있는 문제점을 보완하기 위해 Hochreiter and Schmidhuber[1997]에 의해 고안되었다. Eq. 14-15에서 t는 시간, y는 활성화(activation), f는 미분가능한(differentiable) 활성화 함수(activation function), j는 비출력 유닛(non-output unit), net는 유닛의 현재 순(net) 입·출력을 의미한다. in_j는 j번째 입력 게이트(input gate) 유닛, out_j는 j번째 출력 게이트(output gate) 유닛, c_j는 입·출력 게이트로부터 입력을 받는 j번째 메모리셀(memory cell) 유닛, u는 임의의 유닛이다. w는 유닛 간 연결(connection)에 대한 가중치(weight)이다.

$$y^{out_j}(t) = f_{out_j}(net_{out_j}(t)) \tag{14}$$

$$y^{in_j}(t) = f_{in_j}(net_{in_j}(t)) \tag{15}$$

where $net_{out_j}(t) = \sum_u w_{out_j u} y^u(t-1)$

$$net_{in_j}(t) = \sum_u w_{in_j u} y^u(t-1)$$

$$net_{c_j}(t) = \sum_u w_{c_j u} y^u(t-1)$$

메모리셀의 활성화는 Eq. 16과 같이 계산된다. 여기서 s_{c_j}(t)는 내부 상태(internal state), g는 메모리셀 유닛의 현재 순 입·출력을 결합하는 미분가능한 함수, h는 내부 상태에서 계산된 메모리셀 출력을 정규화(scaling)하는 미분가능한 함수이다. 순 입출력은 메모리셀 유닛 내 정보의 보존 혹은 기각 시점을 결정하기 위해 입력 게이트 유닛을 사용하고, 메모리셀 유닛에 접근하고 메모리셀에 의해 교란되는 것을 방지할 시점을 결정하기 위해 출력 게이트 유닛을 사용한다. 메모리셀의 일정오차순환(constant error carousel, CEC)내 오차 신호(error signals)는 변하지 않지만, 출력 게이트를 통해 다른 시간의 셀로 전이되는 오차 신호가 추가된다. 출력 게이트는 오차를 적절히 정규화하여 어떠한 오차를 CEC에 보존할지 학습한다. 입력 게이트는 다시 오차를 정규화하여 오차를 기각할 시점을 학습한다. LSTM은 저층 DO와 같은 장기적 의존성을 가지는 시계열 자료의 모델링에 적합하다. 이 연구에서는 입력자료 시계열의 순방향과 역방향 정보를 모두 고려하는 양방향(bidirectional) LSTM으로 신경망을 확장하였고 Multi-head Attention 층을 추가하여 시퀀스의 중요한 정보에 가중 학습을 하도록 하였다(Fig. S1).

$$y^{c_j}(t) = y^{out_j}(t) h(s_{c_j}(t)) \tag{16}$$

where $s_{c_j} = 0, s_{c_j}(t) = s_{c_j}(t-1) + y^{in_j}(t) g(\neq t_{c_j}(t))$ for $t > 0$

2.4 모델 학습

이 연구에서는 ARIMA 및 SARIMA 기반 모델의 입력자료는 연속된 시계열 자료가 요구되므로 저층 DO 농도 자료에서 같은 일자의 값들은 평균값으로 교체(replacement)하고 결측 일자의 값은 최근 값으로 대체(imputation)하여 일단위 자료를 생성한다. 또한 모델의 복잡도 감소와 추세 및 잔차(잡음) 성분을 학습하기 유리하도록 Loess를 사용한 계절추세분해(Seasonal trend decomposition using Loess, STL) 기법을 통해 자료의 계절성을 배제(deseaonalization)하고 학습하였다. 하지만, STL 기법은 다중중첩 또는 비선형 계절성을 가지는 자료의 계절성분을 완전히 제거하지 못하는 한계를 가지며 이를 확인하기 위해 잔차성분에 계절성분의 잔류여부를 자기상관함수 도표(autocorrelation function plot)와 Ljung-Box Test를 통해 확인한 결과, 계절성이 약화되었지만 완전히 배제되지 못하여 계절성을 고려한 SARIMA를 모델 학습에 사용하였다. 타깃자료인 저층 DO 농도 시계열 일평균 자료만을 학습하여 평가자료에 대한 예측을 수행하였다. 이후 예측값에 제거한 계절성분을 재결합하였다. RFR, LSTM 모델은 최종적으로 선택된 입력변수에 대해서는 경험 웨이블릿의 차원 축소 자료를 1) 저층 DO 농도 자료와 2) DO 농도 시계열의 경험 웨이블릿 자료와 각각 결합하여 단변량(uni-variate) 및 다변량(multi-variate) 모델 학습을 개별적으로 수행하였다. 전처리(preprocessing)된 자료의 학습을 통해 10, 30, 60, 90 시점(time step)의 저층 DO 농도를 예측하여 성능을 평가하였다. LSTM의 경우 시퀀스의 길이(sequence length)를 10으로 설정하였다. 해당 자료는 표준 정규화(standard scaling)를 통해 개별 변수의 값들이 정규분포(normal distribution)를 갖도록 변환하였다. 개별 시계열 자료 중 70%는 훈련자료(train dataset), 나머지는 평가자료(test dataset)로 분리(split)하여 모델 학습에 사용하였다. SARIMA, RFR, LSTM 모델 학습은 각각 Python(Ver. 3.10.4) 프로그래밍 언어환경의 statsmodels(Ver. 0.13.2), scikit-learn(Ver. 1.2.2), TensorFlow(Ver. 2.8.0) 모듈(module) 및 라이브러리(library)를 이용하여 수행하였다.

2.5 교차검증 및 하이퍼파라미터 튜닝

이 연구에서는 RFR, LSTM의 하이퍼파라미터를 탐색하기 위해 RandomizedSearchCV라는 교차검증(cross validation, CV) 기반 하이퍼파라미터 튜닝(tuning) 방식을 사용하였다. 각 모델 별로 하이퍼파라미터 격자(grid)를 구성하고 교차검증 시 자료 분리 수(fold)는 3으로 설정하였다. SARIMA 모델의 하이퍼파라미터 중 D는 Osborn, Chui, Smith, and Birchenhall(OCSB) Test를 통해 결정하였고 P, Q는 순차적 방식(stepwise approach)을 통해 Akaike information criterion (AIC)를 최소화하는 파라미터(모수)를 선정(Hyndman and Khandakar,

2008)하였으며 자료의 계절성에 따라 계절주기(seasonal period, *S*)를 직접 설정해주었다. RFR 모델에서는 n_estimators(나무의 수), max_depth(나무의 최대 깊이), min_samples_split(내부 노드를 분리하기 위한 최소 표본의 수), mean_samples_leaf(잎 노드를 분리하기 위한 최소 표본의 수), bootstrap(표본 무작위 복원추출)으로 그리드를 구성하였다. LSTM 모델의 그리드는 activation(활성화 함수), optimizer(최적화방법), dropout_rate(입력유닛 누락 비율), leaning_rate (학습률), epochs(전체 자료에 대한 학습 수), batch_size(1 epochs에 사용되는 표본 수)로 구성하였다.

2.6 변수 중요도

이 연구에서는 앞서 WC를 비교하여 선택한 변수에서 RFR 모델 학습 기반 재귀 피쳐 제거(recursive feature elimination, RFE) 방식으로 변수를 추가 스크리닝(screening) 하였다. 선택하려는 변수의 수를 설정하고 설정한 변수의 수가 남을 때까지 모델 학습을 반복하여 변수의 중요도가 낮은(모델 성능에 미치는 영향이 작은) 변수부터 하나씩 제거하여 변수를 추출한다. 여기서는 입력변수의 수를 4-6개 범위로 설정하여 선택한 변수조합의 자료를 RFR 모델 학습에 사용하였다. LSTM 모델의 경우 4-6개 입력변수를 가지는 모든 조합의 자료로 학습한 모델의 예측성능을 비교하여 변수조합별 최종 예측변수(predictor)를 결정하였다.

2.7 모델 성능 평가

이 연구에서는 평가자료에 대한 모델의 예측성능을 평가하기 위해 결정계수(adjusted coefficient of determination, $R^2_{adjusted}$), 스피어만 서열 상관계수(Spearman's rank correlation coefficient, SRCC), 평균제곱근오차(root mean square error, RMSE), 평균절대오차(mean absolute error, MAE)를 평가 지표(evaluation metrics)로 사용하였다. 개별 평가 지표에 대한 계산은 Eq. 17-20와 같다. 여기서 N은 평가자료의 수, y는 실제값(true value), $\hat{y}_i$ 는 예측값(predicted value), $\bar{y}$ 는 타깃(target)값의 평균, $\bar{\hat{y}}$는 예측값의 평균, d_i는 실제값과 예측값의 서열 차, k는 예측변수의 수, R^2는 실제값과 예측값의 결정계수이다.

$$R^2_{adjusted} = 1 - \frac{(1-R^2)(N-1)}{N-k-1} \tag{17}$$

$$SRCC = 1 - \frac{6 \times \sum_{i=1}^{N} d_i^2}{N(N^2-1)} \tag{18}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \tag{19}$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \tag{20}$$

3. 결과 및 고찰

3.1 결과

3.1.1 자료 증식 및 변수 선택

최종 선택된 변수조합 기반 합성자료 기초 통계량(statistics)과 입력변수들과 저층 DO의 상관계수(CC)는 Table 2와 같다.

WC를 비교하였을 때, 기상 변수보다 수질 변수가 저층 DO와 상관성이 좋았고 특히, COD는 저층 DO와 다른 변수보다 좋은 교차상관성을 보였다(Table 2). 예측 시점(10, 30, 60, 90) 별 RFE 방식 변수중요도 계산을 통해 추가로 선택된 변수(4-6개)는 Table 3과 같다. 단변량 자료 변수선택에서는 60 시점 후 예측을 제외하고 모두 저층 암모니아성질소(NH_4-N)를 공통적으로 제거하였다. 또한 예측 시간이 길어질수록 하천영향 변수(FW_TP_60)의 중요도가 낮아졌고 COD의 경우, 해양 기원 COD보다 육상 기원 COD의 중요도가 높았다. 다변량 자료 변수선택에서는 장기 예측 시 하수처리시설 방류수 COD(SDE_COD_0)와 60일전 TP(SDE_TP_60)를 제거하였다. 단기 예측 시 육상 기원 인자가 중요하였으나, 장기 예측 시에는 해양 기원 인자가 중요하였다. LSTM 모델의 경우, 시행착오법(trial and error)을 통해 하천영향 변수(FW_TP_60), 하수처리수 방류수 COD 및 60일전 TP, 저층 NH_4-N 변수조합에서 성능이 가장 우수하여 모델 학습자료 변수로 선택하였다.

3.1.2 최적 하이퍼파라미터 및 모델 성능

이 연구에는 3개의 알고리즘 기반 모델로 마산만 내측의 저층

Table 2. The statistics and correlation coefficients(CC) with bottom DO for synthetic datasets

Statistics	FW_TP_60 [mg·L^{-1}]	SDE_COD_0 [mg·L^{-1}]	SDE_TP_60 [mg·L^{-1}]	COD_sur [mg·L^{-1}]	NH_4-N_bot [μg·L^{-1}]	SiO_2-Si_sur [μg·L^{-1}]	DO_bot [mg·L^{-1}]
Mean	0.150	13.43	0.866	2.85	61.02	530.91	7.87
Std	0.049	2.30	0.260	0.69	34.07	232.48	1.59
Min	0.031	6.40	0.142	1.08	0.25	9.00	1.40
25%	0.123	12.51	0.776	2.49	41.36	409.96	7.09
50% (median)	0.154	13.78	0.888	2.87	60.80	522.11	7.70
75%	0.178	14.36	0.990	3.17	75.92	644.90	8.58
Max	0.346	28.80	2.503	6.08	228.45	1432.85	12.42
CC	-0.039	-0.115	-0.128	0.105	-0.060	-0.086	1.000
WC[a]	0.633	0.726	0.667	0.686	0.665	0.606	1.000

[a]WC indicates the coherence(cross-correlation) between each wavelet of the selected input variable and the bottom DO wavelet.

Table 3. The input variable combinations selected by using RFE

Variate	No. of variable	Time step	Eliminated variables
Uni-	4	10	SDE_COD_0, COD_sur, NH_4-N_bot
		30	FW_TP_60, NH_4-N_bot, SiO_2-Si_sur
		60	SDE_TP_60, SiO_2-Si_sur, DO_bot
		90	SDE_COD_0, COD_sur, NH_4-N_bot
	5	10	SDE_COD_0, NH_4-N_bot
		30	NH_4-N_bot, SiO_2-Si_sur
		60	SDE_TP_60, DO_bot
		90	COD_sur, NH_4-N_bot
	6	10	NH_4-N_bot
		30	NH_4-N_bot
		60	SDE_TP_60
		90	COD_sur
Multi-	4	10	NH_4-N_bot, SiO_2-Si_sur
		30	COD_sur, SiO_2-Si_sur
		60	SDE_COD_0, SDE_TP_60
		90	SDE_COD_0, SDE_TP_60
	5	10	NH_4-N_bot
		30	SiO_2-Si_sur
		60	SDE_TP_60
		90	SDE_COD_0

DO를 예측하였다. 알고리즘 별 최적 모델의 훈련자료, 하이퍼파라미터, 모델성능 평가지표를 Table 4에 나타내었다. 이 연구에서는 예측 결과의 범위를 조정하기 위하여 웨이블릿 형태의 다변량 예측값에 훈련자료 통계량 기반 다변량 동적 규모 조정(scaling)을 수행한 후 실제값과 비교하여 평가하였다.

SARIMA 모델에서 모델 학습 결과가 가장 좋았다. 저층 DO 농도가 급격하게 증가하거나 감소하는 첨두(peak) 구간에서는 큰 예측 오차를 보였다(Fig. 3). 상관계수(SRCC)는 저층 DO의 계절적 특성을 파악하는 성능 지표로 간주할 수 있으며, SARIMA 모델의 경우 자료의 계절성을 제거하고 학습하여 상관관계가 다른 모델보다 좋았다(Table 4). 또한 계절 파라미터 중 계절 이동평균항(Q)이 0으로 수렴하였는데 이는 과거 저층 DO 계절적 예측 오차가 현재 시계열 값 예측에 영향을 미치지 않음을 의미한다. 이는 자료가 계절성을 가지나 과거의 계절적 오차를 고려하지 않는 것이 모델 성능에 더 유리하여 계절 이동평균 모델이 단순화되었다. 오히려 계절 자기회귀항(P)에서 계절성을 내포하여 과거의 계절값이 현재 시계열 값 예측에 선형적 영향을 주었다.

RFR 모델은 단변량 보다 다변량 예측이 더 정확하였으며 단변량 예측의 경우, 예측 시간 60 시점에서 가장 성능이 우수하였고 예측변수의 수가 많을수록 예측성능이 우수하였다. 다변량 예측에서도 60 시점에서 가장 성능이 우수하였으며, 예측변수의 수가 5개 일 때 가장 우수한 성능을 보였다. 결과적으로 최대 성능을 보인 모델에서 학습한 자료의 입력변수는 60일전 하천 TP, 하수처리시설 방류수 COD, 표층 COD, 저층 NH_4-N, 표층 SiO_2-Si였고, 60일 전 하수처리시설 방류수 TP는 제외되었다.

LSTM 모델도 단변량 보다 다변량 자료의 예측에 적합하였다. 예측변수의 수가 5개이면서 예측 시간이 60 시점일 때 모델의 성능이 가장 우수하였다. 앞선 두 모델과 유사하게 저층 DO 농도가 3 $mg \cdot L^{-1}$ 이하인 구간에서는 예측 정확도가 낮았다(Fig. 3). 최종적으로 학습에 사용된 변수는 RFR 모델에서 선택된 변수조합과 동일하였다.

3.2 고찰

RFR, LSTM 모델에서 공통적으로 60일 전 하수처리시설 방류수 TP가 모델 학습에 영향을 주지 못하였지만 해양수질자료 중 표층 COD, 저층 NH_4-N, 표층 SiO_2-Si, 육상 기원 인자는 60일 전 하천 TP, 하수처리시설 방류수 COD가 모델 학습에 중요한 변수였다. 하수처리시설에서 유입되는 점오염원 기인 TP 보다는 하천에서 유입되는 비점오염원 기인 TP가 해역의 저층 DO 예측에 더 큰 영향을 주는 것으로 보인다. 이외 DO 농도에 영향을 주는 수온, 수문기상학적 요인들은 저층 DO 웨이블릿과 교차 상관성이 낮아 이 연구의 학습에서는 배제되었다.

DO를 웨이블릿으로 분해하여 다변량 자료를 예측할 경우, 웨이블릿 분해 특성상 국부적으로 비정상성을 가지는 농도 변화에 민감하였으나, 빈산소 수괴와 같은 급변하는 저층 DO 농도 변화를 예측하지 못하였다. 또한 학습자료 생성을 위한 변환 과정(웨이블릿 변화 및 차원 축소)에서 원 자료의 정보가 많이 손실되었다. 자료변환을 통해 자료의 정상성(stationarity)은 가지나 분해되어 저층 DO 예측값의 진폭(amplitude) 범위가 약화(attenuation)되었으며 농도가 극적으로 변화는 시점에서는 예측에 어려움이 있었다. 이는

Table 4. Performance metrics and learning requirements for the best predictive models in testing phase

Model	No. of variable	Training dataset	Time step	Hyperparameter	$R^2_{adjusted}$	SRCC	RMSE	MAE
SARIMA	1	Monthly deseasonalized DO_bot series	1	P=2, D=0, Q=0, S=7	-0.375	0.233	1.945	1.416
RFR	5	Multi-variate series	60	n_estimators=400, max_depth=30, min_samples_split=15, min_sampled_leaf=16, bootstrap=True	-1.209	-0.078	2.362	1.715
LSTM	5	Multi-variate series	60	activation='tanh', optimizer='RMSprop', dropout_rate=0.3, learning_rate=0.001, epochs=100, batch_size=256	-0.829	0.106	2.147	1.592

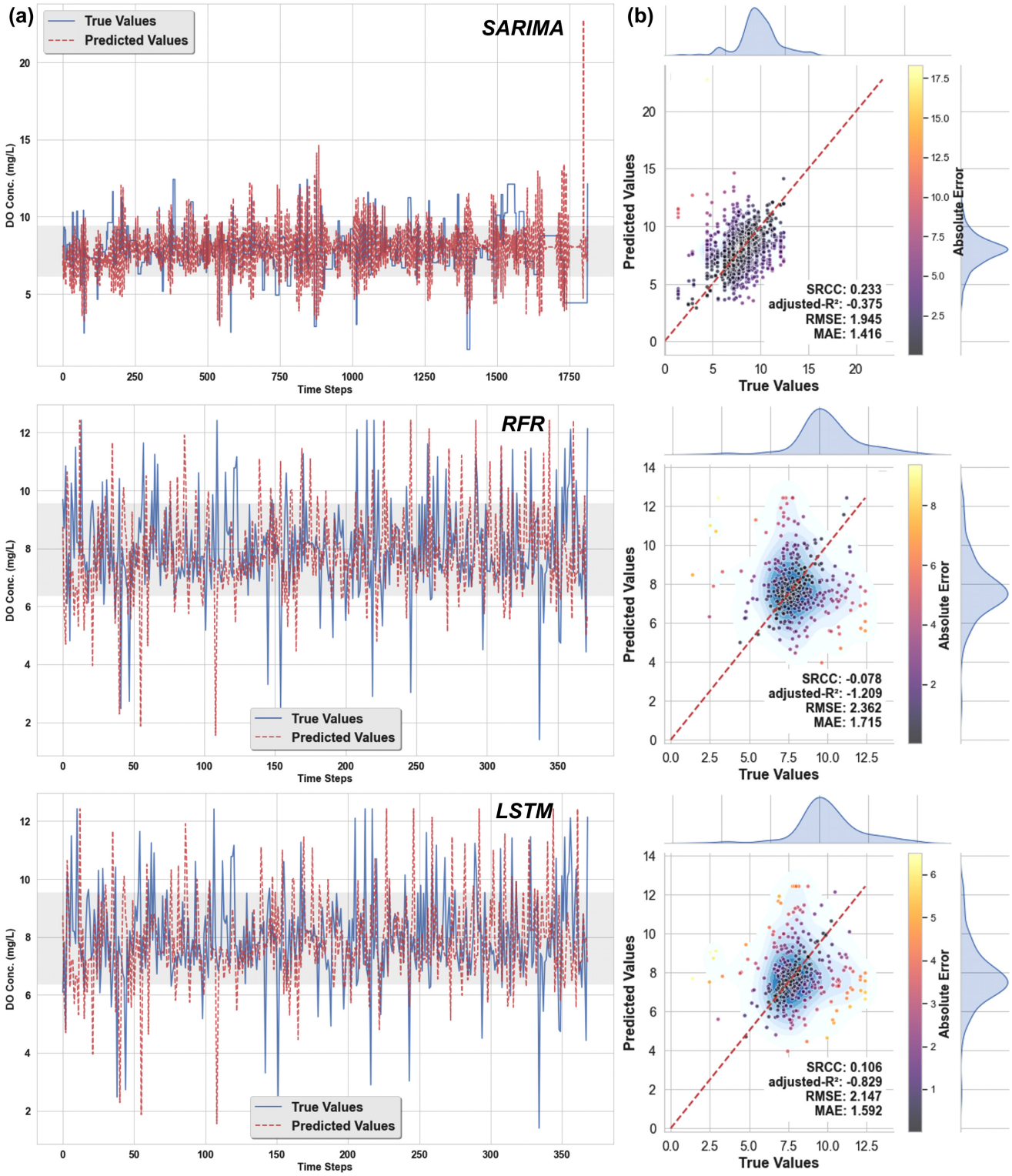

Fig. 3. Plots((a) time series, (b) scatter) for the best fitting results in model performance during testing phase.

원 자료에 대한 웨이블릿 변환 및 주성분 자료의 신호-잡음비(signal-to-noise ratio, SNR) 계산을 통해 확인할 수 있었다(Fig. S2). Lim *et al.*[2021]은 웨이블릿 영역(wavelet domain) PCA를 통해 비정상성을 지닌 다변량 시계열 자료의 시간-규모 의존 정보를 확인하였으며, 이와 같은 기법은 주로 시간에 따라 패턴이 변하는 신호의 오류 탐지(fault diagnosis)에 사용되었다. 시간에 따른 웨이블릿 스펙트럼의 국부적인 요소(local wavelet spectrum)만 포착하여 정상성을 유지하였고 일시적이고 급변하는 신호 패턴을 인식할 수 있

었다. 하지만, 이 연구에서는 시계열 전 기간을 분해하고 차원을 축소한 후 학습에 사용하여 패턴을 파악하기 어려웠다.

SARIMA 모델은 학습 자료가 일정한 추세와 계절성을 가지고 있다고 가정하고 있기에 시계열 자료의 급격한 변화와 같은 이상치(outlier)에 민감한 특성을 가지고 있다. 또한, SARIMA 모델은 다른 변수에 독립적이지 못한 비선형성(non-linearity)을 가지는 복잡한 자료의 학습에는 부적합하다. Coopersmith *et al.*[2011]은 ARMA과 같은 자기회귀 모델의 경우 1-2시간 보다 긴 시간의 자료를 예측할 때는 과거 자료와 현재 자료의 상관관계가 점점 약해져 낮은 성능을 보인다고 보고했다. Ross and Stock[2019]은 DO는 자기상관관계가 좋지 않아 자기회귀 모델에 부적합하며 의사결정나무 기반 모델을 추천하였다. 즉, DO는 시간에 따른 자기상관관계가 낮아 변동이 극심한 변수로 간주된다. 이 연구에서 SARIMA 모델의 예측성능은 다른 모델에 비해 좋았으나, 저층 DO 이외의 변수들까지 개별적으로 학습하고 예측한 결과들을 결합하여 변수들 간의 의존성(dependency)을 고려한 다변량 예측 모델을 구성하거나 다변량 예측 모형(e.g., 벡터오차수정모형)을 사용하여 성능을 개선하는 방법도 고려할 수 있다.

RFR 모델의 예측값은 급격한 DO 변화에서는 일정 시간 간격 뒤로 밀리는(lagged) 경향을 보이는데 이는 이전 연구(Park *et al.*[2022])에서 유사한 결과를 보였고 이는 DO 농도의 비정상성과 복잡성을 원인으로 지적한 바 있다(Fig. 3). 마산만 내측과 같은 반폐쇄성 해역의 수질 자료는 외해와 다르게 강한 비선형성을 보여 패턴을 학습하기 어렵다(Valera *et al.*[2020]). 또한 RFR 모델은 SARIMA, LSTM 모델과 달리 구조상 시간의존성을 가지는 자료에 대해서 과거 정보를 미래 예측에 반영하지 못하는 한계를 가진다.

LSTM 모델에서도 RFR 모델과 같이 단기적인 예측보다 장기적인 예측에서 성능이 우수하였고 예측값이 일정 시간 뒤로 밀리는 경향을 보였다. Kim *et al.*[2024]은 DO와 같이 단주기 동안 변동이 큰 자료의 경우, 예측 시간(lead time)이 길어질수록 LSTM 모델의 예측 성능이 감소하여 학습속도가 빠른 게이트순환유닛(gate recurrent unit, GRU), 1차원 합성곱신경망(1dimension-convolutional neural network, 1D-CNN)이 더 적합하다고 제안하였다. 또한, Park *et al.*[2022]은 예측 시간이 길어질수록 DO 농도의 시계열 변화 특성이 더 강하게 반영되어 LSTM 모델의 입력자료에서 더 긴 시퀀스 길이를 요구한다고 보고하였다.

모든 알고리즘에서 공통적으로 저층 DO가 급격히 변하는 구간에서 예측이 어려웠다. 이는 이전 연구결과에서도 유사한 경향을 보였다(Zhi *et al.*[2021]). 이렇게 수문기상학적 요인, 유역특성, 생지화학적 반응 등이 복합적으로 DO에 영향을 미치므로 DO가 불규칙적으로 변화하는 시기가 발생하고, 이러한 변동성은 DO 자료의 수보다도 모델 성능에 큰 영향을 미친다. 그러므로 DO 자료보다 생지화학 반응(호흡, 광합성 등)에 관련된 변수들을 추가하면 모델의 성능을 보완할 수 있다고 판단된다. 또한 Muller and Muller[2015]는 육상에서 기원되는 비점오염원을 모델에 반영하면 성능이 더 개선될 것이라고 추정하였다. 반면, DO 변화 패턴을 정확하게 학습하기 위해 급격한 변화를 나타나는 DO 자료(특히, 3 $mg \cdot L^{-1}$ 이하)가 충분히 필요하나, 특정 해역의 자료만으로는 부족한 경우가 많다. 게다가 측정 빈도가 적은(intermittent) 해양환경측정망 및 어장환경 모니터링 자료만을 활용하여 저층 DO 농도 자료를 확보하고 자료증식 방법을 적용하였으나, 이는 실제 관측 자료와 차이가 있을 수 있으며 고해상도의 연속관측자료를 확보하지 못한 한계가 있었다. 연속관측자료의 경우, 주기성을 가지며 점진적으로 변화하여 시계열 자료 간의 정보 전이가 용이하나 이 연구에서 합성한 자료는 짧은 주기의 변동성을 보여 패턴을 학습하기 어려웠다. 자료 변환 방법에서도 웨이블릿 분해 자료의 차원 축소 과정 중 손실되는 정보로 인해 모델의 민감도가 감소하는 문제를 개선하기 위해 다양한 차원 축소 방법(e.g., 오토인코더, 동적시간워핑, 국부 정상성 웨이블릿)을 사용할 수 있다. 모델의 예측성능 개선을 위해 Coopersmith *et al.*[2011], Ross and Stock[2019]는 해수유동과 같은 역학적 수치모델(numerical model)과 기계학습(machine learning) 알고리즘을 결합하여 빈산소 수괴 발생 메커니즘을 더욱 정밀하게 해석할 수 있다고 제안하였다. 또한 여러 알고리즘 기반 학습 모델의 출력값을 평균하여 예측을 하는 앙상블 모델은 단일 알고리즘 모델의 불확실성을 해소할 수도 있다(Abba *et al.*[2020]; Yu *et al.*[2020]; Lui *et al.*[2021]). 이 연구에서도 시간과 변수 의존성을 모두 고려한 예측을 할 수 있도록 모델 학습 구조에 대한 추가적인 개선이 필요하다.

4. 결론 및 제언

이 연구에서는 해양수질자료, 기상자료, 육상 기원 오염원 자료를 결합하고 기계학습 기반 모델을 통해 자료의 수를 증식하였다. 자료의 전처리를 위해 EWT을 적용하였고 변수 선택을 위해 XWT, WC, RFE를 사용하였다. SARIMA, RFR, LSTM 기반 모델로 자료를 학습한 결과 저층 DO 변화가 큰 구간(regime shift)에서는 예측성능이 낮았다. 또한 기상 및 육상 기원 오염원 보다는 해양수질 및 생지화학적 특성이 저층 DO 변화와 모델 성능에 큰 영향을 미치는 것으로 판단하였다. 균형 있는 분포를 가진 학습 자료와 빈산소 발생에 영향을 미치는 중요 변수를 자료에 추가하고 자료변환 중 손실을 보완한다면 예측 성능을 개선할 수 있을 것이다(Muller and Muller[2015]). 나아가 이 연구를 발전시켜 저층 DO의 예측값과 실제값을 비교하여 영양염류 변화를 간접적으로 예측할 수 있으며 이는 효율적인 연안환경관리를 위하여 도입한 연안오염총량관리의 오염부하량 삭감 및 규제에 대한 효과를 검증하는 용도로도 활용할 수 있다. 또한, 인공지능 모델을 이용한 빈산소 수괴 발생 예측이 양식장과 같은 경제적 손실을 가져올 수 있는 산업에 활용할 수 있으며, 해양환경 복원 및 규제 정책을 지원할 수 있다(Liu *et al.*[2021]).

후 기

이 연구는 한국해양과학기술원 “남해·제주 연안 해양 환경변화 관리 시스템 개발(PEA0206)”의 지원을 받아 수행된 연구입니다.

이 논문을 면밀히 심사를 하여 개선될 수 있게 해주신 심사위원님께 감사를 드립니다.

References

[1] Abba, S.I., Linh, N.T.T., Abdullahi, J., Ali, S.I.A., Pham, Q.B., Abdulkadir, R.A., Costache, R., Nam, V.T. and Anh, D.T., 2020, Hybrid machine learning ensemble techniques for modeling dissolved oxygen concentration, IEEE, 8, 157218-157237.

[2] Box, G.E.P., Jenkins, G.M., Reinsel, G.C. and Ljung, G.M., 2015, Time series analysis: Forecasting and control, 5th Edition, Wiley.

[3] Breiman, L., 2001, Random forests, Mach. Learn., 45, 5-32.

[4] 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).

[5] Coopersmith, E.V., Minsker, B. and Montagna, P., 2011, Understanding and forecasting hypoxia using machine learning algorithms, J. Hydroinform., 13(1), 64-80.

[6] Diaz, R.J. and Rosenberg, R., 2008, Spreading dead zones and consequences for marine ecosystems, Science, 321, 926-929.

[7] Durbin, J. and Koopman, S.J., 2012, Time series analysis by state space methods, 2nd edition, Oxford Univ. Press, Chapter 8, 198-199.

[8] Gilles, J., 2013, Empirical wavelet transform, IEEE Trans. Signal Process., 61(16), 3999-4010.

[9] Grinsted, A., Moore, J.C. and Jevrejeva, S., 2004, Application of the cross wavelet transform and wavelet coherence to geophysical time series, Nonlinear Process. Geophys., 11, 561-566.

[10] Hyndman, R.J. and Khandakar, Y., 2008, Automatic time series forecasting: The forecast package for R, Journal of Statistical Software, 27(3).

[11] Hochreiter, S. and Schmidhuber, J., 1997, Long short-term memory, Neural Comput., 9(8), 1735-1780.

[12] Huang, N.E., Shen, Z., Long, S.R., Wu, M.C., Shih, H.H., Zheng, Q., Yen, N.C., Tung, C.C. and Liu, H.H., 1998, The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis, in Proc. of Math. Phys. Eng. Sci., 454(1971), 909-995.

[13] Kim, Y.M., Park, S.E., Kim, M.H., Bak, S.H., Kim, C.S., Kim, J.K., and Jang., S.W., 2024, Prediction of the temporal variations in bottom dissolved oxygen using deep learning and continuous observation data of the marine environment, J. Korean Soc. Mar. Environ. Energy, 27(2), 131-145.

[14] Lim, Y.J., Kwon, J.H., and Oh, H.S., 2021, Principal component analysis in the wavelet domain, Pattern Recognition, 119, 108096.

[15] Liu, H., Yang, R., Duan, Z. and Wu, H., 2021, A hybrid neural network model for marine dissolved oxygen concentrations time-series forecasting based on multi-factor analysis and a multi-model ensemble, Engineering, 7, 1751-1765.

[16] Muller, A.C. and Muller, D.L., 2015, Forecasting future estuarine hypoxia using a wavelet based neural network model, Ocean Model., 96, 314-323.

[17] NIFS, 2009, Hypoxia in the coast of Korea.

[18] Park, M.O., Lee, Y.W., Park, J.K., Kim, S.G., Kim, S.S. and Lee, S.M., 2018, Changes in water quality in Masan Bay after the introduction of the total pollution load management system, J. Korean Soc. Mar. Environ. Energy, 21(2), 139-148.

[19] Park, M.O., 2020, Spatiotemporal variation of oxygen deficient water mass in the enclosed bay, Korea, Ph.D. dissertation, Interdiscip. Program of Ocean Ind. Eng., Pukyong Natl. Univ., Busan, S. Korea.

[20] Park, S.S., Kim, B.K., and Kim, K.H., 2022, Prediction in dissolved oxygen concentration and occurrence of hypoxia water mass in Jinhae Bay based on machine learning model, J. Korean Soc. Coast. Ocean Eng., 34(3), 47-57.

[21] Patki, N., Wedge, R. and Veeramachaneni, K., 2016, The synthetic data vault, in Proc. of 2016 IEEE Int. Conf. on Data Sci. Adv. Anal. (DSAA), Montreal, QC, Canada, 399-410.

[22] Ross, A.C. and Stock, C.A., 2019, An assessment of the predictability of column minimum dissolved oxygen concentrations in Chesapeake Bay using a machine learning model, Estuar. Coast. Shelf Sci., 221, 53-65.

[23] Valera, M., Walter, R.K., Bailey, B.A. and Castillo, J.E., 2020, Machine learning based predictions of dissolved oxygen in a small coastal embayment, J. Mar. Sci. Eng., 8, 1007.

[24] Yoo, Y.J. and Kim, S.J., 2019, Analysis of the characteristics of water quality difference occurring between high tide and low tide in Masan Bay, J. Wetl. Res., 21(2), 102-113.

[25] Yu, X., Shen, J. and Du, J., 2020, A machine-learning-based model for water quality in coastal waters, taking dissolved oxygen and hypoxia in Chesapeake Bay as an example, Water Resour. Res., 56(9).

[26] Zhang, J., Gilbert, D., Gooday, A.J, Levin, L., Naqvi, S.W.A., Middelburg, J.J., Scranton, M., Ekau, W., Peña, A., Dewitte, B., Oguz, T., Monteiro, P.M.S., Urban, E., Rabalais, N.N., Ittekkot, V., Kemp, W.M., Ulloa, O., Elmgren, R, Escobar-Briones, E. and Van der Plas, A.K., 2010, Natural and human-induced hypoxia and consequences for coastal areas : synthesis and future development, Biogeosciences, 7, 1443-1467.

[27] Zhang, K., Veeramachaneni, K. and Patki, N., 2022, Sequential models in the synthetic data vault, arXiv, 2207.14406v1.

[28] Zhi, W., Feng, D., Tsai, W.P., Sterle, G., Harpold, A., Shen, C. and Li, L., 2021, From hydrometeorology to river water quality : Can a deep learning model predict dissolved oxygen at the continental scale?, Environ. Sci. Technol., 55, 2357-2368.

Received 11 September 2023
1st Revised 3 November 2023, 2nd Revised 21 August 2024
Accepted 22 August 2024

Appendix

Table S1. Details on the collected data in this study

Sort	Source	Variable	Station	Year	Interval	Transformation
Marine environment	Marine Environment Monitoring Network	Date, time, depth, SD, temp., salinity, pH, DO, COD, NH_4-N, NO_2-N, NO_3-N, DIN, TN, DIP, TP, SiO_2-Si, SS, Chl-a (at surface and bottom layers)	Masan Bay1	1997-2022	4 times/yr (quarterly)	
	Fishery Environment Information System		Masan Bay6	2013, 2015-2018	6 times/yr (bimonthly)	
Freshwater environment	Water Environment Information System	Date, COD, TN, TP	Naedong, Changwon	2005-2022	12 times/ (monthly)	Averaging values from both streams
Hydrodynamic	Ocean Data in Grid Framework	Tidal level	Masan	2012-2022	525,600 times/yr (minutely)	
		Flow rate, flow direction	Masan Harbor	2015-2022	17,520 times/yr (every 30 minutes)	
Climate	Ocean Data in Grid Framework	Wind speed, wind direction, air temp., air pressure	Masan	2011-2022	525,600 times/yr (minutely)	
	Automated Synoptic Observing System	Daily rainfall	Changwon	2011-2022	365 times/yr (daily)	Cumulative rainfalls during 60 days
	Marine Water Quality Automated Monitoring Network	Solar radiation	Masan Bongam	2013-2022	105,120 times/yr (every 5 minutes)	
Effluent	Water Emission Management System	Date, COD, TN, TP	Masan-Changwon	2012-2022	365 times/yr (daily)	

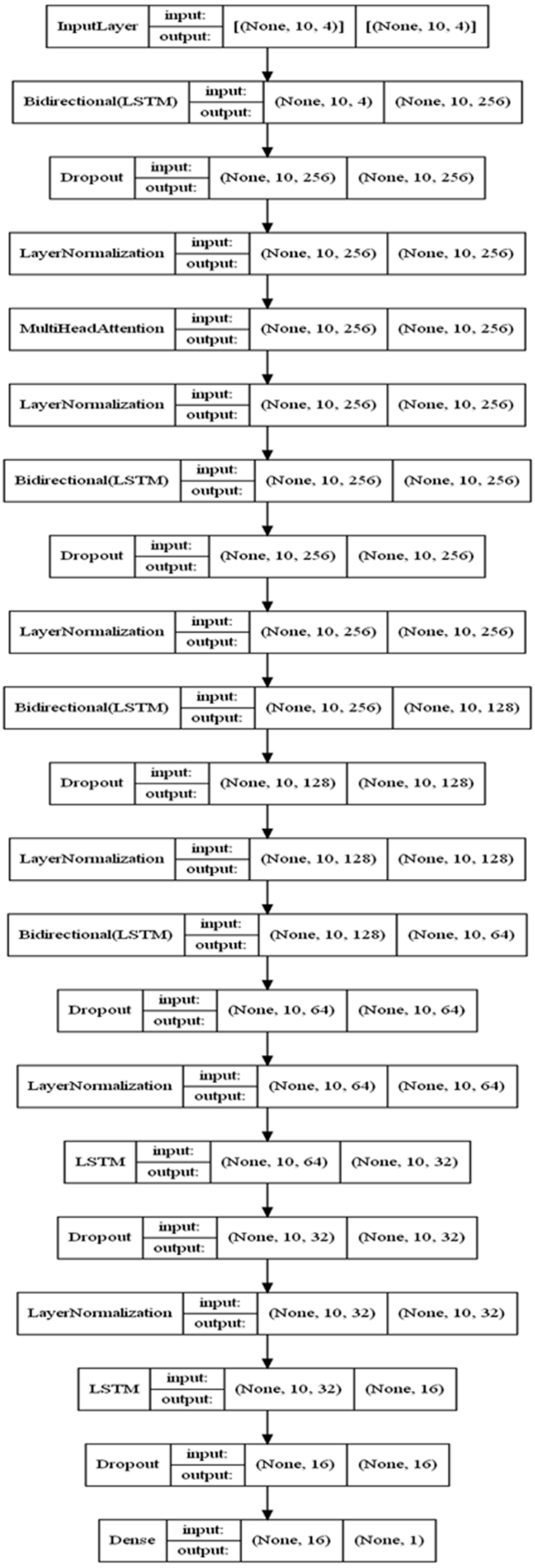

Fig. S1. Graphical description of LSTM model architecture.

Table S2. Signal-to-noise ratio (SNR) of selected variables as learned data to raw data

Variable	FW_TP_60	SDE_COD_0	SDE_TP_60	COD_sur	NH_4-N_bot	SiO_2-Si_sur	DO_bot
SNR	0.236	0.017	0.121	0.063	0.141	0.287	0.039

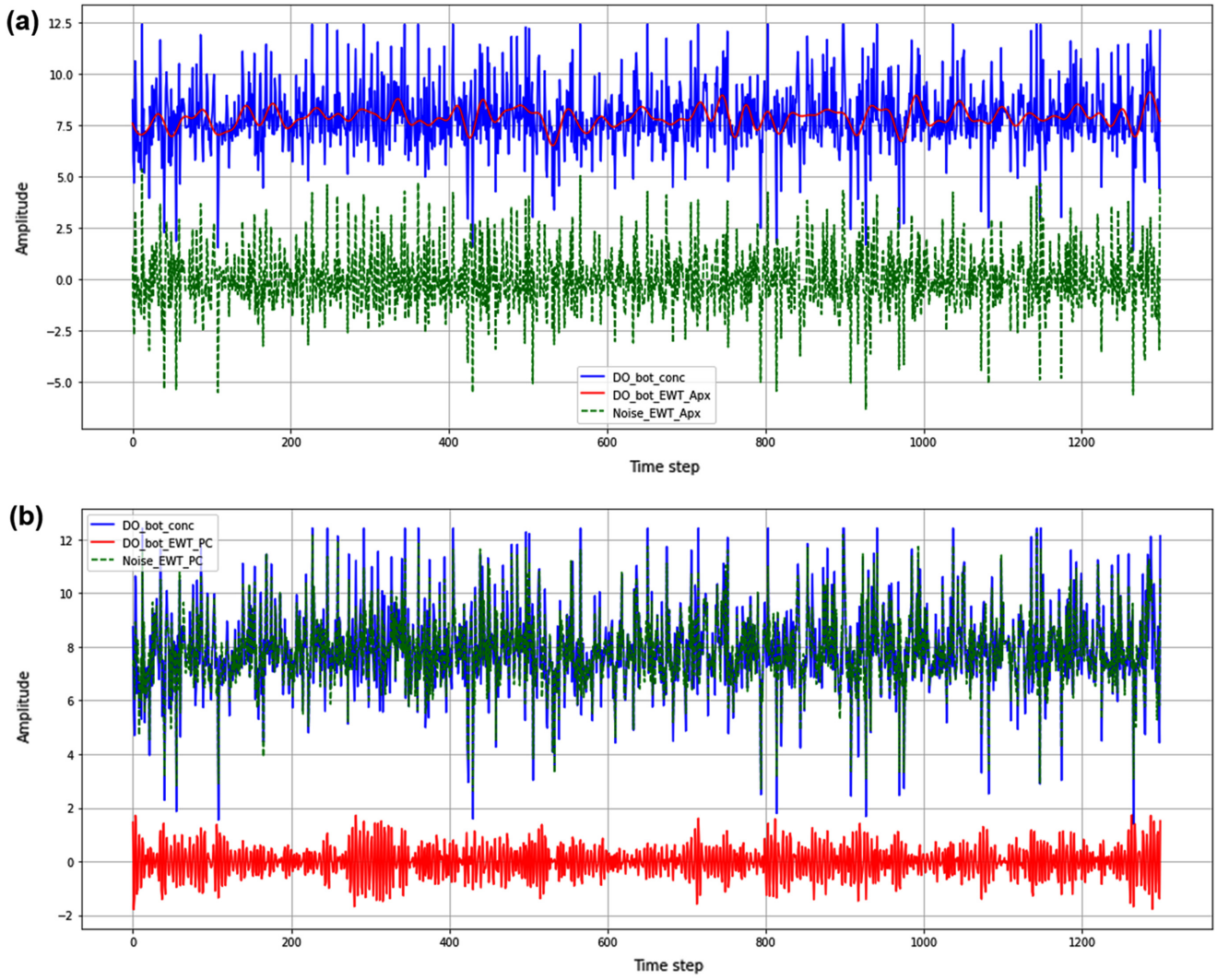

Fig. S2. Plots for signals and noises of bottom dissolved oxygen data before and after wavelet transform and dimension reduction.