RECONSTRUCTING THE BREATHLESS OCEAN WITH SPATIO-TEMPORAL GRAPH LEARNING

Bin Lu[1], **Ze Zhao**[1], **Luyu Han**[2], **Xiaoying Gan**[1]*, **Yuntao Zhou**[2], **Lei Zhou**[2],
Luoyi Fu[3], **Xinbing Wang**[1,3], **Chenghu Zhou**[4], **Jing Zhang**[2]
[1]Department of Electronic Engineering, Shanghai Jiao Tong University
[2]School of Oceanography, Shanghai Jiao Tong University
[3]Department of Computer Science, Shanghai Jiao Tong University
[4]Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences
{robinlu1209, ganxiaoying}@sjtu.edu.cn

ABSTRACT

The ocean is currently undergoing severe deoxygenation. Accurately reconstructing the breathless ocean is crucial for assessing and protecting marine ecosystem in response to climate change. Existing expert-dominated numerical simulations fail to catch up with the dynamic variation caused by global warming and human activities. Besides, due to the high-cost data collection, the historical observations are severely sparse, leading to big challenge for precise reconstruction. In this work, we propose OXYGENERATOR, the first spatio-temporal graph learning model, to reconstruct the global ocean deoxygenation from 1920 to 2023. Specifically, to address the heterogeneity across large temporal and spatial scales, we propose zoning-varying graph message-passing to capture the complex oceanographic correlations between missing values and sparse observations. Additionally, to further calibrate the uncertainty, we incorporate inductive bias from dissolved oxygen (DO) variations and chemical effects. Compared with in-situ DO observations, OXYGENERATOR significantly outperforms CMIP6 numerical simulations, reducing MAPE by 38.77%, demonstrating a promising potential to understand the ocean deoxygenation in data-driven manner.

1 INTRODUCTION

Oxygen is fundamentally essential for all life. Unfortunately, recent research Schmidtko et al. (2017) has shown that the concentration of dissolved oxygen (DO) in the ocean has been steadily decreasing over the past 50 years, indicating the acceleration of global ocean deoxygenation. To quantitatively understand and predict the long-term trend of global ocean deoxygenation, oceanographers simulate the DO concentration based on climate models without utilizing in-situ DO observations. Coupled Model Intercomparison Project Phase 6 (CMIP6) Eyring et al. (2016), a world-wide simulation compasrion project, includes different numerical simulation over a century. However, these models are unable to adjust for DO simulation biases caused by global warming and human activities, leading to error propagation and showing large discrepancies with in-situ observations.

In recent years, with the advancement of ocean observation technologies and international ocean discovery programs, a lot of in-situ observation data has been accumulated. However, the long-term historical observations of DO is severely sparse due to the high-cost and high-risk marine scientific expeditions. For example, the World Ocean Database (WOD) Boyer et al. (2018) is world’s largest and widely-used collection of publicly available ocean profile data[1], and we are extremely surprised to find that more than 96.265% DO observation data are missing in the past 100 years. Therefore, this suggests the following

Question: *Can deep learning methods more accurately reconstruct global ocean deoxygenation over a century under sparse dissolved oxygen observations?*

[*] Correspondence to: Xiaoying Gan (ganxiaoying@sjtu.edu.cn)

[1] We follow the same method of data gridding in He et al. (2019) with 1° × 1° spatial resolution, 1-year temporal resolution and 0-5500 meters (33 depth levels) of the global ocean.

Challenges. This research question drives the design of specific deep learning methods for reconstructing global ocean deoxygenation. Fortunately, despite the severe sparse observations, the missing values contains implicit spatio-temporal correlations with neighboring observations. Meanwhile, multiple physical-biogeochemical factors show strong connections with dissolved oxygen. However, as a double-edged sword, accurately characterizing the complex correlations between missing values and sparse observations involves two main challenges: (1) *Irregular 4D spatio-temporal heterogeneity*. Dynamic ocean is a four-dimensional irregular spatio-temporal area that includes longitude, latitude, depth, and time. Constrained by both tectonic plates and seafloor topography, the ocean is not a regular cube for gridding. In addition, the spatio-temporal correlations in various regions are different due to the influence of ocean circulations and climate changes. (2) *Coupled physical-biogeochemical properties*. The concentration of oceanic dissolved oxygen is influenced by a variety of factors. For one thing, the solubility of oxygen in the water is affected by physical factors, including temperature, salinity and pressure. For another, biological processes, such as photosynthesis and respiration, and organic matter decomposition play a significant role in regulating dissolved oxygen concentrations.

Our Work. To address the aforementioned challenges, we propose OxyGenerator to perform regression prediction on each 4D coordinate, reconstructing global ocean deoxygenation from 1920 to 2023. Besides, we collect more than 6 billion multi-variable oceanic observation records from multiple databases. To summarize, the main contributions of our work are as follows:

- To the best of our knowledge, OxyGenerator is the first deep learning based work to reconstruct global ocean deoxygenation over a century from real-world observations, which significantly outperforms the expert-dominated numerical simulation results with a mean absolute percentage error (MAPE) reduction of 38.77%.
- We propose zoning-varying message-passing via graph hypernetwork to obtain consistent 4D spatio-temporal reconstruction, achieving adaptive ocean zoning.
- We propose the chemistry-informed gradient variance regularization to calibrate the uncertainty of reconstruction, which combines the inductive bias between dissolved oxygen variation and other nutrients (nitrogen, phosphorus) in chemical equations.

2 Problem Definition

Reconstruction of breathless ocean aims to estimate and fill in the missing values within sparse dissolved oxygen observations. Here we consider the four dimensional coordinate to represent oceanic observation, i.e., longitude, latitude, depth, and time. Let $\mathbf{\Omega} = (\omega_{i,j,d,t})_{i,j,d,t} \in \{0,1\}^{L\times G\times D\times T}$be a binary indicator representing observed entries, i.e. $\omega_{i,j,d,t} = 1$ *iff* the entry (i, j, d, t) is observed, otherwise it is missing. We denote the incomplete data matrix $\mathbf{X}$ as follows:

$$\mathbf{X} = \mathbf{X}^{(obs)} \odot \mathbf{\Omega} + \text{NA} \odot (\mathbb{1}_{L\times G\times D\times T} - \mathbf{\Omega}),$$

where $\mathbf{X}^{(obs)} \in \mathbb{R}^{L\times G\times D\times T}$ contains the observed entries, NA denotes the indicator of not available data observation, $\odot$ is the element-wise product and $\mathbb{1}_{L\times G\times D\times T}$ is an $L \times G \times D \times T$ matrix filled with ones. Given the data matrix $\mathbf{X}$, our goal is to construct an estimate $\hat{\mathbf{X}}$ filling the missing entries of $\mathbf{X}$, which can be written as

$$\hat{\mathbf{X}} = \mathbf{X}^{(obs)} \odot \mathbf{\Omega} + \hat{\mathbf{X}}^{(imp)} \odot (\mathbb{1}_{L\times G\times D\times T} - \mathbf{\Omega}),$$

where $\hat{\mathbf{X}}^{(imp)}$ contains the imputed values. However, due to the high scarcity and unavailability of historical dissolved oxygen records (only 3.735% of data are observed), we directly conduct regression prediction in each time frame and compare it with the observed DO. Specifically, we use past T timesteps DO observation $X_{t-T:t-1}$, future T timesteps DO observation $X_{t+1:t+T}$, and some auxiliary variables $\mathcal{D}$ to reconstruct the dissolved oxygen observations at time t. Then, we compare them with the actual observations at time t to achieve performance comparison. In other words, we define the following loss function $\mathcal{L}_r$ of our model to minimize the reconstruction error:

$$\mathcal{L}_r = \sum_t \mathcal{L}(\mathcal{M}(X_{t-T:t-1}, X_{t+1:t+T}, \mathcal{D}) \odot \Omega_t, X_t^{(obs)} \odot \Omega_t), \tag{1}$$

where $\mathcal{M}$ is our proposed OxyGenerator, $\mathcal{L}$ is the mean squared error loss.

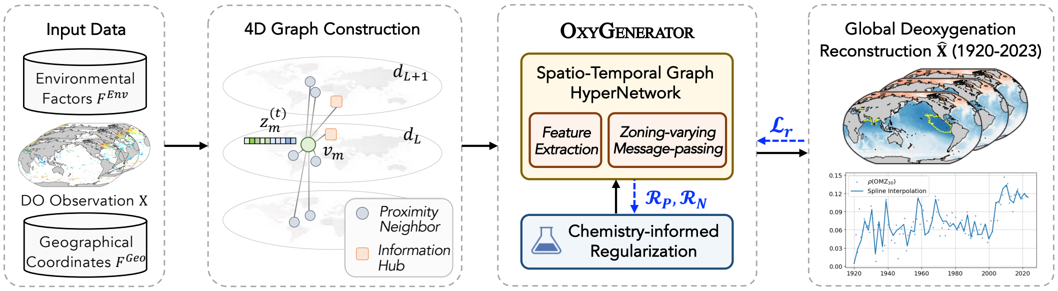

Input Data
Environmental Factors F^{Env}
DO Observation X
Geographical Coordinates F^{Geo}
4D Graph Construction
d_{L+1}
$z_m^{(t)}$
v_m
d_L
Proximity Neighbor
Information Hub
OxyGenerator
Spatio-Temporal Graph HyperNetwork
Feature Extraction
Zoning-varying Message-passing
$\mathcal{R}_P, \mathcal{R}_N$
Chemistry-informed Regularization
$\mathcal{L}_r$
Global Deoxygenation Reconstruction $\hat{X}$ (1920-2023)
0.15
0.12
0.09
0.06
0.03
0.00
1920
1940
1960
1980
2000
2020
$\rho(OMZ_{30})$
Spline Interpolation

Figure 1: Framework of breathless ocean reconstruction via our proposed OxyGenerator.

3 Methodology

In this section, we introduce the methodology of ocean deoxygenation reconstruction and detailed architecture of our proposed OxyGenerator in Figure 1. Specifically, we first propose graph-based modeling to connect both local and remote observations in irregular four-dimension space. To capture the *spatio-temporal heterogeneity* in different regions, inspired by the zoning strategy in oceanography, we propose zoning-varying graph message-passing mechanisms via hypernetwork. Moreover, to fuse the knowledge of *physical-biogeochemical properties*, we integrate multiple environmental factors and geographical coordinates for nonlinear feature extraction. Especially for the chemical effects in ocean deoxygenation, we leverage the thermodynamic equilibrium among dissolved oxygen (O), nitrogen (N) and phosphorus (P) in Equation 2 to calibrate the uncertainty of reconstruction. For detailed information on the method, please refer to Appendix D.

$$\underbrace{(CH_2O)_{106}(NH_3)_{16}H_3PO_4}_{\text{Organic Matter}} + \underbrace{138O_2}_{\text{DO}} \rightarrow 106CO_2 + 122H_2O + \underbrace{16HNO_3 + H_3PO_4}_{\text{Inorganic Nutrients}} \tag{2}$$

4 Experiment

Evaluation Methods. Due to the irreproducibility of historical ocean observations, only existing observations can be utilized for evaluation. Meanwhile, due to the data scarcity, further data partitioning for training/validation/testing would render the interpolation and imputation algorithms inapplicable. Therefore, we treat the deoxygenation reconstruction as a regression task independent of current-time observations and perform 4-fold cross testing of the collected data. For each fold, we randomly choose 25% observation data as the test data and the rest as training and validation. We report on the performance on each fold test data and the average performance on 4 folds.

Evaluation Metrics. We employ four metrics for performance evaluation, including Root Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE), Mean Absolute Error (MAE), and Coefficient of Determination (R^2).

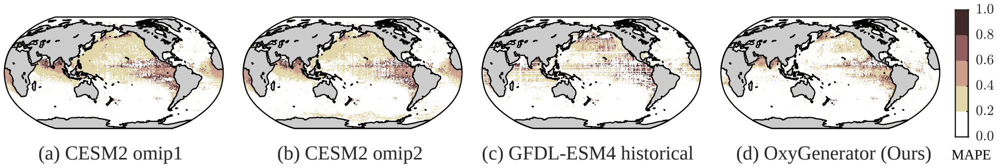
Gold-standard Benchmark. As we are the first research to use deep learning methods for global dissolved oxygen reconstruction over a century, our method is benchmarked against three advanced expert-dominated simulation results from Coupled Model Inter-comparison Project Phase 6 (CMIP6): **CESM2 omip1** Danabasoglu et al. (2020), **CESM2 omip2** Danabasoglu et al. (2020) and **GFDL-ESM4 historical** Dunne et al. (2020).

4.1 Experimental Result Comparison

In Table 1, we illustrate the comparison results of 4-fold cross testing. Our OxyGenerator achieves the best performance on all four metrics, with a 38.77% reduction in MAPE compared to suboptimal numerical simulation methods. It proves that deep learning methods can more accurately reconstruct ocean deoxygenation trends. Figure 2 depicts the average MAPE of four methods in spatial distribution. Owing to the strength of capturing heterogeneous spatio-temporal correlations, OxyGenerator provides more accurate reconstruction in the North Pacific, Indian Ocean, Equatorial Atlantic and other regions.

Table 1: Comparison with simulation results from CMIP6 numerical simulation methods. The best results are highlighted in bold, and the second best is underlined.

Benchmark	$k = 1$	$k = 2$	$k = 3$	$k = 4$	Average Performance			
	MAPE	MAPE	MAPE	MAPE	MAPE	R2	RMSE	MAE
CESM2 omip1	23.63	23.15	$\underline{23.67}$	$\underline{22.87}$	$\underline{23.32_{\pm 0.38}}$	$0.7966_{\pm 0.0064}$	$37.37_{\pm 0.34}$	$25.98_{\pm 0.31}$
CESM2 omip2	$\underline{23.62}$	$\underline{23.00}$	24.60	23.13	$23.58_{\pm 0.72}$	$0.7947_{\pm 0.0096}$	$38.22_{\pm 0.55}$	$27.12_{\pm 0.32}$
GFDL-ESM4 historical	26.13	24.01	26.68	24.33	$25.28_{\pm 1.31}$	$\underline{0.8228_{\pm 0.0051}}$	$\underline{35.45_{\pm 0.65}}$	$\underline{23.69_{\pm 0.38}}$
OXYGENERATOR (Ours)	**14.72**	**13.48**	**15.72**	**13.20**	$\mathbf{14.28_{\pm 1.16}}$	$\mathbf{0.9026_{\pm 0.0072}}$	$\mathbf{26.31_{\pm 1.23}}$	$\mathbf{17.57_{\pm 1.10}}$
Improvement	37.67%	41.38%	33.59%	42.28%	38.77%	9.70%	25.78%	25.83%

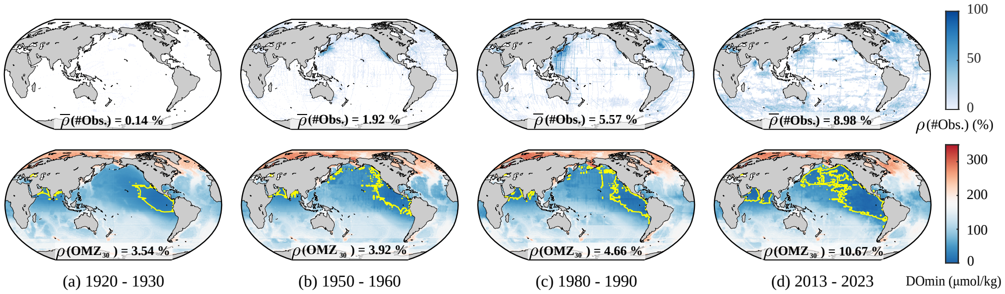


(a) CESM2 omip1 (b) CESM2 omip2 (c) GFDL-ESM4 historical (d) OxyGenerator (Ours)

Figure 2: The spatial distribution of MAPE in global ocean deoxygenation reconstruction using different methods (The darker the color, the greater the error).

4.2 OCEAN DEOXYGENATION RECONSTRUCTION ANALYSIS

In Figure 3, we show the results before and after reconstruction using OXYGENERATOR. The above figures shows the proportion of ocean data observed in four-dimensional coordinates $\rho(\#.\mathrm{Obs})$. Overall, dissolved oxygen (DO) observation are very sparse in each interval, and many areas have no observations. In the below, the figure shows the minimum DO reconstructed by OXYGENERATOR. The yellow line envelopes the oxygen minimum zone (OMZ) where $\mathrm{DO}_{\mathrm{min}} \leq 30\mu$mol/kg. $\rho(\mathrm{OMZ}_{30})$ indicates the proportion of OMZ30 regions to global oceans, which clearly shows a significant increase over a century.

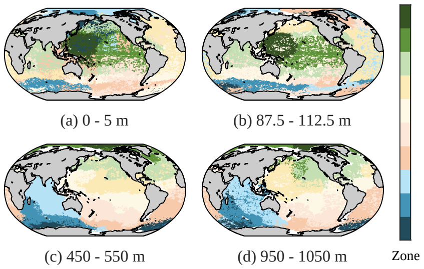

$\overline{\rho}$(#Obs.) = 0.14 %
$\overline{\rho}$(#Obs.) = 1.92 %
$\overline{\rho}$(#Obs.) = 5.57 %
$\overline{\rho}$(#Obs.) = 8.98 %
100
50
0
ρ(#Obs.) (%)
$\rho(\mathrm{OMZ}_{30})$ = 3.54 %
$\rho(\mathrm{OMZ}_{30})$ = 3.92 %
$\rho(\mathrm{OMZ}_{30})$ = 4.66 %
$\rho(\mathrm{OMZ}_{30})$ = 10.67 %
300
200
100
0
DOmin (μmol/kg)

(a) 1920 - 1930 (b) 1950 - 1960 (c) 1980 - 1990 (d) 2013 - 2023

Figure 3: Global ocean deoxygenation reconstruction via OXYGENERATOR from 1920 to 2023 .

4.3 EFFECT OF ADPATIVE ZONING VIA SPATIO-TEMPORAL HYPERNETWORK

As shown in Figure 4, OXYGENERATOR can adaptively carry out spatial zoning, which is quantified into 10 zones. Moreover, compared with WOD zoning strategy Boyer et al. (2018), we observe that: (1) The adaptive zoning is closely related to latitude distribution and presents a clear oceanic partitioning. Especially below 450 meters depth, we show similar zoning results with WOD in the Indian Ocean, Atlantic Ocean, and Pacific Ocean. (2) Compared to depth-independent zoning in WOD, our method depicts a finer grained zoning that varies at different depths and intra-ocean. Surface seawater is influenced more by human activities and exhibits a more intricate zoning pattern, whereas deep seawater is closely associated with geographical location.



(a) 0 - 5 m (b) 87.5 - 112.5 m (c) 450 - 550 m (d) 950 - 1050 m

Figure 4: Adaptive zoning of 4 representative depth level in 2016. Different colors represent different zones.

5 Conclusion

In this paper, we propose OxyGenerator, a deep learning model that effectively reconstructs global ocean deoxygenation based on sparse observation data in 1920-2023. In the future, we will continue to collaborate with oceanographers to further improve the compliance with physical-biogeochemical mechanisms and investigate its impacts for marine ecosystem.

Acknowledgement

This work was partially supported by National Key R&D Program of China (No.2022YFB3904204), National Natural Science Foundation of China (No. 62272301, 42125601, 42276201, 42050105, 62020106005, 62061146002, 61960206002).

References

Karianne J. Bergen, Paul Allan Johnson, Maarten V. de Hoop, and Gregory C. Beroza. Machine learning for data-driven discovery in solid earth geoscience. *Science*, 363, 2019.

Kaifeng Bi, Lingxi Xie, Hengheng Zhang, Xin Chen, Xiaotao Gu, and Qi Tian. Accurate medium-range global weather forecasting with 3d neural networks. *Nature*, 619(7970):533–538, 2023.

L. Bopp, L. Resplandy, J. C. Orr, S. C. Doney, J. P. Dunne, M. Gehlen, P. Halloran, C. Heinze, T. Ilyina, R. Séférian, J. Tjiputra, and M. Vichi. Multiple stressors of ocean ecosystems in the 21st century: projections with cmip5 models. *Biogeosciences*, 10(10):6225–6245, 2013.

L. Bopp, L. Resplandy, A. Untersee, P. Le Mezo, and M. Kageyama. Ocean (de)oxygenation from the last glacial maximum to the twenty-first century: insights from earth system models. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 375(2102):20160323, 2017.

T.P. Boyer, O.K. Baranova, C. Coleman, H.E. Garcia, A. Grodsky, R.A. Locarnini, A.V. Mishonov, C.R. Paver, J.R. Reagan, D. Seidov, I.V. Smolyar, K. Weathers, and M.M. Zweng. World ocean database 2018. Technical Report 87, NOAA Atlas NESDIS, 2018. `https://www.ncei.noaa.gov/sites/default/files/2020-04/wod_intro_0.pdf`.

G. Danabasoglu, J. F Lamarque, J. Bacmeister, D. A. Bailey, K. A, J. Edwards, L. K. Emmons, J. Fasullo, R. Garcia, A. Gettelman, C. Hannay, M. M. Holland, W. G. Large, P. H. Lauritzen, D. M. Lawrence, J. T. M. Lenaerts, K. Lindsay, W. H. Lipscomb, M. J. Mills, R. Neale, K. W. Oleson, B. Otto-Bliesner, A. S. Phillips, W. Sacks, S. Tilmes, L. Van Kampenhout, M. Vertenstein, A. Bertini, J. Dennis, C. Deser, C. Fischer, B. Fox-Kemper, J. E. Kay, D. Kinnison, P. J. Kushner, V. E. Larson, M. C. Long, S. Mickelson, J. K. Moore, E. Nienhouse, L. Polvani, P. J. Rasch, and W. G. Strand. The community earth system model version 2 (cesm2). *Journal of Advances in Modeling Earth Systems*, 12(2), 2020. ISSN 1942-2466. doi: 10.1029/2019ms001916. URL `https://dx.doi.org/10.1029/2019MS001916`.

J. P. Dunne, L. W. Horowitz, A. J. Adcroft, P. Ginoux, I. M. Held, J. G. John, J. P. Krasting, S. Malyshev, V. Naik, F. Paulot, E. Shevliakova, C. A. Stock, N. Zadeh, V. Balaji, C. Blanton, K. A. Dunne, C. Dupuis, J. Durachta, R. Dussin, P. P. G. Gauthier, S. M. Griffies, H. Guo, R. W. Hallberg, M. Harrison, J. He, W. Hurlin, C. McHugh, R. Menzel, P. C. D. Milly, S. Nikonov, D. J. Paynter, J. Ploshay, A. Radhakrishnan, K. Rand, B. G. Reichl, T. Robinson, D. M. Schwarzkopf, L. T. Sentman, S. Underwood, H. Vahlenkamp, M. Winton, A. T. Wittenberg, B. Wyman, Y. Zeng, and M. Zhao. The gfdl earth system model version 4.1 (gfdl-esm 4.1): Overall coupled model description and simulation characteristics. *Journal of Advances in Modeling Earth Systems*, 12(11), 2020. ISSN 1942-2466. doi: 10.1029/2019ms002015. URL `https://dx.doi.org/10.1029/2019MS002015`.

Veronika Eyring, Sandrine Bony, Gerald A Meehl, Catherine A Senior, Bjorn Stevens, Ronald J Stouffer, and Karl E Taylor. Overview of the coupled model intercomparison project phase 6 (cmip6) experimental design and organization. *Geoscientific Model Development*, 9(5):1937–1958, 2016.

AR Fay and GA McKinley. Global open-ocean biomes: mean and temporal variability. *Earth System Science Data*, 6(2):273–284, 2014.

Thomas L. Frölicher, Keith B. Rodgers, Charles A. Stock, and William W. L. Cheung. Sources of uncertainties in 21st century projections of potential ocean ecosystem stressors. *Global Biogeochemical Cycles*, 30(8):1224–1243, 2016. ISSN 0886-6236.

H. E. Garcia, R. A. Locarnini, T. P. Boyer, J. I. Antonov, O. K. Baranova, M. M. Zweng, and D. R. Johnson. *World Ocean Atlas 2009, Volume 3: Dissolved Oxygen, Apparent Oxygen Utilization, and Oxygen Saturation*. Number 70 in NOAA Atlas NESDIS. U.S. Government Printing Office, Washington, D.C., 2010.

Hongjing Gong, Chao Li, and Yuntao Zhou. Emerging global ocean deoxygenation across the 21st century. *Geophysical Research Letters*, 48(23), 2021. ISSN 0094-8276.

David Ha, Andrew M. Dai, and Quoc V. Le. Hypernetworks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*, 2017.

Garcia He, Weathers Kw, Paver Cr, Smolyar I, Boyer Tp, Locarnini Mm, Zweng Mm, Mishonov Av, Baranova Ok, Seidov D, and Reagan Jr. World ocean atlas 2018, volume 3: Dissolved oxygen, apparent oxygen utilization, and dissolved oxygen saturation. Report, 2019. URL `https://archimer.ifremer.fr/doc/00651/76337/`.

Remi Lam, Alvaro Sanchez-Gonzalez, Matthew Willson, Peter Wirnsberger, Meire Fortunato, Ferran Alet, Suman Ravuri, Timo Ewalds, Zach Eaton-Rosen, Weihua Hu, Alexander Merose, Stephan Hoyer, George Holland, Oriol Vinyals, Jacklynn Stott, Alexander Pritzel, Shakir Mohamed, and Peter Battaglia. Learning skillful medium-range global weather forecasting. *Science*, pp. eadi2336, 2023.

Tung Nguyen, Johannes Brandstetter, Ashish Kapoor, Jayesh K. Gupta, and Aditya Grover. Climax: A foundation model for weather and climate. In *International Conference on Machine Learning*, 2023.

Lena Podina, Brydon Eastman, and Mohammad Kohandel. Universal physics-informed neural networks: Symbolic differential operator discovery with sparse data. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pp. 27948–27956. PMLR, 2023.

Maziar Raissi, Paris Perdikaris, and George E. Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *J. Comput. Phys.*, 378:686–707, 2019.

Markus Reichstein, Gustau Camps-Valls, Bjorn Stevens, Martin Jung, Joachim Denzler, Nuno Carvalhais, and Prabhat. Deep learning and process understanding for data-driven earth system science. *Nature*, 566:195 – 204, 2019.

Gabriel Reygondeau, William W. L. Cheung, Colette C. C. Wabnitz, Vicky W. Y. Lam, Thomas L. Frölicher, and Olivier Maury. Climate change-induced emergence of novel biogeochemical provinces. In *Frontiers in Marine Science*, 2020.

Sunke Schmidtko, Lothar Stramma, and Martin Visbeck. Decline in global oceanic oxygen content during the past five decades. *Nature*, 542:335–339, 2017.

Maike Sonnewald, Stephanie Dutkiewicz, Chris N. Hill, and Gael Forget. Elucidating ecological complexity: Unsupervised learning determines global marine eco-provinces. *Science Advances*, 6, 2020.

Waldo R Tobler. A computer movie simulating urban growth in the detroit region. *Economic geography*, 46:234–240, 1970.

Qi Zeng, Yash Kothari, Spencer H. Bryngelson, and Florian Schäfer. Competitive physics informed networks. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*, 2023.

Yuntao Zhou, Hongjing Gong, and Feng Zhou. Responses of horizontally expanding oceanic oxygen minimum zones to climate change based on observations. *Geophysical Research Letters*, 49, 2022.

A Background of Ocean Deoxygenation

Drivers and Impacts. Ocean deoxygenation is a complex environmental issue characterized by the significant reduction in dissolved oxygen levels in the ocean. Several interconnected factors, especially the climate change and anthropogenic activities, contribute to ocean deoxygenation. Rising temperatures impact oxygen circulation and diminish the capacity of seawater to hold dissolved oxygen, resulting in decreased oceanic oxygen levels. Excessive nutrient input from human activities, particularly from agriculture and industrial processes, lead to the proliferation of algae, creating algal blooms. When these blooms eventually decompose, they consume large amounts of oxygen, creating localized areas with reduced oxygen levels. In recent years, with the intensification of global warming and human influence, ocean deoxygenation has presented an accelerating trend. The consequences of ocean deoxygenation are far-reaching. Reduced oxygen levels can harm marine biodiversity, leading to disruptions in food webs and ecosystems. Fisheries, which rely on oxygen-rich environments to support commercially important species, may experience declines. Hence, there is an urgent need to understand the mechanisms driving deoxygenation, assess its ecological consequences, and develop effective strategies for sustainable development.

El Niño-Southern Oscillation (ENSO). Anomalous climate events influence the distribution and rates of ocean deoxygenation. Taking the El Niño-Southern Oscillation (ENSO) as an example, this climatic phenomenon greatly disturbs oceanic dissolved oxygen levels. During El Niño years, the warming of the central and eastern equatorial Pacific reduces the solubility of oxygen, leading to a decline in dissolved oxygen concentrations in affected areas. Understanding the intricate relationship between ENSO and oceanic dissolved oxygen is crucial for comprehending the broader implications of climate-related variations on marine ecosystems.

Oxygen Minimum Zones (OMZs). Oxygen Minimum Zones (OMZs), standing for low oxygen zones, plays a pivotal role in characterizing the issue of oceanic oxygen deficiency. OMZs are emerging in various regions, posing challenges for marine organisms adapted to higher oxygen concentrations. In various studies, OMZs are bounded by different thresholds of dissolved oxygen levels. In our work, we select a 30 μmol/kg threshold to define OMZs, referred to as OMZ30. To be specific, areas with dissolved oxygen concentrations below 30 μmol/kg at any depth are labeled as OMZ30. In Figure 3, OMZ30 are highlighted by yellow lines and our work indicates the significant expansion of OMZ30 over the past century. This observation signifies a clear trend of ocean deoxygenation, emphasizing the urgent need for in-depth investigations into deoxygenation rates and driving factors.

B Data Collection and Quality Control

We collect over 6 billion historical observation records from 1920 to 2023 of dissolved oxygen and relevant environmental factors, including temperature, salinity, nitrates, phosphates, silicates, and chlorophyll, from multiple public databases in Table 2.

Table 2: Detailed Information of Data sources for global ocean observations.

Database	Time	Institution	Source	Access Date	Variables
World Ocean Database (WOD 2018)	1900-2023	National Centers for Environmental Information	`https://www.ncei.noaa.gov/`	2023-05	temperature, salinity, dissolved oxygen, biogeochemical major elements (P, N, Si) and chlorophyll.
CLIVAR and Carbon Hydrographic Database (CCHDO)	1922-2023	CLIVAR and Carbon Hydrographic Data Office	`https://cchdo.ucsd.edu/`	2023-05	temperature, salinity, dissolved oxygen, and biogeochemical major elements (P, N, Si).
Argo	2001-2023	Argo Global Data Assemby Center	`https://argo.ucsd.edu/`	2023-05	temperature, salinity, dissolved oxygen, biogeochemical major elements (N) and chlorophyll.
Global Ocean Data Analysis Project version2.2022 (GLODAPV2_2022)	1972-2021	NOAA's National Centers for Environmental Information (NCEI)	`https://glodap.info/`	2023-05	temperature, salinity, dissolved oxygen, biogeochemical major elements (P, N, Si) and chlorophyll.
Geotraces IDP	2007-2018	GEOTRACES International Data Assembly Centre (GDAC)	`https://www.geotraces.org`	2023-10	temperature, salinity, dissolved oxygen, biogeochemical major elements (P, N, Si) and chlorophyll.

We follow the data preprocessing in existing works Schmidtko et al. (2017) and conduct formatting standardization and spatio-temporal tagging correction for all data. Besides, we establish unified quality control standards to ensure the availability and reliability of observations. Hereby, we obtain data matrix of dissolved oxygen $\mathbf{X}$, environmental factors $\mathbf{F}^{\text{Env}}$ and geographical coordinates $\mathbf{F}^{\text{Geo}}$. In our work, we follow the spatial and temporal resolution setting in Garcia et al. (2010); Eyring et al. (2016), defined as the annual average temporal resolution from 1920 to 2023, spatial resolution of $1^{\circ} \times 1^{\circ}$, and a total of 33 depth levels from 0 to 5500 meters. Hence, the dimension range of the data is $L = 360$, $G = 180$, $D = 33$ and $T = 104$.

C Related Work

In this section, we briefly review the related research lines to our work.

C.1 Ocean Deoxygenation

Global warming and excessive nutrient inputs caused by human activities have led to significant ocean deoxygenation in recent years. To better understand the oxygen cycling mechanism and assess the overall impact of human activities on the marine system since the 20th century, a comprehensive global analysis of ocean deoxygenation is particularly important. Existing research has made two ways of positive attempts, but still leaves some limitations: (1) *Numerical Simulation Models*. Existing expert-dominated studies utilize numerical simulations based on climate models to probe the drivers and predict oxygen loss Gong et al. (2021); Bopp et al. (2017). For example, Coupled Model Intercomparison Project Phase 6 (CMIP6) Eyring et al. (2016) includes three experiments (CESM2-omip1, CESM2-omip2 and GFDL-ESM4-historical) on dissolved oxygen simulation. Nevertheless, most simulations entirely rely on knowledge of the climate system and fail to leverage observations for correction, thereby showing inferior performance Frölicher et al. (2016); Bopp et al. (2013). (2) *Spatial Interpolation Methods*. Due to the severe sparse DO observations, another aspect of studies attempt data reconstruction through distance based weighted average Schmidtko et al. (2017) and geostatistical regression Zhou et al. (2022) for spatial interpolation. However, these methods are only smoothing of the existing DO observation data and yield inaccurate results for areas that lack observation data. They ignore the spatio-temporal heterogeneity of different regions and fail to make full use of auxiliary variables for reconstruction. In this paper, to the best of our knowledge, we are the first to propose a deep learning method to reconstruct global ocean deoxygenation over a century, considering the spatio-temporal hetegeneity in different regions (Section D.2) and chemical properties across dissovled oxygen and nutrients (Section D.3).

C.2 Data-Driven Earth System

Existing superior earth system methods are mostly expert-dominated numerical models, which rely on complicated physics processes, sensitive initial conditions and suitable forcing. Moreover, many numerical models are computationally intensive and costly for fine-grained spatio-temporal resolution or long-range simulation. With the rapid growth of artificial intelligence, *AI for Science*, or more specifically data-driven Earth system, has become a hot topic in both Computer Science and Earth Geoscience Bergen et al. (2019); Reichstein et al. (2019). Given the accumulation of scientific data, data-driven deep learning models are attempting to learn complex and nonlinear correlations in the Earth system, while reducing computational and application costs. Especially in the field of numerical weather prediction (NWP), deep learning techniques are particularly suitable for improving its prediction performance due to their large amount of data, e.g. recent state-of-the-art methods Pangu-Weather Bi et al. (2023), GraphCast Lam et al. (2023). For another, Nguyen et al. propose a foundation model for weather and climate modeling called ClimaX, which extends Transformer architecture for more general weather and climate tasks. Overall, the data-driven Earth system is still in its early stages, and more scenarios and technologies are worth further exploration.

D Detailed Methodology

D.1 Four-Dimensional (4D) Graph Construction

Ocean observation has three-dimensional spatial coordinates (longitude, latitude and depth), while constrained by land plates and seabed topography, making the entire data observation an irregular cube. Meanwhile, due to the sparsity of observations, when there is a lack of observation information in adjacent receptive fields, localized convolution filters cannot aggregate sufficient information. Therefore, instead of CNN-based methods, we propose graph modeling to define two types of neighbors, i.e., *proximity neighbor* and *information hub*, for each data grid. We consider 3D spatial proximity under irregular boundaries. At the same time, we connect a wider range of related nodes with respect to the observation completeness, hereby improving the richness of information.

Proximity Neighbor. The *First Law of Geography* Tobler (1970) states that ”everything is related to everything else, but near things are more related than distant things.” According to that, we consider proximity neighbors as data grids within the range of $[-\delta^\circ, \delta^\circ]$ in the horizontal longitude and latitude range, as well as $\pm d_m$ depth layer in the vertical direction. Regarding the irregular oceanic boundaries, we adopt bedrock elevation data published by NOAA to ensure the rationality.

Information Hub. Due to the insufficient observations in proximity neighbors, we further expand the spatial proximity range and utilize observation completeness $\mathcal{C}_{\text{obs}}$ as an indicator to measure the richness of information within a time range from $t - T$ to $t + T$:

$$\mathcal{C}_{\text{obs}} = \frac{\|\omega_{i,j,d,t-T:t+T}\|_1}{2T} = \frac{\sum_\tau \|\omega_{i,j,d,\tau}\|}{2T}.$$

When $\mathcal{C}_{\text{obs}} \geq \varepsilon_{\mathcal{C}}$, we deem this data grid an information hub and the neighbor of node $v_{i,j,d,t}$, otherwise it is not. Thanks to the flexibility of graph modeling, information hub can effectively enlarge the receptive field, which is crucial for reconstruction under sparse observations. We define the adjacency matrix A_t and node set $\mathcal{V}_t$ at time t, and the corresponding edge feature matrix $\mathcal{E}_t$ as the difference of attributes, including distance, density, pressure, etc.

D.2 Spatio-Temporal Graph HyperNetwork

Feature Extractor. The concentration changes of dissolved oxygen are influenced by different factors. Thus, we first want to construct a feature extractor $f(\theta)$ that can fully characterize the attribute features of nodes. In order to capture the temporal variation of dissolved oxygen, we adopt a bidirectional LSTM model $f_{\text{Bi-LSTM}}$ to encode both historical and future T timesteps DO observation as follows:

$$\overrightarrow{Z_\tau^{\text{DO}}} = \text{Bi-LSTM}(X_\tau, \overrightarrow{Z_{\tau-1}^{\text{DO}}}; \overrightarrow{\theta}_{\text{Bi-LSTM}}),$$
$$\overleftarrow{Z_\tau^{\text{DO}}} = \text{Bi-LSTM}(X_\tau, \overleftarrow{Z_{\tau-1}^{\text{DO}}}; \overleftarrow{\theta}_{\text{Bi-LSTM}}),$$

where X_τ denotes the DO observation at time τ, $\overrightarrow{Z_\tau^{\text{DO}}}$ and $\overleftarrow{Z_\tau^{\text{DO}}}$ denote the hidden states from two directions, $\overrightarrow{\theta}_{\text{Bi-LSTM}}$ and $\overleftarrow{\theta}_{\text{Bi-LSTM}}$ denote the Bi-LSTM parameters. Correspondingly, the DO temporal feature at reconstruction time t is $Z_t^{\text{DO}} = \overrightarrow{Z_t^{\text{DO}}} \| \overleftarrow{Z_t^{\text{DO}}}$, thereby capturing the temporal evolution of DO in two directions. In addition, dissolved oxygen concentration is also related to a series of geographical coordinates F_t^{Geo} (including latitude, longitude, depth, time) and environmental factors F_t^{Env} (including temperature, salinity, nutrients, chlorophyll). In order to comprehensively consider the coupling correlations between these factors and dissolved oxygen, we leverage a multi-layer perceptron (MLP) to embed into a latent space: $Z_t^F = \text{MLP}(F_t^{\text{Geo}} \| F_t^{\text{Env}}; \theta_{MLP})$. To sum up, the overall latent feature embedding is the composition of both DO temporal feature Z_t^{DO} and multi-factor feature Z_t^F as $Z_t = Z_t^{\text{DO}} \| Z_t^F$.

Zoning-Varying Message-Passing. The global ocean over a century shows heterogeneous spatio-temporal correlations in different historical periods and regions due to the varying impacts of human activities and climate change. In oceanographic research, many studies partition the global ocean into different zones Fay & McKinley (2014); Reygondeau et al. (2020); Sonnewald et al. (2020) in order to better investigate the complex oceanic processes. Inspired by zoning strategy in oceanography, a naive method is to train multiple models through pre-determined zones. However, there have no theoretical basis for the partition of deoxygenation areas, and training multiple models increase the computational costs. Therefore, we propose *Spatio-Temporal Graph HyperNetwork*, which adaptively generate zoning-varying parameters via hypernetwork Ha et al. (2017) for graph message-passing. We can obtain a dynamic partition more efficiently through a globally shared parameter generator and the low dimensional context information of each node.

To simplify the explanation, we denote the m-th node in the node set $\mathcal{V}_t = \{v_{i,j,d,t}\}_{i,j,d,t}$ at time t as v_m. For each node v_m, we generate zoning-varying parameter $\mathcal{Z}_m = [\mathcal{Z}_m^\alpha, \mathcal{Z}_m^\beta]$ based on context

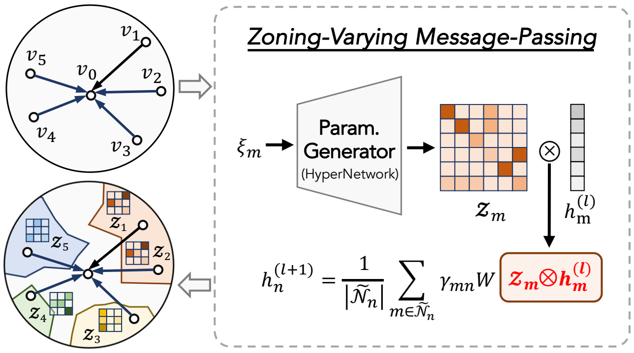


Figure 5: Framework of zoning-varying message-passing.

information ξ_m which includes geographical coordinates and physical elements:

$$\mathcal{Z}_m^{\alpha} = \mathrm{MLP}_{\alpha}(\xi_m; \phi_{\alpha}), \mathcal{Z}_m^{\beta} = \mathrm{MLP}_{\beta}(\xi_m; \phi_{\beta}),$$

where $\mathcal{Z}_m^{\alpha} \in \mathbb{R}^{d_z \times d_z}$ denotes rotation matrix, $\mathcal{Z}_m^{\beta} \in \mathbb{R}^{d_z}$ denotes bias vector. Afterwards, compared to vanilla graph neural networks (GNN), we perform non-shared feature scaling on the latent feature embedding:

$$\tilde{h}_m^{(l)} = h_m^{(l)} \otimes \mathcal{Z} \triangleq \mathcal{Z}_m^{\alpha\, T} \cdot h_m^{(l)} + \mathcal{Z}_m^{\beta},$$

where $h_m^{(l)}$ is the l-th layer GNN input of node v_m, $\tilde{h}_m^{(l)}$ is the node-wise adaptive latent feature of $h_m^{(l)}$. Thus, the nonlinear correlation between different geographic information and physical elements will serve as a criterion for dynamic partitioning. Nodes with similar parameter $\mathcal{Z}$ will belong to the same zone. Unlike unsupervised clustering or manually setting partition rules, for ocean deoxygenation problems, hypernetworks can transform discrete partitions into continuous ones. We further use graph neural networks for message passing on adaptive latent features and consider the edge features of different neighbors. Take node v_n as an example, the $l+1$-th layer GNN output is denoted as:

$$h_n^{(l+1)} = \frac{1}{|\tilde{\mathcal{N}}_n|} \sum\nolimits_{m \in \tilde{\mathcal{N}}_n} \gamma_{mn} W \tilde{h}_m^{(l)},$$

where $\tilde{z}_m^l$ is the adaptive latent feature, W is the model parameter of message-passing, and $\tilde{\mathcal{N}}_n$ is the neighbor set of node v_n with self-loop. The edge weight γ_{mn} is derived by nonlinear projection of edge feature by a multi-layer perceptron: $\gamma_{mn} = \mathrm{MLP}_{\gamma}(e_{mn})$.

D.3 Chemistry-informed Regularization

The global ocean deoxygenation is a complex system that couples physical and biochemical effects. The embedding of domain knowledge can calibrate the uncertainty of neural networks and eliminate abnormal reconstruction. However, despite some research on physics-informed neural networks Raissi et al. (2019); Podina et al. (2023); Zeng et al. (2023), there is still little research on how to express chemical dynamic equilibrium as shown in Equation 2. Here, we consider the dynamic transition equilibrium between dissolved oxygen and nitrate (phosphate), which means the gradient between dissolved oxygen and nitrate (phosphate) concentration is a constant. For example, for the observation of nitrate F_m^N (phosphate F_m^P) at node v_m (node v_n), we can calculate the corresponding gradient based on the reconstructed dissolved oxygen $\hat{x}_m$ ($\hat{x}_n$) which is approximately a constant, i.e., $\frac{\partial \hat{x}_m}{\partial F_m^N} = \text{const.}$, $\frac{\partial \hat{x}_n}{\partial F_n^P} = \text{const.}$

Therefore, we propose the chemistry-informed gradient variance regularization as one of the supervised signals. Specifically, in a batch of training data, we select nodes with nitrate (phosphate) observations and calculate their corresponding gradients. Since the dynamic equilibrium between nitrate (phosphate) and dissolved oxygen is approximately consistent, the norm of gradient variance should be small:

$$\mathcal{R}_N = \left\| \sigma\left(\left\{ \frac{\partial \hat{x}_m}{\partial F_m^N} \right\}_m \right) \right\|_2^2, \mathcal{R}_P = \left\| \sigma\left(\left\{ \frac{\partial \hat{x}_n}{\partial F_n^P} \right\}_n \right) \right\|_2^2, \tag{3}$$

where $\sigma(\cdot)$ denotes the calculation of the gradient variance.

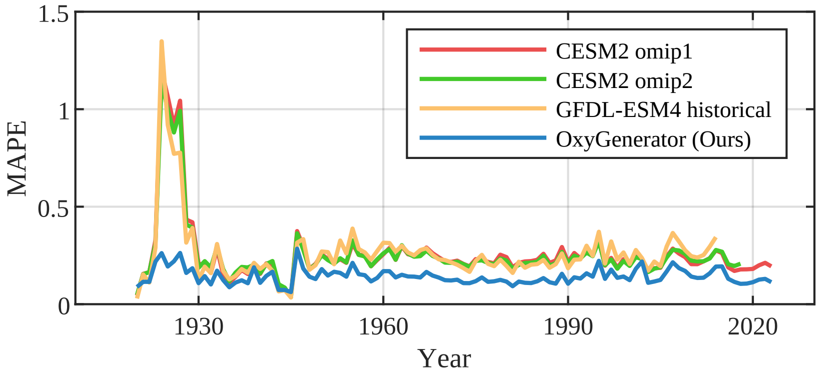


Figure 6: Performance Comparison of MAPE over time.

D.4 Optimization Algorithm

Due to the memory burden of large scale graphs for global ocean, we divide it into different training groups based on World Ocean Database 2018 Boyer et al. (2018) ocean division. Since different groups show varying complexity, each group is optimized using different iterations in one epoch of training. We periodically evaluate the reconstruction performance of different group using the validation dataset, and then reset the iterations number for each group. During the learning process, we integrate the reconstruction loss $\mathcal{L}_r$ in Equation 1 and two chemical-informed gradient variance regularization in Equation 3 as follows:

$$\mathcal{L}_{\text{OxyGenerator}} = \mathcal{L}_r + \lambda(\mathcal{R}_N + \mathcal{R}_P), \tag{4}$$

where λ is the ratio coefficient of two losses. We conclude the algorithm in Algorithm 1.

Algorithm 1 Optimization algorithm of OxyGenerator

```
Require: Observed data $\mathbf{X}$, geographical factor $\mathbf{F}^{\text{Geo}}$, environmental factor $\mathbf{F}^{\text{Env}}$
Ensure: OxyGenerator $\mathcal{M}$ with parameter $\theta^*, \phi^*$
$\theta, \phi \leftarrow$ random initialization
Split training data $\mathcal{B}_{\text{train}}$ into different batches via $N_{\text{area}}$ areas. {For each batch of data, it has a corresponding area ID.}
Initialize the iteration numbers per area $T_{\text{area}} = [1/N_{\text{area}}, \cdots, 1/N_{\text{area}}]$
while not converged or max. epochs not reached do
    for a batch of training data $B_i$ from $\mathcal{B}_{\text{train}}$ do
        Determine the area ID $n$ of $B_i$ and corresponding iteration number $T_n = T_{\text{area}}[n]$.
        for iteration number $T_n$ do
            Calculate loss $\mathcal{L}^{(train)}_{\text{OxyGenerator}}$ via Equation 4.
            Update the model parameter $\theta, \phi$.
        end for
    end for
    Update the iteration numbers per area via average validation loss on different areas, i.e. $T_{\text{area}} = \text{softmax}\left[\mathcal{L}^{(val)}_{\text{OxyGenerator}}\right]$.
end while
Return the optimized model parameter $\theta^*, \phi^*$.
```

E Additional Experiment Results

E.1 Temporal Distribution of Reconstruction Error.

We analyze the variation of reconstruction error over time, as shown in Figure 6. OxyGenerator shows the consistent reconstruction performance among different years. Notably, there exist

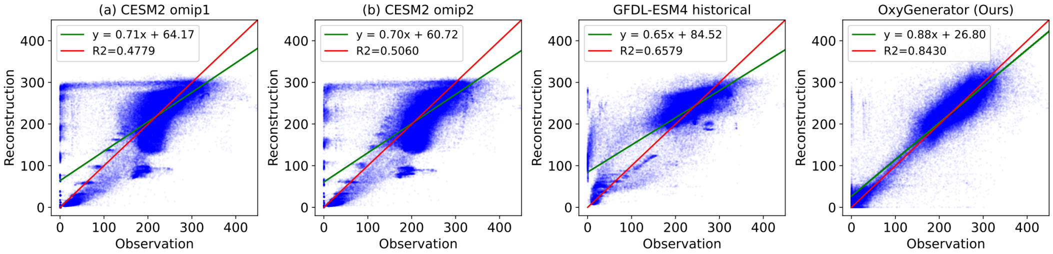
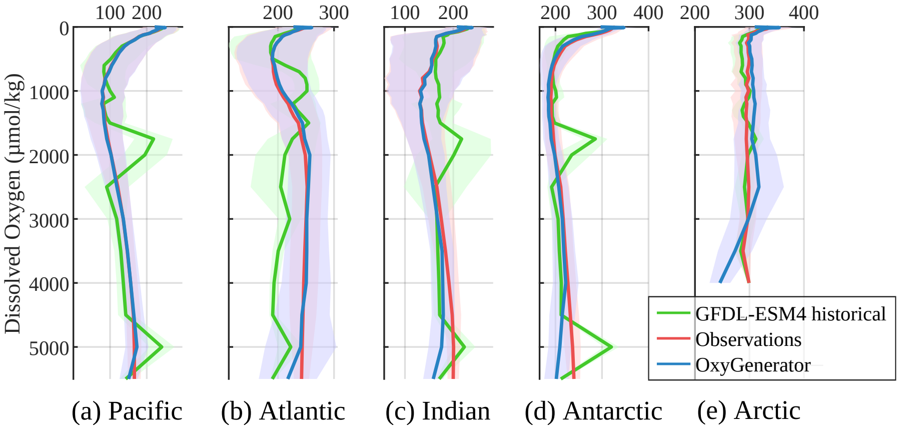


Figure 7: Performance Comparison of deoxygenation reconstruction of different numerical simulation methods and OXYGENERATOR in the Black Sea.



(a) Pacific (b) Atlantic (c) Indian (d) Antarctic (e) Arctic

Figure 8: We divide the global ocean into five oceans and plot the changes of reconstruction results of our OXYGENERATOR, GFDL-ESM4 historical, and actual observations at different depths.

an error peak in 1923-1928. This large estimation error in three simulation models comes from a severe underestimation of the Black Sea deoxygenation. Figure 7 illustrates a scatter plot depicting the relationship between observed values and reconstructed values from different methods. It can be observed that many data points with low dissolved oxygen concentrations in the observational data are estimated as relatively high dissolved oxygen concentrations in the modeled data. OXYGENERATOR, based on the correction and correlation learning of observations, effectively captures the changes in extreme deoxygenation areas.

E.2 COMPARISON OF RECONSTRUCTION PERFORMANCE WITH VERTICAL DEPTH

Figure 8 depicts the variation of reconstruction performance as the ocean depth changes. We visualize the observed values of dissolved oxygen, OXYGENERATOR reconstruction values, and GFDL-ESM4 historical numerical simulation results by partitioning them across the five oceans. The dissolved oxygen concentration in seawater generally exhibits a pattern of initially decreasing followed by an increase. We observe a high concordance between the reconstruction results of our method and the observed values, demonstrating consistency with the corresponding patterns. In contrast, the results based on the GFDL-ESM4 historical numerical simulation method exhibit significant fluctuations across different oceans, indicating a relatively high variability. This suggests that the modeling of simulation method in vertical depth layers lacks comprehensive consideration.